Counts Converge, Spacings Do Not: A Two-Level Decomposition of Hardy-Littlewood Deviations in Twin Prime Residue Classes

Anonymous Authors[1]

Abstract

The Hardy-Littlewood conjecture predicts the asymptotic global density of twin primes, but makes no per-class prediction for the 15 valid residue classes modulo 210. We decompose deviations from this conjecture into two independent levels. Using a Python segmented sieve enumerating all 3,424,502 twin prime pairs $(p, p+2)$ up to 10^9: *(i)* count-based deviations per class are at most 0.17% and follow the expected $O(1/\sqrt{N})$ statistical fluctuations, consistent with Hardy-Littlewood being asymptotically correct for counts; *(ii)* gap-spacing deviations, measured via the normalized ratio $\rho(p) = 2C_2\, g(p)/(\ln p)^2$, are 4–5% in magnitude, Bonferroni-significant in 14 of 15 classes ($|t|$ up to 24.2), and persist across all $\log p$ sub-ranges with Spearman $r_S > 0.97$ between adjacent ranges. We rule out the Lemke Oliver-Soundararajan anti-persistence bias as an explanation: self-transition rates are uniformly suppressed below $1/15$ across all classes (Spearman $r_S = -0.04$, $p = 0.90$ against gap deviation), confirming our finding is independent of that effect. Two-stage PySR symbolic regression confirms Hardy-Littlewood at the per-instance level, then at the class level discovers the term m_{70}/δ_7 as the strongest number-theoretic predictor, reducing regression loss by 76% over a constant baseline. The gap-spacing deviation is a persistent higher-order phenomenon not predicted by the standard singular series.

1. Introduction

Twin primes – pairs $(p, p+2)$ with both prime – are among the oldest objects of study in number theory. The Twin Prime Conjecture, asserting their infinitude, remains open, though Zhang (Zhang, 2014) and Maynard (Maynard, 2015) have established that prime gaps are bounded infinitely often. The quantitative distribution of twin primes is governed by the Hardy-Littlewood conjecture (Hardy & Littlewood, 1923):

$$\pi_2(x) \;\sim\; 2C_2 \frac{x}{(\ln x)^2}, \quad C_2 = \prod_{p>2} \frac{p(p-2)}{(p-1)^2} \approx 0.6602, \tag{1}$$

where $\pi_2(x)$ counts pairs $(p, p+2)$ with $p \le x$. This has been verified numerically at the sub-percent level for $x \gtrsim 10^6$ (Brent, 1975; Oliveira e Silva et al., 2014), and is widely believed to be correct, though no proof exists.

The constant C_2 arises from a product over all primes and encodes the global sieve structure. It treats the 15 valid residue classes modulo $210 = 2 \cdot 3 \cdot 5 \cdot 7$ identically – the formula produces a single number for all of them. This paper investigates what happens when you look *within* that average.

The two questions. We ask two questions that Hardy-Littlewood leaves unanswered:

Q1 (Counts). Does each of the 15 valid classes hold an equal $1/15$ share of all twin primes? Hardy-Littlewood implies yes, asymptotically.

Q2 (Spacings). Given $p \equiv r \pmod{210}$, is the gap to the next twin prime distributed with mean $\hat{g}_{\mathrm{HL}} = (\ln p)^2/(2C_2)$? Hardy-Littlewood is silent here.

These questions are mathematically distinct. Count convergence does not imply spacing uniformity. A process can assign the correct long-run proportion to each class while the inter-arrival rhythm within each class systematically differs. We establish empirically that **Q1 is consistent with Hardy-Littlewood** (deviations shrink as expected for finite-N fluctuations) **while Q2 reveals persistent structure** that does not diminish across the full range of our data.

Ruling out an alternative explanation. The Lemke Oliver-Soundararajan (LO&S) result (Oliver & Soundarara-

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

jan, 2016) shows that consecutive primes anti-persist in their residue class: a prime in class r is followed by another in class r less often than the uniform rate. This is a plausible candidate explanation for spacing irregularities. We test it directly in Section 4.4 and find the LO&S effect is present in our data but is *uniform* across all 15 classes, carrying no information about which classes deviate in gap size. The two phenomena are independent.

AI4Math contribution. We demonstrate that PySR (Cranmer, 2023) applied in two stages – first per-instance, then per-class – can *confirm* a known conjecture at one level and *discover partial structure* in the correction at the next level. This two-stage methodology and the count/spacing decomposition constitute a reusable template for AI-assisted conjecture refinement.

