Yalla: Estimating Muscle Growth, Maintenance, and Decline from Training Logs

Oliver Frings

Version 2.3 (v123) 19 July 2026

Abstract

Yalla estimates, per muscle and for the body overall, whether recent training is enough to grow, hold, or slowly lose muscle size. The estimate reads *training stimulus* from logged sets, loads, and reps; the app never sees the body itself. Two axes drive it: a *dose* axis (effort-weighted weekly sets against volume landmarks) and a *trend* axis (progressive overload measured as the better of a set trend and a load trend). Set effort is captured with a three-level control whose default is estimated from the load–rep relationship against the lifter's own best set. This document defines every quantity, the equations that combine them, the decision rules that produce a state, and the evidence each parameter rests on. It also states the model's assumptions, its validation status, and where it is deliberately conservative. A second, statistical layer — the prediction ledger (Section 8) — closes the loop on the strength axis: the model writes down a probabilistic strength forecast per exercise *before* the outcome exists, grades it when the weeks have passed, and re-parameterizes itself per user from the errors. The specification corresponds to the functions `growthStatus()`, `inferEffort()`, `planForecast()`, and the `lg*` ledger functions in `app.js`.

1 Introduction

Most training apps record what was lifted; few say what it means for muscle. Yalla estimates, per muscle and for the body overall, whether recent training is enough to grow, maintain, or slowly lose size, using only the sets, loads, and reps a user logs. This document specifies that model: the quantities it computes, the equations that combine them, the decision rules that produce a per-muscle verdict, and the evidence behind each parameter. The goal is an auditable feedback signal whose parameters trace to the resistance-training literature.

Sections 3–7 define the training-stimulus model — a mechanistic classifier whose parameters come from the literature and are fitted to nothing — in three layers: measurement (attribution and effort, Sections 3–5), classification (dose, trend, and state, Section 6), and forecasts (Section 7). Section 8 closes the loop the other way: the prediction ledger, a statistical layer that forecasts each lift's strength change, records the forecast before the outcome exists, grades it, and learns per-user parameters from the errors — Section 8.1 states precisely what is measured and what is estimated there. The closing sections collect the evidence with a grade per source (Section 9), the verification suites (Section 10), and the assumptions and limitations.

2 Background and related work

The model synthesises several strands of resistance-training research.

Volume and its counting. Schoenfeld, Ogborn & Krieger [16] established a graded dose–response between weekly set volume and hypertrophy, with each added set contributing on average about +0.37% of muscle across the studied range. The larger meta-regression of Pelland et al. [12] (67 studies, 2,058 subjects) confirmed rising-but-diminishing returns, found that

counting indirect (secondary) sets at about half best fit the data, and found frequency largely neutral once weekly volume is matched. Yalla adopts fractional set counting (1) and derives its volume landmarks from a continuous dose–response curve (9) built on these findings.

Effort and proximity to failure. Sets far from failure contribute little; Refalo et al. [14] report that hypertrophy improves as sets approach failure, with roughly comparable outcomes across zero to four reps in reserve. Yalla scales each set by an effort factor (2) and, because reps in reserve cannot be read from load and reps alone, estimates it by default from the load–rep relationship [3] against the lifter's own best set.

Landmarks, maintenance, and decline. The volume-landmark framework of minimum effective volume (MEV) and maximum adaptive volume (MAV) popularised by Israetel et al. [8] informs the landmark structure, though as a practitioner synthesis it carries the weakest evidence grade here. Bickel et al. [1] showed that built muscle is retained on a small fraction of the volume that built it, which sets the maintenance floor; Mujika & Padilla [9] documented the timeline over which an insufficient stimulus leads to loss, which motivates treating a sustained low dose as decline. Plotkin et al. [13] found that progressing load or reps drives near-identical growth, which is why the trend axis credits progress from either.

Positioning. These landmarks and principles are widely used as coaching guidance but are rarely turned into a per-log computed signal. Yalla operationalises them into a transparent, per-muscle estimate with stated uncertainty and a reproducible verification suite (Section 10).

3 Model overview

The unit of analysis is a muscle group g (chest, lats, quads, and so on). Training history is a set of logged entries; each entry e records one exercise in one session with a date d_e, a top-set load w_e (kg), top-set reps r_e, a set count n_e, an effort code $f_e \in \{0, 1, 2\}$, and a stored volume-load v_e.

Each muscle is placed in one of three states, combined from the two axes: *growing* (adequate dose and progressing), *holding* (a maintenance dose, or flat), and *under-stimulated* (below the volume needed to hold size). The dose axis answers whether there is enough stimulus this week; the trend axis answers whether the stimulus is rising, flat, or falling. A muscle grows only when both are satisfied: enough sets and a rising trend. This mirrors the requirement that hypertrophy needs both sufficient volume and progressive overload.

4 Volume attribution

4.1 Fractional set counting

Each exercise contributes to its primary muscle at full weight and to each secondary muscle at half. Writing the attribution weight of exercise x to muscle g as $\alpha(x, g)$:

$$\alpha(x, g) = \begin{cases} 1 & \text{if } g \text{ is the primary mover of } x, \\ 0.5 & \text{if } g \text{ is a secondary mover of } x, \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

The intuition: a bench press trains the chest first, but the triceps also work hard in it, so each pressing set counts as half a set of triceps work. The 0.5 weight is the fractional counting scheme that best fit the hypertrophy data in the Pelland et al. dose–response meta-regression [12].

4.2 Effort weighting

Only sets taken reasonably close to failure drive much growth, so each set is scaled by an effort factor $\varepsilon(f)$ before it counts, where $f \in \{0, 1, 2\}$ is the entry's effort code:

$$\varepsilon(f) = \begin{cases} 0.5 & \text{Easy } (f{=}0,\ \geq 4 \text{ reps in reserve}), \\ 1 & \text{Hard } (f{=}1,\ \sim 1\text{–}3 \text{ in reserve; the default}), \\ 1 & \text{Max } (f{=}2,\ 0\text{–}1 \text{ in reserve}). \end{cases} \tag{2}$$

The intuition: a set stopped far short of failure delivers roughly half the growth signal of a hard set, so it counts half. Max carries no growth premium over Hard: hypertrophy is comparable across roughly 0–4 reps in reserve, and training to failure mainly adds fatigue [14]. Max still helps the load trend indirectly through the extra reps it buys. Entries with no recorded effort (all logs predating the feature) default to Hard, so $\varepsilon = 1$ and historical counts are unchanged.