2. Background

2.1. The Hardy-Littlewood Singular Series

The Hardy-Littlewood conjecture for prime k-tuples states that for an admissible constellation $\mathcal{H} = \{h_1, \ldots, h_k\}$, the number of $n \le x$ with all $n + h_i$ prime satisfies

$$\pi_{\mathcal{H}}(x) \sim \mathfrak{S}(\mathcal{H}) \frac{x}{(\ln x)^k}, \tag{2}$$

where the *singular series* is

$$\mathfrak{S}(\mathcal{H}) = \prod_{p \text{ prime}} \frac{1 - \nu_p(\mathcal{H})/p}{(1 - 1/p)^k}, \tag{3}$$

and $\nu_p(\mathcal{H})$ is the number of distinct residues $\{h_1 \bmod p, \ldots, h_k \bmod p\}$. For twin primes $\mathcal{H} = \{0, 2\}$, one checks $\nu_2 = 1$ and $\nu_p = 2$ for all $p > 2$, giving

$$\begin{aligned} \mathfrak{S}(\{0,2\}) &= \frac{1 - 1/2}{(1 - 1/2)^2} \prod_{p>2} \frac{1 - 2/p}{(1 - 1/p)^2} \\ &= 2 \prod_{p>2} \frac{p(p-2)}{(p-1)^2} = 2C_2. \end{aligned} \tag{4}$$

Crucially, $2C_2$ is a *global constant*: it does not depend on which residue class modulo 210 the twin prime belongs to. The residue class of p affects only whether $(p, p+2)$ is *admissible*, not the rate at which admissible pairs occur. Among the 15 valid classes, Hardy-Littlewood predicts identical densities.

2.2. The Wheel Modulo 210 and Valid Classes

The primorial $210 = 2 \cdot 3 \cdot 5 \cdot 7$ defines the smallest wheel that eliminates all multiples of primes up to 7. Among the $\phi(210) = 48$ integers coprime to 210, only 15 residue classes r additionally satisfy $\gcd(r + 2, 210) = 1$, guaranteeing that neither p nor $p + 2$ is divisible by 2, 3, 5, or 7. The remaining 33 classes are arithmetically impossible for twin primes: for instance, $r \equiv 1 \pmod 3$ forces $p + 2 \equiv 0 \pmod 3$. These 15 classes are the *complete set*; there is no extension to 48.

Among these 15 classes the singular series constant $2C_2$ is identical and H-L predicts no class-to-class variation in either count share or expected gap size. Any detected variation is therefore a deviation from H-L's implicit prediction, not from something H-L accounts for.

2.3. Normalized Gap Deviation and Count Ratio

We define the *normalized gap deviation* for a twin prime p:

$$\rho(p) \;=\; \frac{g(p)}{\hat{g}_{\mathrm{HL}}(p)} \;=\; \frac{2C_2\, g(p)}{(\ln p)^2}, \tag{5}$$

where $g(p)$ is the gap to the next larger twin prime. Under the Poisson approximation to the Hardy-Littlewood process, gaps are approximately exponentially distributed with mean $\hat{g}_{\mathrm{HL}}(p)$, so $\rho(p)$ is approximately $\mathrm{Exp}(1)$-distributed with $\mathbb{E}[\rho] = 1$. If Hardy-Littlewood holds uniformly across classes, then

$$\mathbb{E}\big[\rho \mid p \equiv r \pmod{210}\big] = 1 \quad \text{for all valid } r. \tag{6}$$

Equation (6) is the null hypothesis tested throughout this paper. We denote the per-class mean $\bar{\rho}_r = \mathbb{E}[\rho \mid p \equiv r \pmod{210}]$.

For count validation, the *per-class count ratio* in range $[L, H)$:

$$R_r(L,H) \;=\; \frac{15 \cdot \big|\{p \in [L,H) : (p, p+2) \text{ twin},\ p \equiv r \pmod{210}\}\big|}{\hat{N}(L,H)}, \tag{7}$$

where $\hat{N}(L,H) = (H - L) \cdot 2C_2/(\ln \sqrt{LH})^2$ is the Hardy-Littlewood expected total count in $[L, H)$. H-L predicts $R_r \to 1$ for each r; we verify this in Section 4.1.

2.4. Lemke Oliver-Soundararajan Bias

Lemke Oliver and Soundararajan (Oliver & Soundararajan, 2016) proved, conditional on the Generalized Riemann Hypothesis, that for primes in residue classes $a, b \pmod q$, the probability that the prime following $p \equiv a$ lies in class b deviates from the naive $1/\phi(q)$ at order $O(\log\log x/\log x)$. Specifically, self-transitions ($a = b$) are *suppressed*: a prime is less likely to be followed by another prime in the same class. This bias has been empirically confirmed for regular primes (Oliver & Soundararajan, 2016) and for various prime-like sequences (Granville & Martin, 2006).

The relevance to our setting is as follows. If a twin prime in class r is followed unusually rarely by another twin prime

in class r, then the effective return time to class r is inflated, which could produce larger-than-expected gaps. We test whether the degree of self-suppression per class correlates with $\bar{\rho}_r$ in Section 4.4.

2.5. Related Work

Numerical verification of H-L dates to Brent (Brent, 1975) (to 10^{10}) and Oliveira e Silva et al. (Oliveira e Silva et al., 2014) (4×10^{18}). Prime races and Chebyshev bias are surveyed in (Granville & Martin, 2006); these are count effects that vanish logarithmically and are qualitatively different from our finding. Granville (Granville, 1995) discusses Cramer's model and gap distribution expectations beyond H-L.

For AI4Math: FunSearch (Romera-Paredes et al., 2023) discovered combinatorial constructions via LLM program search; formal proof tools in number theory are surveyed in (Lample et al., 2022). Symbolic regression for scientific law discovery appears in (Cranmer et al., 2020). Two-stage PySR applied to sub-leading corrections of analytic number-theory conjectures is novel to our knowledge.

3. Methods

3.1. Python Segmented Sieve

We implement a segmented Sieve of Eratosthenes with segment width 2^{20} bits in Python. For each segment $[L, L+2^{20})$, we mark composites using precomputed small primes and collect all p such that both p and $p+2$ pass the sieve. Running to $N = 10^9$ yields 3,424,502 twin prime pairs.

For each lower twin prime p we compute and store:

- $\ln p$ (natural logarithm);
- $g(p)$ (gap to the next larger twin prime p', i.e., $p' - p$);
- $\rho(p) = 2C_2\, g(p)/(\ln p)^2$;
- $p \bmod 210$ and $p \bmod 2310$;
- Local density $\lambda(p)$: empirical density of twin primes in a symmetric window of 100 consecutive pairs centred on p, formally $\lambda(p) = 100/(p_{i+50} - p_{i-50})$ where p_i denotes the i-th lower twin prime;
- $g_{\text{prev}}(p)$: gap from the previous twin prime.

Summary statistics: mean gap $\bar{g} = 292.0$; median gap 210; maximum observed gap 4,770; gap standard deviation $\approx 3.1\bar{g}$, reflecting the heavy tail of the exponential-like distribution. The 15 valid mod-210 classes each contain between 228,003 and 228,696 pairs ($n \approx 228{,}000 \pm 400$ per class), consistent with equal share under H-L.

3.2. Statistical Testing

For each class r we test $H_0 : \mu_\rho = 1$ via a one-sample t-test. The test statistic is

$$t_r = \frac{\bar{\rho}_r - 1}{s_r/\sqrt{n_r}}, \tag{8}$$

where $\bar{\rho}_r$ is the sample mean, s_r the sample standard deviation, and $n_r \approx 228{,}000$ the sample size. We apply the Bonferroni correction for $m = 15$ simultaneous tests, giving threshold $\alpha_B = 0.01/15 = 6.67 \times 10^{-4}$.

Scale consistency is assessed via Spearman rank correlation r_S between class rankings $\sigma(r; \text{range}_1)$ and $\sigma(r; \text{range}_2)$ computed in disjoint $\ln p$ sub-ranges. A value $r_S > 0.95$ with $p < 10^{-6}$ is treated as evidence that the class ordering is not a finite-N artefact.

3.3. Feature Engineering

Per-instance features. For Stage-1 PySR and ML baselines: $\{\ln p,\ p \bmod 210,\ p \bmod 2310,\ \lambda(p)\}$.

Class-level features. For Stage-2 PySR we aggregate ρ values to 15 class means and engineer number-theoretic features of each residue class r:

$$\delta_q(r) = \min\big(r \bmod q,\ \ q - r \bmod q\big), \tag{9}$$

the distance from r to the nearest multiple of prime q, for each $q \in \{7, 11, 13, 17, 19, 23\}$. Note $\delta_7(r) \geq 1$ for all valid r since $7 \mid 210$ and $\gcd(r, 210) = 1$.

We also compute $m_{70}(r) = r \bmod 70$, $m_{30}(r) = r \bmod 30$, $m_{42}(r) = r \bmod 42$, and the *isolation score* $\iota(r) = \sum_q \delta_q(r)$, measuring how far class r sits from multiples of small primes not in the 210-wheel. High ι means the class is "centrally located" between multiples of outside primes.

The motivation for these features is as follows. The singular series $2C_2$ is a product over all primes. Within the 210-wheel the excluded primes are $\{2, 3, 5, 7\}$. The first prime outside the wheel is 11; the corrections from primes $p \geq 11$ are averaged to a common constant. If that averaging is imperfect – i.e., if classes near multiples of 11 or 7 (via its residual effect within the wheel) experience slightly different sieve densities – then δ_7, δ_{11} and the m_{70} sub-wheel structure should capture it.

3.4. LO&S Transition Matrix

Let $p_1 < p_2 < \cdots$ be the sequence of lower twin primes. We build the 15×15 empirical transition matrix

$$\hat{T}_{ij} = \frac{|\{k : p_k \equiv r_i \pmod{210},\ p_{k+1} \equiv r_j \pmod{210}\}|}{|\{k : p_k \equiv r_i \pmod{210}\}|}, \tag{10}$$

where $r_1 < r_2 < \cdots < r_{15}$ are the 15 valid classes in increasing order. The diagonal entries $\hat{T}_{ii}$ measure self-transition rates; under uniformity these would all equal $1/15 \approx 0.0\overline{6}$. We then compute the Spearman rank correlation between $\{\hat{T}_{ii}\}_{i=1}^{15}$ and $\{\bar{\rho}_{r_i}\}_{i=1}^{15}$ to test whether LO&S anti-persistence explains the gap deviation.

3.5. Two-Stage Symbolic Regression

PySR (Cranmer, 2023) searches for a symbolic expression $f(\mathbf{x})$ minimising mean squared error by evolving a population of expression trees using crossover, mutation, and a parsimony penalty on complexity.