4.3 Weekly effective sets per muscle

Time is bucketed into $N = 10$ rolling weeks; week index $k = 9$ is the current week. The effective weekly sets $S_{g,k}$ for muscle g in week k sum every contributing entry:

$$S_{g,k} = \sum_{e \,\in\, \text{week } k} n_e \, \varepsilon(f_e) \, \alpha(x_e, g), \tag{3}$$

where the sum runs over all entries e logged in week k, and n_e, f_e, and x_e are the entry's set count, effort code, and exercise. In plain terms, the week's training for a muscle is counted set by set, with easy sets discounted and helper-muscle work counted at half. This same effort-weighted, fractional count feeds the weekly-sets chart, the balance radar, and the "hard sets" ring, so every surface in the app reads consistently.

5 Effort inference

Proximity to failure cannot be recovered from load and reps alone: the same 100 kg × 8 can be an all-out set or an easy one. So effort is a self-report. To keep friction to one optional tap, its default is estimated from the numbers and shown pre-selected; the lifter overrides only when the estimate is wrong.

5.1 Estimated one-rep max (anchor)

Set strength is summarised by the Epley one-rep-max (1RM) estimate $\varphi(w, r)$ [3]:

$$\varphi(w, r) = w \left(1 + \frac{r}{K}\right), \tag{4}$$

where w is the set's load (kg), r its reps, and K the rep denominator (population default $K = 30$; Section 5.3 calibrates it per lifter). K is the exchange rate between reps and load: every rep completed at a weight raises the implied one-rep ceiling by $1/K$ of that weight, so at $K = 30$ a set of ten signals a max about a third above the bar weight. A small K describes a lifter whose reps fall off quickly as the bar approaches their limit; a large K, one who can grind out many reps near it.

Epley is one of several standard one-rep-max formulas (Brzycki, Lombardi, and Wathan are common alternatives), and all agree closely over the one-to-ten-rep range where the estimate is used. It is preferred here because it is the simplest — linear in reps, with one parameter — and that simplicity pays twice: the formula inverts in closed form to predict the reps a load allows (6), and K can be recovered from two of the lifter's own sets (8) without curve fitting. A

measured max would be more direct, but deliberate max attempts are risky, rarely performed, and go stale as the lifter progresses; the estimate reads strength from the sets a lifter logs anyway.

The anchor A_x for exercise x is the best estimate over the lifter's *prior* sessions for that exercise:

$$A_x = \max_{e \in \text{hist}(x)} \varphi(w_e, r_e), \tag{5}$$

where $\text{hist}(x)$ is the set of the lifter's prior logged entries for exercise x (the current, in-progress session is excluded) and w_e, r_e are each entry's load and reps. The anchor is simply the strongest set the lifter has shown on that exercise so far — the yardstick today's sets are judged against.

5.2 Reps in reserve and classification

Inverting (4) gives the reps a lifter should complete at load w before failure; subtracting the reps actually done estimates the reps in reserve (RIR):

$$\text{RIR} = K\left(\frac{A_x}{w} - 1\right) - r, \tag{6}$$

where A_x is the anchor (5), w and r are the load and reps of the set under evaluation, and K is the rep denominator of (4); the estimate is evaluated on the session's hardest set for that exercise (the one with the largest φ). In words: from the anchor the model predicts how many reps today's load should allow; whatever the lifter left undone is the estimated reps in reserve. It maps to an effort code f:

$$f = \begin{cases} 2\ (\text{Max}) & \text{if RIR} \le 1, \\ 0\ (\text{Easy}) & \text{if RIR} \ge 4, \\ 1\ (\text{Hard}) & \text{otherwise.} \end{cases} \tag{7}$$

The bands are deliberately coarse: day-to-day strength varies by a rep or two, so finer distinctions would mostly be noise. The anchor is recomputed on every estimate. Because it rolls, it adapts to deloads and off days and cannot lock in a wrong baseline, the failure mode of a one-time calibration. When there is no prior baseline for the lift, or the exercise carries no external load (bodyweight or timed holds), (6) is not evaluated and the default stays Hard.

5.3 Self-calibrated rep denominator

The denominator K — the reps-to-load exchange rate of (4) — varies by lifter and by lift. A set logged as Max is a near-true failure point, so two Max sets of the same exercise at different loads pin K: setting φ equal for both, $w_1(1 + r_1/K) = w_2(1 + r_2/K)$, which solves to

$$K = \frac{w_2 r_2 - w_1 r_1}{w_1 - w_2}, \tag{8}$$

where (w_1, r_1) and (w_2, r_2) are the loads and reps of the two Max sets. Intuitively, two all-out sets at different weights trace the lifter's personal trade-off between load and reps, and K is the steepness of that trade-off. Valid within-exercise pairs are pooled across the whole log, the median is taken and clamped to $[20, 40]$, and $K = 30$ is used until enough failure data exists. This grounds the effort estimate in the lifter's own data.

6 The growth signal: dose, trend, and state

With the measurement layer in place — effective weekly sets (3) built from attribution, effort weighting, and the inferred effort defaults — this section defines the classifier: the landmarks that calibrate the dose axis, the windows and ratios that form the trend axis, and the decision rules that combine both into a per-muscle state and an overall verdict.

6.1 Volume landmarks

Four thresholds, in effective sets per muscle per week, define the dose axis. They follow the MEV/MAV framework and the dose–response meta-analyses.

Symbol (value)	Name	Meaning
MV (2)	Maintenance volume	Floor to hold existing size. Below it, a previously-trained muscle reads as under-stimulated.
MEV (6)	Minimum effective volume	Lower edge of the productive growth window; at or above it, an adequate dose is present.
target (10)	Practical target	Where most of the dose–response benefit has accrued; the balance-radar goal.
MAV (20)	Maximum adaptive volume	Upper productive bound; beyond it, added sets mostly add fatigue with little further growth.

A single set of landmarks is applied to every muscle. Real per-muscle landmarks differ, but one baseline keeps the output legible. The maintenance floor is deliberately low: built muscle is held on very little (as little as roughly one-ninth of the building volume in young trained adults [1]), so a 2-set floor avoids flagging loss too early.

6.2 Continuous backbone

The four landmarks discretise a continuous relationship [16, 12]. We model the fraction of the attainable per-muscle growth stimulus at s effective sets per week as a saturating curve $\rho(s)$, calibrated so the practical target captures about 80% of what is attainable:

$$\rho(s) = 1 - e^{-s/s_0}, \qquad s_0 = \frac{\text{target}}{\ln 5} \approx 6.21, \tag{9}$$

where s is the muscle's effective weekly set count and the scale s_0 is fixed by requiring $\rho(\text{target}) = 0.8$ at the practical target of 10 sets. The curve encodes diminishing returns: the first few weekly sets buy the most stimulus, each added set buys less, and past twenty there is almost nothing left to buy. The landmarks are read off this single curve: $\rho(\text{MV}) \approx 0.28$, $\rho(\text{MEV}) \approx 0.62$, $\rho(\text{target}) \approx 0.80$, and $\rho(\text{MAV}) \approx 0.96$. So MEV is the ≈ 0.6 knee, target the 0.8 point, and MAV the ≈ 0.95 near-plateau. This replaces four independent thresholds with one meta-analytic function; the growth signal also reports $\rho(D_g)$ as a per-muscle share of the attainable growth stimulus.

6.3 Temporal windows

Both axes compare a recent window against an earlier one, each a three-week mean over the 10-week series. For any weekly series X_g, the window means $\text{recent}(X_g)$ and $\text{earlier}(X_g)$ are:

$$\text{recent}(X_g) = \frac{1}{3}\sum_{k=7}^{9} X_{g,k}, \qquad \text{earlier}(X_g) = \frac{1}{3}\sum_{k=4}^{6} X_{g,k}, \tag{10}$$

where $X_{g,k}$ is the series value for muscle g in week k, with week indices running $k = 0, \ldots, 9$ and $k = 9$ the current week. The comparison simply asks how the last three weeks stack up against the three before them. The three-week mean smooths single missed or heavy sessions. The recent window is weeks 8–10 (the last three), the earlier window is weeks 5–7; the oldest three weeks provide history for charts but do not enter the comparison. Because the window is recent, the signal responds within about three weeks of a genuine change in training.

6.4 Dose

The current dose D_g for muscle g is the recent mean of effective weekly sets:

$$D_g = \text{recent}(S_g), \tag{11}$$

where S_g is the effective-set series (3) and recent the window mean of (10). This is the model's answer to "how much is this muscle being trained right now?"

6.5 Trend: progressive overload two ways

Overload can come from more effective sets or from heavier load; either counts as progress [13, 15]. A set trend T_g^S and a load trend T_g^L are formed as recent-over-earlier ratios. The load series $L_{g,k}$ is the best estimated one-rep max reached in week k on exercises for which g is the primary mover:

$$T_g^S = \frac{\text{recent}(S_g)}{\text{earlier}(S_g)}, \qquad T_g^L = \frac{\text{recent}(L_g)}{\text{earlier}(L_g)}, \tag{12}$$

where S_g is the effective-set series (3), L_g the load series just defined, and recent and earlier the window means of (10). Each ratio is defined only when its earlier window is positive; otherwise it is treated as absent. The combined trend τ_g takes the better of the two:

$$\tau_g = \max\left(T_g^S,\, T_g^L\right), \tag{13}$$

with an absent ratio contributing 0 to the maximum, and τ_g itself absent only when both ratios are absent (a muscle just started, with no earlier baseline). In words: is the muscle doing more sets, or lifting heavier, than it did a month ago? Whichever answer is more favourable is the trend. Taking the maximum prevents classic strength progression (add load, drop reps) from being misread as a decline: raw tonnage would fall there while T^L rises.

An earlier version used tonnage, $V = w\,r\,n$, as the sole trend signal. Tonnage rewards total kilograms moved, so lighter high-rep work inflates it while heavier low-rep work deflates it; it is a poor proxy for stimulus, and (13) replaces it.

6.6 Per-muscle state

The state σ_g is a function of the dose D_g and the trend τ_g, evaluated top to bottom (first matching rule wins):

1. **Under-stimulated** if $D_g < \text{MV}$ (below the maintenance floor of 2).
2. **Under-stimulated** if τ_g exists and $\tau_g < 0.75$ (training eased off sharply).
3. Otherwise, if $D_g \geq \text{MEV}$ (adequate dose, ≥ 6): **growing** if τ_g is absent or $\tau_g \geq 1.08$; **holding** otherwise ($0.75 \leq \tau_g < 1.08$).
4. Otherwise (maintenance band, $2 \leq D_g < 6$): **growing** if $\tau_g \geq 1.15$ and $D_g \geq 5$ (climbing toward a growth dose); **holding** otherwise.

The asymmetry is intentional: a muscle grows only with both an adequate dose and a rising trend, but it can be flagged under-stimulated by a low dose alone. Holding is the wide middle band: enough volume to keep size, below the dose that adds it.

6.7 Borderline states

The inputs carry known noise: effort-weighted sets to about ± 1 set, and each trend ratio to about ± 0.06. When a muscle's dose or trend sits within that range of a decision boundary, it is flagged *borderline* and shown with a tilde to mark the verdict as provisional. The sensitivity analysis (Section 10) shows the growth-trend boundary is where this matters most.

6.8 Overall verdict

Let n_{grow}, n_{hold}, n_{shrink} count trained muscles in each state, and m be the number of trained muscles. Let τ_{all} be the recent-over-earlier ratio of whole-body weekly volume. The headline is:

1. **Need more data** if fewer than 3 weeks show any training, or $m = 0$.
2. **At risk** if $n_{\text{shrink}} > n_{\text{grow}}$ and $n_{\text{shrink}} \geq \lceil m/3 \rceil$.
3. **Growing** if $n_{\text{grow}} \geq \max(1,\ \text{round}(0.4\,m))$ and τ_{all} is absent or ≥ 0.95.
4. **Holding** otherwise.

The "at risk" headline requires under-stimulated muscles to both outnumber growing ones and reach a third of those trained, so a few neglected small muscles do not flip the whole read. It resolves to "at risk" chiefly when most of recent training is genuinely sparse.

7 Forecasts

The signal of Section 6 reads history. Two forward-looking companions complete the mechanistic model: the plan forecast checks a plan's prescribed dose against the growth window before any of it is logged, and the Monte Carlo projection turns the same mechanisms into a probabilistic muscle-gain trajectory.

7.1 Plan forecast

Separately from the history-based signal, the plan forecast projects a plan's *prescribed* weekly sets before any of it is logged. Prescribed sets per rotation are scaled to a weekly rate $\hat{s}_g$ by how often the plan is run, capped so an ambitious frequency on a short rotation cannot over-promise:

$$\hat{s}_g = s_g \cdot \min\left(1.2,\ \frac{p}{W}\right), \tag{14}$$

where s_g is prescribed sets per rotation for muscle g (primary plus half of secondary, as in (1)), p is planned sessions per week, and W is the number of distinct rotation workouts. In words: cycling through the rotation more often than once a week delivers proportionally more weekly sets, and the cap keeps an ambitious schedule from projecting more weekly work than anyone sustains. Each trained muscle is then sorted into under-dosed ($<$ MV), growth window (MEV to 20), past-productive ($>$ 20), or maintenance, and the counts drive a plain-language outlook. The forecast promises no centimetres or kilograms; it reports the dose against the growth window and what to change.