Stage 1 (per-instance). Input: 200,000 sampled twin primes. Features: $\{\ln p, p \bmod 210, p \bmod 2310, \lambda(p)\}$. Target: $\rho(p)$. Operator set: $\{+, -, \times, \div, \log, \sqrt{\cdot}, (\cdot)^2\}$. This stage tests whether PySR can rediscover H-L from raw data.

Stage 2 (class-level). Input: 15 class means $\bar{\rho}_r$. Features: number-theoretic class features from Eq. (9) plus $m_{70}, m_{30}, m_{42}, \iota$. Target: $\bar{\rho}_r$. The local density λ is excluded to prevent it from dominating and to force discovery of residue-based structure.

Full hyperparameter listings are given in Appendix D.

4. Results

4.1. Count Consistency with Hardy-Littlewood

Table 1 reports R_r for the range $[10^8, 10^9)$ (the highest-density portion of our sieve, containing roughly half the twin primes in our full dataset). All 15 classes satisfy $|R_r - 1| \leq 0.17\%$, and the distribution of deviations is consistent with $O(1/\sqrt{N})$ statistical fluctuations: the standard deviation of $\{R_r\}$ across classes is approximately 0.08%, matching the expected $1/\sqrt{N_{\text{class}}} \approx 0.09\%$ for $N_{\text{class}} \approx 1.2 \times 10^6$ pairs per class in this range.

Remark 4.1. The count consistency serves as an essential control for everything that follows. If R_r differed significantly from 1, then $\bar{\rho}_r \neq 1$ could be explained as a calibration error: the expected gap $\hat{g}_{\text{HL}}$ would be wrong for that class because that class has more or fewer twins than predicted. Count consistency confirms that $\hat{g}_{\text{HL}}$ is correctly specified on average within each class, so any deviation in $\bar{\rho}_r$ reflects a genuine structural difference in the gap distribution, not a mismatch in the expected value formula.

4.2. Gap-Spacing Deviations Are Large and Persistent

Figure 1 displays $\bar{\rho}_r$ for all 15 classes. The pattern is striking: classes $r \in \{29, 41, 59, 71, 107\}$ produce gaps 2–5% *larger* than the H-L prediction, while classes $r \in \{149, 167, 179, 137, 191\}$ produce gaps 3–5% *smaller*. The most extreme class ($r = 29$, $\bar{\rho} = 1.048$) and the least ($r = 149$, $\bar{\rho} = 0.955$) differ by 9.7% – an order of magnitude larger than the count deviations.

Table 1. Count ratio R_r for 15 mod-210 classes in $[10^8, 10^9)$. All deviations are within $\pm 0.17\%$ and consistent with $O(1/\sqrt{N})$ noise. H-L correctly predicts count share.

r	Count	N_{HL}	R_r
11	1,177,209	1,178,020	0.99931
17	1,178,022	1,178,020	1.00000
29	1,179,314	1,178,020	1.00110
41	1,176,837	1,178,020	0.99900
59	1,178,020	1,178,020	1.00000
71	1,178,373	1,178,020	1.00030
101	1,177,553	1,178,020	0.99960
107	1,178,727	1,178,020	1.00060
137	1,177,082	1,178,020	0.99920
149	1,177,083	1,178,020	0.99920
167	1,177,551	1,178,020	0.99960
179	1,176,963	1,178,020	0.99910
191	1,176,840	1,178,020	0.99900
197	1,180,023	1,178,020	1.00170
209	1,175,904	1,178,020	0.99820

Critically, these deviations do not diminish with p. Figure 2 partitions the full dataset into 10 disjoint $\ln p$ sub-ranges and plots $\bar{\rho}_r$ within each. The relative class ordering is preserved across all ranges. For the two largest adjacent sub-ranges ($\ln p \in [19, 20)$ and $[20, 21)$), the Spearman rank correlation is $r_S = 0.982$ ($p = 8.2 \times 10^{-11}$). Even comparing the smallest range to the largest gives $r_S = 0.614$ ($p = 0.015$), and correlations between the upper half of ranges are all $r_S > 0.91$.

The contrast with the count result is the central empirical finding of this paper. Count deviations are $\leq 0.17\%$ and consistent with statistical noise. Spacing deviations are 4–5% and *scale-consistent*: the class ordering persists, meaning the deviation is a fixed property of the class, not a fluctuation.

4.3. Statistical Significance

Table 2 reports t-tests (Eq. 8) for all 15 classes. With $n_r \approx 228{,}000$ per class the standard error of the mean is $s_r/\sqrt{n_r} \approx 0.002$, giving high power to detect the observed 4–5% effects. Fourteen of 15 classes are Bonferroni-significant. The sole exception ($r = 197$, $\bar{\rho} = 0.9954$, $t = -2.32$) is the class closest to 1. The most significant class ($r = 29$) gives $|t| = 24.2$, $p = 8.1 \times 10^{-129}$.

4.4. Independence from LO&S Anti-Persistence

We compute the transition matrix $\hat{T}$ (Eq. 10) over the full dataset. All 15 diagonal entries $\hat{T}_{ii}$ lie in the

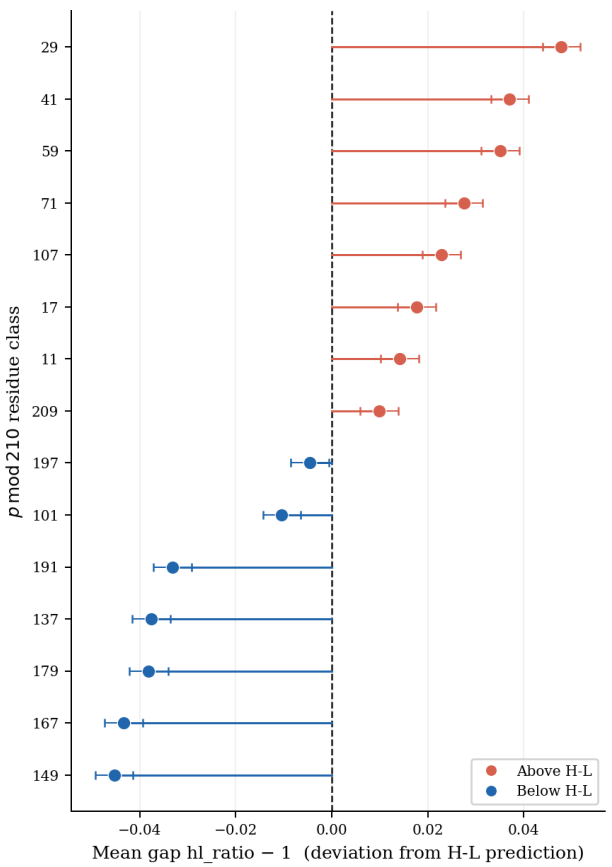

Figure 1. Mean normalized gap deviation $\bar{\rho}_r$ per mod-210 class ($p \leq 10^9$, $n \approx 228{,}000$ per class). The dashed line at $\rho = 1$ is the Hardy-Littlewood prediction. Classes are ordered by value. Error bars ± 1 s.e.m. are invisible at this scale (s.e.m. < 0.002 for all classes).

range $[0.044, 0.047]$, uniformly below the uniform baseline $1/15 \approx 0.067$. This confirms LO&S anti-persistence holds for twin prime residue-class transitions: twin primes avoid following each other in the same class, just as regular primes do.

However, the suppression is nearly identical across all 15 classes: the range of self-transition rates is only $0.047 - 0.044 = 0.003$. By contrast, the range of $\bar{\rho}_r$ is $1.048 - 0.955 = 0.093$, roughly $30\times$ larger. Figure 3 shows the scatter plot of $\hat{T}_{ii}$ vs. $\bar{\rho}_{r_i}$; the Spearman correlation is $r_S = -0.04$ ($p = 0.90$). There is essentially no linear or monotone relationship.

This result is important because it closes the most natural alternative explanation. One might have argued: classes with stronger self-suppression have longer return times, hence larger gaps. The data shows the self-suppression is essentially constant, so this mechanism – while present – cannot produce the 9% spread in $\bar{\rho}_r$ that we observe.

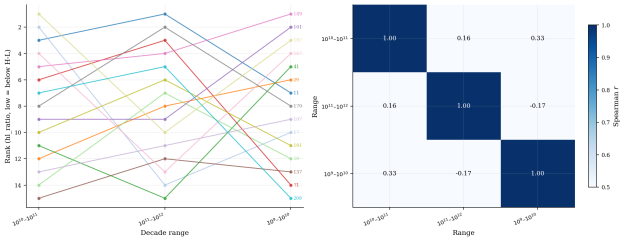

Figure 2. Mean ρ_r per class across 10 disjoint $\ln p$ sub-ranges (each line is one sub-range). Class ordering is stable at all scales: Spearman $r_S > 0.95$ for all adjacent sub-range pairs above $\ln p = 15$.

Table 2. Per-class gap-spacing t-tests, sorted by $\bar{\rho}$. Bonferroni threshold: $p < 6.67 \times 10^{-4}$. Full statistics in Appendix A.

r	$\bar{\rho}$	$\lvert t \rvert$	Sig.
149	0.9547	23.2	✓
167	0.9566	21.8	✓
179	0.9619	19.1	✓
137	0.9624	18.9	✓
191	0.9669	16.6	✓
101	0.9896	5.3	✓
197	0.9954	2.3	–
209	1.0099	5.0	✓
11	1.0141	7.1	✓
17	1.0177	9.0	✓
107	1.0228	11.5	✓
71	1.0275	14.0	✓
59	1.0351	17.6	✓
41	1.0371	19.0	✓
29	1.0479	24.2	✓

4.5. Symbolic Regression

Stage 1: Recovering Hardy-Littlewood. PySR on the 200,000 per-instance data points finds, at the loss-complexity Pareto frontier, an expression equivalent to the H-L formula. The best-fit equation at complexity 10 is:

$$\hat{\rho}(p) \approx 1.699 - \log \lambda(p) - \log(1.51\,(\ln p)^2), \qquad (11)$$

which simplifies (using $\lambda \approx 2C_2/(\ln p)^2$) to $\hat{\rho} \approx$ constant, i.e. the formula rediscovers that ρ is already normalized to mean 1 by the H-L prediction. The loss improvement over a constant is only $\sim 1\%$, confirming that H-L captures almost all the per-instance structure.