7.2 Monte Carlo growth projection

A point prediction of muscle gain would overstate what the model knows, so the in-app forecast is probabilistic. Each of 500 runs accumulates a weekly fractional gain Δ_t

$$\Delta_t = b \cdot m \cdot a \cdot \rho(D) \cdot \eta \cdot o \cdot e^{-t/32}, \tag{15}$$

where t counts weeks, b is a training-age base rate [2], m is the inter-individual response multiplier, drawn lognormal to span the range Hubal et al. [7] observed (cross-sectional-area (CSA) change from about -2% to $+59\%$), a is an age factor that attenuates hypertrophy past about 30 [10], $\rho(D)$ is the dose–response of (9) [16, 12] evaluated at the scenario's effective weekly dose D, η scales with proximity to failure [14], o is near 1 while progressing and drops without added load [13], and the factor $e^{-t/32}$ tapers the weekly rate toward the lifter's ceiling with a

32-week time constant. In words: each week adds a small slice of growth — bigger with a larger dose, harder sets, and continued progression; how big varies a lot between people; and every slice shrinks as the lifter nears their ceiling. Sex is a separate lever set near neutral: relative hypertrophy is comparable between sexes [11], so it bears on absolute mass while the percentage gain here stays about the same for either. Below maintenance the increment turns negative (atrophy [9]), so a starved scenario trends down and the axis extends below zero. Most terms are well constrained by the literature; b carries a wide prior, and m dominates the spread, so the width of the band is the honest output. The multiplier m is where the strength ledger feeds back in: once it has learned a lifter's strength response, m is drawn from the bridge conditional (19) rather than the population prior (Section 8.6), so a projection that starts population-generic personalises as strength data accrues.

7.3 Simulated regimes and sensitivity

The figures that follow project muscle gain under different training regimes from the Monte Carlo forecast just defined. A seeded script (`research/simulate.mjs`) generates them and reproduces every value on re-run.

At a novice base rate, an adequate plan trained with progressive overload gains most over 24 weeks; the same volume with no added load reaches about half that, an irregular schedule (∼55% of sessions) less again, and training below maintenance loses tissue. More weekly volume across the three standard plans gives more gain, with diminishing returns and heavily overlapping bands (Fig. 1).

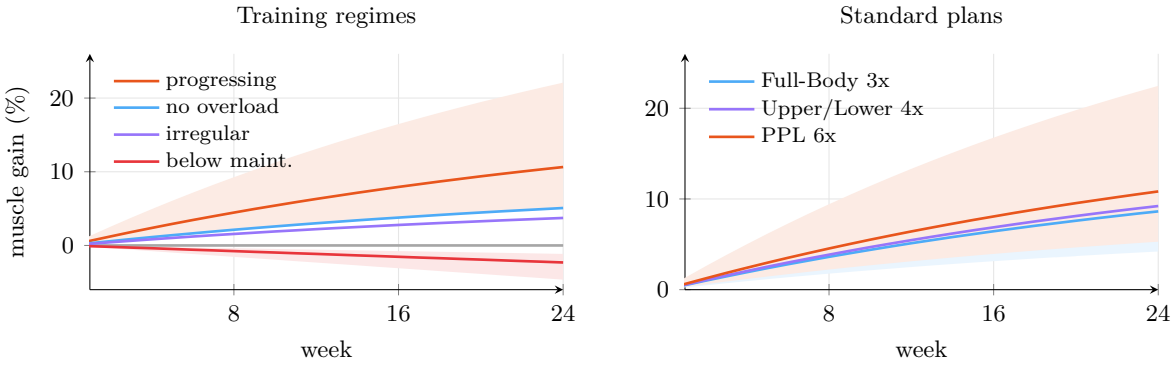

Figure 1: Projected muscle gain over 24 weeks (median; 10–90% bands). *Left:* four regimes at a novice base rate — the progressing and no-overload lines share a volume, so their gap is progressive overload, and below maintenance the median crosses zero. *Right:* the three standard plans — Full-Body, Upper/Lower, and Push/Pull/Legs (PPL) — each progressing.

On the same progressing plan, a novice gains most and an advanced lifter least across three training-age base rates [2]. Across a plausible range for each input, training experience and progression move the 16-week median gain most and age least; Section 10 gives the matching threshold sensitivity (Fig. 2).

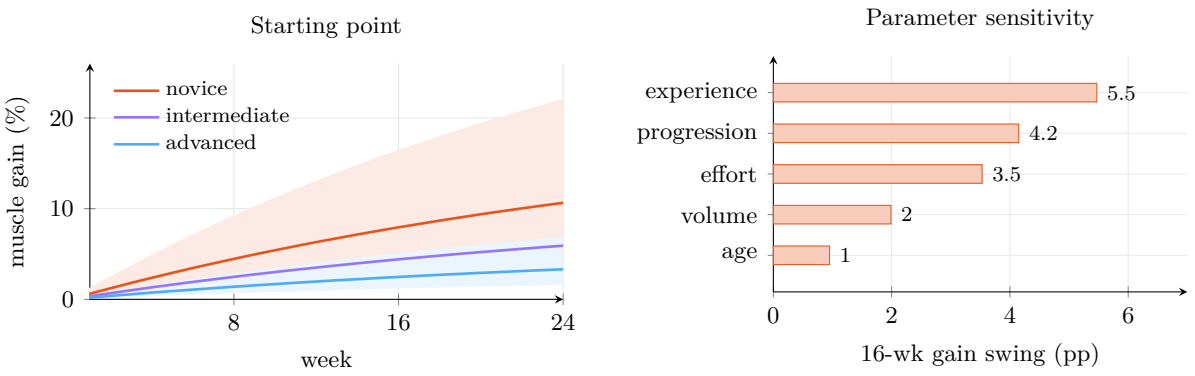

Figure 2: *Left:* the same progressing plan from three starting points (training-age base rates); newer lifters gain faster, all bands wide and overlapping. *Right:* swing in the 16-week median gain, in percentage points (pp), as each input moves across a plausible range, sorted by influence.

8 The prediction ledger and self-calibration

Everything up to here is a mechanistic model: its parameters come from the literature and none is fitted to an outcome. This section adds the missing loop. The app cannot observe muscle size, but it can observe *strength* — the estimated one-rep max φ of (4) is computed on every logged set — so strength is where prediction can meet observation. The protocol is fixed ex ante in `CALIBRATION-PLAN.md` (which names this layer Part II, as does the code); this section summarises it.

8.1 What is measured and what is estimated

This layer is easy to misread as "a model of each lift's one-rep max", so the division of labour is stated up front.

The one-rep max is a measurement, not a model output. $\varphi(w, r)$ is a deterministic transformation of a logged set — a way to put 100 kg × 8 and 110 kg × 4 on one scale. The only calibrated quantity inside it is the rep denominator K (Section 5.3). No latent strength state is estimated, and no trajectory is fitted through the one-rep-max series: the weekly peak $P_{x,k}$ defined next is simply the best measured value of the week.