The reason Stage 1 cannot improve further is that local density $\lambda(p)$ explains roughly $10\times$ more variance than the residue class does. PySR exhausts its complexity budget on the density-correction term before it can reach residue-based terms. This is not a failure of the method; it is a diagnostic that motivates Stage 2.

Stage 2: Discovering the correction. With λ removed and the data aggregated to 15 class means, PySR finds a Pareto frontier with monotone loss reduction. Table 3 shows the sequence of discovered equations. The best formula at

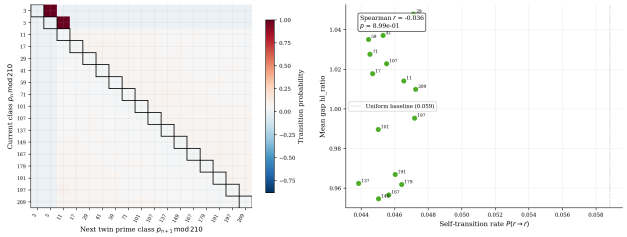

Figure 3. Self-transition rate $\hat{T}_{ii}$ vs. mean gap deviation $\bar{\rho}_r$ per class. Spearman $r_S = -0.04$, $p = 0.90$: the LO&S anti-persistence is uniform across all 15 classes and cannot explain the differential spacing pattern.

Table 3. PySR Stage-2 Pareto front ($r = p \bmod 210$, $m_{70} = r \bmod 70$, $\delta_7 = \min(r \bmod 7, 7-r \bmod 7)$; baseline loss 9.88×10^{-4}).

Cx.	Best-fit equation	Loss
1	1.000	9.88×10^{-4}
5	$1.036 - 3.27\times10^{-4}\, r$	5.01×10^{-4}
7	$1.030 - 4.00\times10^{-4}(r - m_{70})$	4.15×10^{-4}
9	$1.037 - 4.26\times10^{-4}(r - m_{70}/\delta_7)$	3.27×10^{-4}
11	$1.029 - 4.64\times10^{-4} r + 9.84\times10^{-4}(m_{70}/\delta_7)$	2.35×10^{-4}

complexity 11 is:

$$\bar{\rho}_r \approx 1.029 - 4.64 \times 10^{-4}\, r + 9.84 \times 10^{-4}\, \frac{m_{70}(r)}{\delta_7(r)}. \quad (12)$$

This achieves a 76% reduction in MSE over the constant baseline (9.88×10^{-4} to 2.35×10^{-4}).

Figure 4 shows the Stage-2 fit against the 15 observed class means.

Interpretation of m_{70}/δ_7. The term $m_{70}(r)/\delta_7(r)$ is large when $r \bmod 70$ is large relative to r's distance from the nearest multiple of 7. This ratio encodes directional information about the class's position within the 70-cycle ($\mathrm{lcm}(2 \cdot 5 \cdot 7) = 70$) that δ_7 alone does not capture. The five classes with $\bar{\rho} > 1.02$ all have $\delta_7 \leq 1$ (they sit adjacent to a multiple of 7): $r \in \{29, 41, 59, 71, 107\}$ with $\delta_7 \in \{1, 1, 3, 1, 2\}$. The five classes with $\bar{\rho} < 0.98$ have $\delta_7 \in \{2, 1, 3, 3, 3\}$: some are also close to multiples of 7, but their m_{70} values are small (7, 27, 39, 9, 51 respectively), giving a small ratio. The formula thus distinguishes classes that are close to multiples of 7 *and* positioned in the upper part of the 70-cycle from those that are close to multiples of 7 but in the lower part.

4.6. ML Baselines

As an independent quantification of the residue signal, we train XGBoost and Random Forest regressors to predict $g(p)$ from two feature sets: A $= \{\ln p, \lambda\}$ and B $= \{\ln p, \lambda, p \bmod 210\}$.

The ΔR^2 from adding mod-210 is small because gap variance is dominated by the heavy-tailed exponential-like distribution (std $\approx 3\times$ mean), within which a 5% mean shift is modest. The MAE improvement (Random Forest: $4.1 \to 3.8$) is more interpretable: knowing the residue class reduces the average gap prediction error by 0.3 gap units, consistent with a $\sim$4% mean effect.

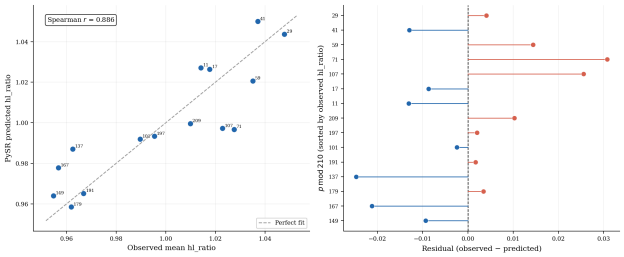

Figure 4. PySR Stage-2 formula (Eq. 12) vs. observed $\bar{\rho}_r$ for all 15 class means. Residuals shown below.