What the model estimates is a rate. The forecast target $Y_{x,k}$ (16) is the four-week log-*change* of a lift's weekly peak — how fast strength is moving, not where it stands. The forecast mean (17) is a fixed mechanistic function of the training inputs (dose, effort, overload — all frozen from the mechanistic model) multiplied by essentially one learned number per lifter: the response multiplier m_u, plus small per-muscle deviations around it. A statistician would recognise a regression in which everything except the scale is a known offset; only the scale is learned.

Learning is sequential, not batch. Each graded forecast updates the posterior for m_u in closed form (18) — a Kalman filter with a static state. Nothing is refitted over the history; the ledger record carries everything the update needs.

The ledger itself is not a model. It is the evaluation protocol — the same discipline as weather-forecast verification: state the predictive distribution before the outcome exists, then score it with proper scoring rules and coverage checks (Section 8.7).

Why this shape. A lift yields at most one usable observation per week, noisy and gapped — far too little to identify a per-exercise dynamic model without it mostly reproducing its own assumptions. Borrowing the structure from the literature and learning only "how strongly does this lifter respond" is the largest parameter set the data honestly supports; per-muscle and per-exercise effects are added only with shrinkage, only as data accumulates.

8.2 The observable

Time is bucketed into absolute calendar weeks (week k means the same seven days no matter when the computation runs). For exercise x, the weekly peak $P_{x,k}$ is the best estimated one-rep max the lifter reached in week k. The outcome of a forecast made at week k with horizon $H = 4$ weeks is the realised log-gain

$$Y_{x,k} = \log \frac{\max\left(P_{x,k+H-1},\ P_{x,k+H}\right)}{\max\left(P_{x,k-1},\ P_{x,k}\right)}, \tag{16}$$

where the numerator is the best weekly peak over the two *final* weeks of the horizon and the denominator the best over the two weeks up to and including the forecast week. In words: where the lift ended up against where it stood, each read from a two-week window so one bad day does not decide either end. The two windows must be the same length: the maximum of four noisy weeks is systematically higher than the maximum of two, and an asymmetric definition would bias every score upward.

8.3 The forecast

At most once per exercise per week — when a workout is saved — the model emits a forecast of $Y_{x,k}$ as a normal distribution with mean

$$\mu_{x,k} = H\, r_0\, m_u\, \rho_S(D_g)\, \eta\, o(\tau_g), \tag{17}$$

where H is the horizon (4 weeks), r_0 the population base rate of weekly log-strength gain at full stimulus (0.0075, i.e. about 0.75% per week), m_u the lifter's personal response multiplier (the learned quantity, prior centred on 1), $\rho_S(D_g)$ the *strength* dose–response at the primary muscle's recent effective weekly dose D_g, η the mean effort factor (2) of the recent sets, and $o(\tau_g)$ an overload factor that is 1 while the trend τ_g (13) shows progression, 0.5 when flat, and 0.15 after a sharp drop. The dose–response here is *not* the hypertrophy curve (9): strength saturates with volume faster than size (Pelland et al. [12] report "considerably more pronounced diminishing returns" for strength), so the ledger uses its own $\rho_S(s) = 1 - e^{-s/s_{0,S}}$ with $s_{0,S} = \text{target}_S / \ln 5 \approx 3.11$ at a strength practical target of $\text{target}_S = 5$ effective sets/wk (versus 10 for hypertrophy). Thus $\rho_S(5) = 0.8$ where $\rho(5) \approx 0.55$: the same volume reads as "most of the available strength benefit" but only "more than half" the size benefit. The effort factor η is kept identical to Part I — strength is somewhat less failure-sensitive than hypertrophy, but that difference sits below the resolution of the three coarse effort bands, and strength's load-dependence already enters through the e1RM outcome and $o(\tau_g)$. The forecast's standard deviation combines the posterior uncertainty about m_u, a residual term, and the measurement noise of the weekly peak (estimated per exercise from median-detrended week-to-week variation). In words: the predicted gain is the base rate, scaled by who you are, how much you train the muscle, how hard the sets are, and whether the load is still moving; the band around it says how sure the model is entitled to be. The multiplier m_u is the learned counterpart of the response multiplier the Monte Carlo forecast draws at random (Section 7.2): what the growth projection samples from the population distribution because it cannot know the individual, the ledger estimates for the individual from their own graded forecasts.

8.4 The ledger

Each forecast is appended to a ledger with its inputs and the parameter posterior *frozen* at emission. Scoring, when week $k+H$ completes, writes only the outcome block: the realised Y, the probability integral transform (PIT — where the outcome landed in the forecast distribution), the continuous ranked probability score (CRPS — a proper scoring rule that rewards being both right and appropriately confident [4]), and whether the central 80% interval covered the outcome.

Predictions are never edited after the fact; records that cannot be scored (the lifter stopped logging that exercise) are labelled unscorable and kept, because silently dropping them would flatter the statistics. Backfilled "predictions" over pre-ledger history are labelled retrodictions and never feed calibration.

8.5 Re-parameterization

Let $\theta_u = \log m_u$ with prior $\theta_u \sim \mathcal{N}(0,\, 0.35^2)$, wide enough to span the inter-individual range of Hubal et al. [7]. Each scored forecast updates the posterior by a conjugate normal step, linearised at the current posterior mean:

$$\theta_{\text{obs}} = \hat{\theta} + \frac{y - \mu}{\mu}, \qquad p_{\text{obs}} = \frac{c\,\mu^2}{\sigma^2_{\text{tot}}}, \qquad \hat{\theta} \leftarrow \frac{p_0\,\hat{\theta} + p_{\text{obs}}\,\theta_{\text{obs}}}{p_0 + p_{\text{obs}}}, \tag{18}$$

where y is the realised log-gain (16), μ the forecast mean at the current estimate $\hat{\theta}$, σ^2_{tot} the non-parameter part of the forecast variance, p_0 the prior (incoming) precision, and p_{obs} the observation's precision. The factor $c = 1/(H{+}2)$ is an effective-information correction: forecasts are emitted weekly but their four-week outcomes overlap, so consecutive scores are far from independent — $1/H$ of the correction follows from the overlap itself, the remainder from residual dependence, calibrated in simulation. In words: each graded forecast nudges the personal multiplier toward whatever ratio of "happened over predicted" it showed, with a weight that honestly reflects how little new information one overlapping week carries. A weak training week moves it barely at all (μ small $\Rightarrow$ precision small); a posterior clamp at $[\times 0.2, \times 3]$ keeps data errors from claiming impossible physiology. A second, tighter stage learns per-muscle deviations around the personal multiplier once a muscle has enough scored records.