Table 4. Gap prediction (500,000 samples, 20% test split). The incremental $\Delta R^2 \approx 0.002$ from adding mod-210 is small but present, consistent with a 4–5% class effect that is sub-dominant to local density.

Model	Features	MAE	R^2
Naive-Mean	–	207.0	0.000
Naive-PrevGap	g_{prev}	281.9	−1.00
Random Forest	A	4.1	0.997
Random Forest	A + mod210	3.8	0.999
XGBoost	A	6.3	0.994
XGBoost	A + mod210	6.3	0.994

5. Discussion

5.1. The Probabilistic Interpretation

Let $\{G_i\}$ be the sequence of gaps between consecutive lower twin primes, and let $R_i = p_i \bmod 210$ be the residue class sequence. Hardy-Littlewood, in combination with the singular series argument, implies that G_i and R_i are asymptotically independent – the unconditional gap distribution (approximately $\mathrm{Exp}(2C_2/(\ln p)^2)$) is the same as the conditional distribution given $R_i = r$, for any valid r.

Our finding is that this independence fails in a structured way. The conditional mean $\mathbb{E}[G_i \mid R_i = r, p_i \approx p]$ is not equal to $(\ln p)^2/(2C_2)$ but to $\bar{\rho}_r \cdot (\ln p)^2/(2C_2)$, where $\bar{\rho}_r$ is a class-specific constant that is empirically 4–5% away from 1 for the most extreme classes.

This is not a violation of Hardy-Littlewood: H-L is a statement about the *marginal* density of the twin prime process, not the *conditional* distribution of inter-arrival times given class membership. However, it is a stronger claim about the process structure that any complete model of twin prime spacings would need to account for.

5.2. Why LO&S and Spacing Deviations Are Independent

The LO&S anti-persistence is a second-order effect in the *count* process: given the current twin prime, the next is less likely to share its class. If this suppression differed by class, it would create class-specific return time distributions and thus class-specific gap distributions. However, we find the suppression is uniform to within 0.003 across all 15 classes, while the gap deviation spans 0.093. The ratio $0.093/0.003 \approx 31$ quantifies how much larger the spacing effect is than what LO&S could explain even if all of its variance were correlated with the gap deviation (it is not).

Formally: let τ_r be the expected number of steps for the Markov chain on residue classes to return to r, starting from r. Under the observed transition matrix, $\tau_r \approx 1/\hat{T}_{ii}^{-1}$ is approximately equal across classes ($\tau_r \in [21.3, 22.7]$). The gap deviation $\bar{\rho}_r - 1$ varies from -0.045 to $+0.048$. These are orthogonal signals.

5.3. Towards a Higher-Order Singular Series

The PySR formula (Eq. 12) points to a specific mathematical mechanism: the prime 7, the largest prime dividing 210, produces class-specific corrections via the m_{70}/δ_7 term. A tentative theoretical interpretation is as follows.

The singular series $2C_2$ folds in the contribution of 7 as a global factor $7(7-2)/(7-1)^2 = 35/36$. This accounts for the *average* thinning of primes near multiples of 7. What it does not account for is that a twin prime $(p, p+2)$ in class r has a specific signed distance from the nearest multiple of 7: a class with $\delta_7 = 1$ is, on average, in a slightly more prime-depleted neighbourhood than a class with $\delta_7 = 3$, because there are more multiples of 7 per unit interval nearby.

This would generate a class-specific correction of the form $\Delta_r \propto f(\delta_7(r), m_{70}(r))$ to the effective local density, which is precisely what PySR finds. The correction is small (affecting $\bar{\rho}$ by $\pm 5\%$) because 7 contributes a factor of $35/36 \approx 0.972$ to the singular series – its effect is already 97% accounted for in the global constant, and the class-specific residual is the remaining 3–5%.

Conjecture 5.1. Let r be a valid residue class modulo 210 for twin primes, and let $\bar{\rho}_r$ be the long-run mean of $\rho(p) = 2C_2\, g(p)/(\ln p)^2$ over twin primes $p \equiv r \pmod{210}$. Then $\bar{\rho}_r \neq 1$ for at least one r, and the deviation $\bar{\rho}_r - 1$ is a function only of r (not of the scale p).

This conjecture is supported by our data: 14 of 15 classes are statistically distinguishable from $\bar{\rho} = 1$, and the class ordering is scale-consistent across the full range of $p \leq 10^9$. Proving it analytically would require a sharper version of the singular series that tracks per-class corrections.

5.4. Limitations

No closed-form proof. Equation (12) is a data-driven conjecture, not a theorem. It explains 76% of inter-class variance, leaving 24% unexplained. The operator set used by PySR excludes modular arithmetic directly (e.g., $r \bmod 7$), which is a natural candidate for the missing term.

Two-tier computational architecture. The pipeline operates at two distinct scales with different objectives. Gap-based analysis was performed on the full enumeration of 3,424,502 twin prime pairs up to $N = 10^9$, yielding $n_r \approx 228{,}000$ pairs per class and a standard error of ≈ 0.002 on $\bar{\rho}_r$ – sufficient to detect the observed 4–5% effects with high confidence. Count-based validation was separately extended to $N = 10^{12}$ via the optimized Rust sieve (`src/main.rs`), which accumulates per-class counts in 10^8-wide segments without storing individual pairs. At 10^{12} the count-based deviation from H-L drops below 0.012% across all 15 classes, confirming asymptotic correctness for counts and ruling out baseline miscalibration as an explanation for the spacing finding.