8.6 From strength back to size: the bridge

The ledger learns a lifter's *strength* response; the Monte Carlo projects *muscle size* and, on its own, draws the individual response multiplier at random because it cannot see the person. The bridge lets the learned strength response inform that draw. Write $\theta_S = \log m_u$ (learned, posterior $\mathcal{N}(\hat{\theta}_S, v_S)$) and $\theta_H = \log(\text{indiv})$ (the hypertrophy multiplier the projection needs). Treat them as bivariate normal with correlation ρ_{sh}; conditioning θ_H on the strength posterior and integrating over it gives, again normally,

$$\theta_H \mid \text{data} \sim \mathcal{N}\Big(\beta\,\hat{\theta}_S,\; \sigma^2_H(1-\rho^2_{sh}) + \beta^2 v_S\Big), \qquad \beta = \rho_{sh}\,\frac{\sigma_H}{\sigma_S}, \tag{19}$$

and the Monte Carlo draws $\text{indiv} = e^{\theta_H}$ from this instead of from $\mathcal{N}(0, \sigma^2_H)$. The correlation is set deliberately low, $\rho_{sh} = 0.3$, because strength is a *biased, weak* proxy for size: the individual-level correlation between strength change and size change is small (Ahtiainen et al. [5], $r \approx 0.16$; Balshaw et al. [6] find hypertrophy explains ∼19% of strength-gain variance, behind neural drive at ∼31%), and early strength gain is largely neural and skill-based. At $\rho_{sh} = 0.3$ the bridge *shifts* the muscle-gain centre by $\beta\hat{\theta}_S$ (a strong strength responder nudges the size forecast up ∼18%) but the band $\sigma_H\sqrt{1-\rho^2_{sh}}$ barely narrows: strength informs the size forecast without pretending to pin it. At $\rho_{sh} = 0$ the draw reverts exactly to the unconditional one. The bridge activates only after ten scored forecasts (the same gate as the card), so it never personalises on no data. This is the pragmatic connection; the principled successor — a per-muscle latent factor with a fast exercise-specific skill term and a slow shared-capacity term that loads on both strength and size — is noted for future work (`CALIBRATION-PLAN.md` §12).

8.7 Validation before belief

Three reproducible checks accompany the ledger (`research/ledger.test.mjs`): known-answer tests of the ledger mechanics; *simulation-based calibration* — users drawn from the prior, the

full loop run on their synthetic logs, then verifying that interval coverage matches the nominal 80%, that the truth is uniformly distributed under the reported posterior (final z-scores: mean 0.18, sd 1.07 against targets 0 and 1), and that recovery reaches the ceiling the posterior width permits; and numeric parity between the app implementation and the reference module. A walk-forward backtest (`research/ledger-backtest.mjs`) replays any exported history with an expanding window — no future data can leak into a forecast — and reports coverage, CRPS, and skill against two fixed baselines: "no change" and "personal drift" (the lifter's own trailing mean gain). The model earns its keep only if it beats the second. The forecast card in the app shows numbers only after ten scored forecasts; until then it says it is calibrating.

The bridge (19) is checked the same way: synthetic lifters are generated with a known (θ_S, θ_H) pair at correlation ρ_{sh}, the ledger learns θ_S from their logs, and the conditioned size forecast is scored against the true θ_H. At $\rho_{sh} = 0$ it reproduces the unconditional forecast exactly; at a strong coupling it clearly reduces error; at the production $\rho_{sh} = 0.3$ the conditioned centre tracks the truth (small but positive) and the interval stays at or above nominal coverage — i.e. the bridge is never overconfident.

8.8 Which exercises drive strength

With outcomes and exposures defined, a second question opens: which exercises contribute most? Self-progression is uninformative here — practising a lift raises its own φ through skill alone — so the estimand is *transfer*: the association between one weekly effective set of exercise x and the H-week log-gain of the *other* lifts sharing x's muscle, within person, with the lift's own practice, the pre-window dose state, the trend, and time held fixed. Per-exercise coefficients are shrunk toward zero (partial pooling as the multiplicity control), uncertainty comes from a moving-block bootstrap that respects the weekly serial correlation, and the output language stays associational: this is an observational design with known limits, not a randomised trial. The analysis ships as a research script (`research/effectiveness.mjs`) with a known-truth synthetic self-check; pooled estimation across consenting users is designed but deferred.

9 Scientific basis

Each parameter and mechanism traces to a specific source. Effect sizes are context-dependent; the model uses the literature to set its structure and thresholds, and promises no outcomes. Grades follow the GRADE approach (Grading of Recommendations, Assessment, Development and Evaluation) to rating the underlying evidence: High (meta-analyses of randomised controlled trials, RCTs), Moderate (individual RCTs or focused reviews), Low (practitioner framework). A Low grade flags a parameter that rests on weaker evidence and is the first candidate for revision (Table 1).

10 Verification

No body-composition study validates the model (see the limitations that follow). Three reproducible checks accompany the training-stimulus signal (the `research/` folder in the repository), each raising rigor without new measurement; the prediction ledger carries its own verification suite (Section 8.7), including simulation-based calibration and a walk-forward backtest.

- **Known-answer tests.** Canonical scenarios — a returning lifter, a deload, easy versus hard sets, a hard stop — paired with domain-expected verdicts, so miscalibration or later drift is caught in code.
- **Sensitivity analysis.** Perturbing each threshold by ±10% over a synthetic population shows the growth-trend cutoff is the most load-bearing parameter (about a quarter of muscles change

Mechanism / parameter	Basis	Grade
Graded volume dose–response; ∼10 sets as a practical target	Schoenfeld, Ogborn & Krieger [16]: each added weekly set ≈ +0.37% muscle, benefit rising across the studied range.	High
Fractional (0.5) counting of indirect sets; diminishing returns; frequency neutral at equal volume	Pelland et al. [12] (peer-reviewed; 67 studies, 2,058 subjects).	High
Continuous backbone $\rho(s)$ and landmark derivation	Derived from the dose–response curves above [16, 12].	High
Effort factor ε; hypertrophy comparable across ∼0–4 RIR, dropping off when easier	Refalo et al. [14], systematic review and meta-analysis of proximity-to-failure.	High
MEV/MAV landmark structure (∼10 to 20 sets)	Israetel et al. [8], volume-landmarks framework.	Low
Maintenance floor MV; size held on ∼1/3 (older) to ∼1/9 (young) of building volume	Bickel, Cross & Bamman [1] (RCT, 16-wk build plus 32-wk reduced dose).	Moderate
Decline below maintenance stimulus; ∼2–4-week strength retention, then loss	Mujika & Padilla [9] detraining review; cross-sectional-area loss from ∼2–3 weeks of cessation.	Moderate
Overload via added load *or* reps; the trend tracks both	Plotkin et al. [13] (RCT; load- and rep-progression gave near-identical growth); mechanism in Schoenfeld [15].	Moderate
Epley one-rep-max φ; self-calibrated denominator K	Standard load–rep relationship [3]; K fitted from the lifter's own Max sets (8).	Method
Monte Carlo forecast: training-age base rate	Damas et al. [2]; hypertrophy rate depends on training status.	Moderate
Monte Carlo forecast: inter-individual response spread	Hubal et al. [7] (n=585); 12-wk CSA change ≈ −2% to +59%.	High
Forecast age factor	Peterson et al. [10]; lean-mass gains attenuate with age.	Moderate
Forecast sex lever (≈ neutral for % gain)	Roberts et al. [11]; relative hypertrophy comparable between sexes.	High
Ledger prior spread of m_u	Hubal et al. [7]; inter-individual strength-response range sets the prior width (18).	High
Ledger base rate r_0	Prior centred near 0.75%/wk, consistent with the strength arm of Pelland et al. [12]; the posterior learns the product $r_0 m_u$ per user.	Prior
Ledger scoring (CRPS, PIT, coverage)	Gneiting & Raftery [4], proper scoring rules; protocol in `CALIBRATION-PLAN.md`.	Method

Table 1: Evidence basis for each model parameter, with GRADE-style ratings.

state), while the maintenance floor and sharp-drop cutoff are nearly cosmetic. This is why the borderline band sits on the trend boundary.

- **Predictive-coherence backtest.** On a training log, muscles labelled growing should precede larger subsequent gains in estimated 1RM than those labelled holding or under-stimulated. This one-off check matured into the prediction ledger (Section 8), where the same idea runs continuously. Strength change is only a proxy for hypertrophy; a body-composition study remains the standard these checks cannot replace.

11 Assumptions and limitations

- **Validation status.** The training-stimulus model is mechanistic, grounded in the literature and internally consistent. It has no validation against direct body-composition measurements (dual-energy X-ray absorptiometry, ultrasound, or magnetic resonance imaging): its thresholds are derived from published dose–response findings, none are fitted to an outcome dataset, and it claims no predictive-accuracy interval. The prediction ledger (Section 8) fits per-user parameters to *strength* outcomes under a pre-registered protocol and reports its own calibration; strength remains a pro
- **A stimulus estimate.** Every output estimates the training stimulus from logs. The app cannot see the body; no measurement of muscle mass is implied.
- **Effort is self-reported.** Reps in reserve is not recoverable from load and reps alone; (6) sets only a default. Day-to-day max-rep variability ($\pm$1–3 reps) sits at the resolution of the effort bands, so the estimate is noisy near failure and the manual override remains the ground truth.
- **Top-set inference.** History stores each exercise's top set and its set count, without the reps of every set, so within-session rep drop-off is not yet used as an effort cue.
- **One landmark set for all muscles.** Chosen for legibility; real landmarks vary by muscle and training age.
- **Load trend needs external load.** Bodyweight and timed work do not contribute to T^L; those muscles rely on the set trend alone.
- **Residual tonnage use.** The overall verdict's τ_{all} still uses whole-body volume; the per-muscle trend does not.
- **Conservative by design.** The maintenance floor and the "at risk" threshold are set conservatively, biasing the model toward under-reporting muscle loss.

References

[1] Bickel CS, Cross JM, Bamman MM. Exercise dosing to retain resistance-training adaptations in young and older adults. *Med Sci Sports Exerc.* 2011;43(7):1177–1187. doi:10.1249/MSS.0b013e318207c15d

[2] Damas F, Phillips SM, Libardi CA, et al. Resistance training-induced changes in integrated myofibrillar protein synthesis are related to hypertrophy only after attenuation of muscle damage. *J Physiol.* 2016;594(18):5209–5222. doi:10.1113/JP272472

[3] Epley B. Poundage chart. *Boyd Epley Workout.* Lincoln, NE: Body Enterprises; 1985.

[4] Gneiting T, Raftery AE. Strictly proper scoring rules, prediction, and estimation. *J Am Stat Assoc.* 2007;102(477):359–378. doi:10.1198/016214506000001437

[5] Ahtiainen JP, Walker S, Peltonen H, et al. Heterogeneity in resistance training-induced muscle strength and mass responses in men and women of different ages. *Age (Dordr).* 2016;38(1):10. doi:10.1007/s11357-015-9870-1

[6] Balshaw TG, Massey GJ, Maden-Wilkinson TM, et al. Changes in agonist neural drive, hypertrophy and pre-training strength all contribute to the individual strength gains after resistance training. *Eur J Appl Physiol.* 2017;117(4):631–640. doi:10.1007/s00421-017-3560-x

[7] Hubal MJ, Gordish-Dressman H, Thompson PD, et al. Variability in muscle size and strength gain after unilateral resistance training. *Med Sci Sports Exerc.* 2005;37(6):964–972. doi:10.1249/01.mss.0000170469.90461.5f

[8] Israetel M, Hoffmann J, Smith CW. Training volume landmarks for muscle growth (MEV/MAV/MRV). *Renaissance Periodization*; 2017. `https://rpstrength.com/blogs/articles/training-volume-landmarks-muscle-growth`

[9] Mujika I, Padilla S. Detraining: loss of training-induced physiological and performance adaptations. Part I — short-term insufficient training stimulus. *Sports Med.* 2000;30(2):79–87. doi:10.2165/00007256-200030020-00002

[10] Peterson MD, Sen A, Gordon PM. Influence of resistance exercise on lean body mass in aging adults: a meta-analysis. *Med Sci Sports Exerc.* 2011;43(2):249–258. doi:10.1249/MSS.0b013e3181eb6265

[11] Roberts BM, Nuckols G, Krieger JW. Sex differences in resistance training: a systematic review and meta-analysis. *J Strength Cond Res.* 2020;34(5):1448–1460. doi:10.1519/JSC.0000000000003521

[12] Pelland JC, Remmert JF, Robinson ZP, Hinson SR, Zourdos MC. The resistance training dose response: meta-regressions exploring the effects of weekly volume and frequency on muscle hypertrophy and strength gains. *Sports Med.* 2026. doi:10.1007/s40279-025-02344-w

[13] Plotkin D, Coleman M, Van Every D, et al. Progressive overload without progressing load? The effects of load or repetition progression on muscular adaptations. *PeerJ.* 2022;10:e14142. doi:10.7717/peerj.14142

[14] Refalo MC, Helms ER, Trexler ET, Hamilton DL, Fyfe JJ. Influence of resistance training proximity-to-failure on skeletal muscle hypertrophy: a systematic review with meta-analysis. *Sports Med.* 2023;53(3):649–665. doi:10.1007/s40279-022-01784-y

[15] Schoenfeld BJ. The mechanisms of muscle hypertrophy and their application to resistance training. *J Strength Cond Res.* 2010;24(10):2857–2872. doi:10.1519/JSC.0b013e3181e840f3

[16] Schoenfeld BJ, Ogborn D, Krieger JW. Dose-response relationship between weekly resistance training volume and increases in muscle mass: a systematic review and meta-analysis. *J Sports Sci.* 2017;35(11):1073–1082. doi:10.1080/02640414.2016.1210197

A Symbol reference

Symbol	Definition
$g,\ x,\ e$	Muscle group, exercise, logged entry
$w,\ r,\ n,\ f$	Top-set load (kg), top-set reps, set count, effort code $\{0, 1, 2\}$
$\alpha(x, g)$	Muscle attribution weight: 1 primary, 0.5 secondary, 0 otherwise (1)
$\varepsilon(f)$	Effort factor: 0.5 / 1 / 1 for Easy / Hard / Max (2)
$S_{g,k}$	Effective weekly sets, muscle g, week k (3)
$\varphi(w, r),\ K$	Epley estimated one-rep max and its rep denominator, the reps-to-load exchange rate (K=30 default, self-calibrated) (4), (8)
A_x	Anchor: best prior φ over hist(x), the prior logged entries for exercise x (5)
RIR	Estimated reps in reserve of a set (6)
$L_{g,k}$	Weekly peak φ on primary exercises for g
D_g	Recent dose: recent mean of S_g (11)
$\rho(s),\ s_0$	Hypertrophy dose–response stimulus fraction and its scale (9)
$\rho_S(s),\ s_{0,S}$	Strength dose–response and its scale ($s_{0,S} \approx 3.11$; earlier plateau) (17)
$T^S,\ T^L,\ \tau_g$	Set trend, load trend, combined trend (12)–(13)
σ_g	Per-muscle state (growing / holding / under-stimulated)
MV, MEV, target, MAV	Landmarks 2, 6, 10, 20 effective sets/muscle/week
Prediction ledger (Section 8)	
$P_{x,k}$	Weekly peak: best measured φ for exercise x in absolute week k
H	Forecast horizon, 4 weeks
$Y_{x,k}$	Realised 4-week log-gain, symmetric 2-week endpoint windows (16)
r_0	Population base rate of weekly log-strength gain at full stimulus
$m_u,\ \theta_u$	Personal response multiplier and its log, the learned parameter
$\mu_{x,k}$	Forecast mean (17): mechanistic offset $\times\ m_u$
$o(\tau_g)$	Overload factor: 1 progressing / 0.5 flat / 0.15 after a sharp drop
c	Effective-information correction $1/(H{+}2)$ in (18)
σ^2_{tot}	Non-parameter forecast variance: residual + 2× peak measurement noise
θ_S, θ_H	Strength / hypertrophy log response-multipliers; bridged by (19)
$\rho_{sh},\ \beta$	Strength–size response correlation (0.3) and bridge slope $\rho_{sh}\sigma_H/\sigma_S$ (19)
σ_S, σ_H	Log-scale spreads of the strength / hypertrophy multipliers (0.35 / 0.55)

B Parameter values

Every numeric constant in the model, its value, and where it enters. These are the values used in production (v122); the sensitivity analysis (Section 10) quantifies how much each one matters.

Parameter	Value	Where
Rolling window N	10 weeks	series length (3)
Recent / earlier windows	weeks 8–10 / weeks 5–7	(10)
Secondary-muscle weight	0.5	α (1)
Effort factors (Easy/Hard/Max)	0.5 / 1 / 1	ε (2)
RIR $\rightarrow$ effort bands	Max $\leq$1, Easy $\geq$4, else Hard	(7)
Epley denominator K	30 default, fitted then clamped to $[20, 40]$	(4), (8)
Dose–response scale s_0	target$/\ln 5 \approx 6.21$	$\rho(s)$ (9)
Maintenance volume MV	2 effective sets/wk	state rule 1
Min. effective volume MEV	6 effective sets/wk	state rules 3–4
Practical target	10 effective sets/wk	radar goal; sets s_0
Upper bound MAV	20 effective sets/wk	plan forecast
Sharp-drop trend cutoff	$\tau_g < 0.75$	state rule 2
Growth trend cutoff	$\tau_g \geq 1.08$	state rule 3
Maintenance-band climb	$\tau_g \geq 1.15$ and $D_g \geq 5$	state rule 4
Borderline margins	$\pm$1 set; trend within 0.06 of $\{0.75, 0.92, 1.08\}$	Section 6.7
Overall "at risk"	$n_{\text{shrink}} > n_{\text{grow}}$ and $n_{\text{shrink}} \geq \lceil m/3 \rceil$	Section 6.8
Overall "growing"	$n_{\text{grow}} \geq \max(1, \text{round}(0.4m))$, τ_{all} absent or $\geq$0.95	Section 6.8
"Need more data"	fewer than 3 active weeks	Section 6.8
Plan frequency cap	$\min(1.2,\ p/W)$	(14)
Prediction ledger (Section 8; protocol and changelog in `CALIBRATION-PLAN.md`)		
Horizon H	4 weeks	(16)
Base rate r_0	0.0075 /wk ($\approx$0.75%)	(17)
Strength dose–response $s_{0,S}$	target$_S/\ln 5 \approx 3.11$ (target$_S$=5)	ρ_S (17)
Prior on $\theta_u = \log m_u$	$\mathcal{N}(0,\ 0.35^2)$	(18)
Posterior clamp on m_u	$[\times 0.2,\ \times 3]$	(18)
Effective-information c	$1/(H{+}2) = 1/6$	(18)
Per-muscle deviation prior	$\mathcal{N}(0,\ 0.15^2)$, applied from 6 scored records	second update stage
Overload factor o	1 / 0.5 / 0.15 (progressing / flat / sharp drop)	(17)
Residual variance σ_u^2	starts 0.025^2, re-estimated every 8 scored, floor 0.005^2	forecast sd
Peak measurement sd	median-detrended MAD of weekly log-diffs$/\sqrt{2}$; default 0.02, floor 0.005	forecast sd
Emission gate	$\geq$3 prior loaded weeks; trained this week; baseline exists	ledger protocol
Scoring gate	$\geq$1 loaded entry in weeks $k{+}H{-}1..k{+}H$	(16)
Forecast numbers shown from	10 scored forecasts	app card
Bridge correlation ρ_{sh}	0.3 (conservative; range 0.15–0.5)	(19)
Bridge spreads σ_S, σ_H	0.35 / 0.55 (log scale)	(19)
Bridge activation gate	$\geq$10 scored forecasts and $v_S <$ prior	Section 8.6