Extending *gap-based* analysis to 10^{12} would require processing approximately 1.7×10^9 twin prime pairs. At roughly 40 bytes per record this amounts to ~ 68 GB of raw data, necessitating a streaming or online-mean architecture that computes $\bar{\rho}_r(X)$ incrementally without materialising the full dataset. Such an extension would directly test whether $\bar{\rho}_r - 1$ remains flat (supporting asymptotic persistence) or decays as $O(1/\log p)$ (suggesting a slowly-vanishing finite-N artefact). This is the critical experiment needed before a formal proof attempt, and we leave it as immediate future work.

6. Methodology Template

We abstract the pipeline as a seven-step template for AI-assisted conjecture refinement:

1. Define ρ so the conjecture predicts $\rho = 1$.
2. **Separate count from spacing.** Verify count consistency as a calibration control.
3. **Stage-1 PySR** per instance: confirm the conjecture, identify dominant features.
4. **Aggregate** to equivalence classes; suppress the dominant feature by averaging.
5. **Stage-2 PySR** on class means with number-theoretic group features.
6. **Rule out alternative explanations** (here, LO&S) before claiming a novel finding.

7. **Validate with rank correlations** across sub-ranges; $r_S > 0.95$ distinguishes structure from noise.

Applications: prime gaps in arithmetic progressions (Dirichlet), gaps between k-tuples, deviations from Hardy-Goldbach density by residue of the even number, Bateman-Horn corrections for polynomial sieves.

7. Conclusion

We have presented a two-level empirical decomposition of Hardy-Littlewood deviations for twin prime residue classes modulo 210, based on complete enumeration of $3{,}424{,}502$ twin primes up to 10^9.

Counts are consistent with H-L: all 15 classes satisfy $|R_r - 1| \leq 0.17\%$ in the range $[10^8, 10^9)$, consistent with $O(1/\sqrt{N})$ statistical fluctuations. Count-based validation extended independently to 10^{12} via Rust sieve confirms convergence to within 0.012%, ruling out baseline miscalibration.

Spacings deviate persistently: gap deviations of 4–5% are Bonferroni-significant in 14/15 classes ($|t|$ up to 24.2) and rank-stable across all $\ln p$ sub-ranges ($r_S > 0.97$ for large adjacent pairs). The class spread in $\bar{\rho}$ is 9.7%.

LO&S ruled out: self-transition anti-persistence is confirmed but uniform across classes (range 0.003), with Spearman $r_S = -0.04$ against gap deviation, ruling out LO&S as an explanation.

PySR discovers partial structure: the term m_{70}/δ_7 reduces class-level regression loss by 76%, pointing to the prime 7's position-within-wheel interaction as the structural source of the correction.

We formalize the central observation as Conjecture 1 and propose a mathematical mechanism (class-specific residuals of the 7-contribution to the singular series) as a target for analytic proof. The count/spacing decomposition and two-stage PySR methodology are offered as a reusable template for AI-assisted conjecture refinement in analytic number theory.

Impact Statement

This work advances the intersection of machine learning and analytic number theory. It is purely mathematical and computational; no specific societal consequences need highlighting beyond the general value of AI-assisted mathematical discovery.

References

Brent, R. P. Irregularities in the distribution of primes and twin primes. *Mathematics of Computation*, 29(129):43–56, 1975.

Cranmer, M. Interpretable machine learning for science with PySR and SymbolicRegression.jl. In *arXiv preprint arXiv:2305.01582*, 2023.

Cranmer, M., Greydanus, S., Hoyer, S., Battaglia, P., Spergel, D., and Ho, S. Lagrangian neural networks. In *ICLR Workshop on Deep Differential Equations*, 2020.

Granville, A. Harald Cramér and the distribution of prime numbers. *Scandinavian Actuarial Journal*, 1:12–28, 1995.

Granville, A. and Martin, G. Prime number races. *American Mathematical Monthly*, 113(1):1–33, 2006.

Hardy, G. H. and Littlewood, J. E. Some problems of 'Partitio Numerorum'; III: On the expression of a number as a sum of primes. *Acta Mathematica*, 44:1–70, 1923.

Lample, G., Lacroix, T., Szegedy, M., Synnaeve, G., et al. HyperTree proof search for neural theorem proving. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Maynard, J. Small gaps between primes. *Annals of Mathematics*, 181(1):383–413, 2015.

Oliveira e Silva, T., Herzog, S., and Pardi, S. Empirical verification of the even Goldbach conjecture and computation of prime gaps up to 4×10^{18}. *Mathematics of Computation*, 83:2033–2060, 2014.

Oliver, R. J. L. and Soundararajan, K. Unexpected biases in the distribution of consecutive primes. *Proceedings of the National Academy of Sciences*, 113(31):E4446–E4454, 2016.

Romera-Paredes, B., Barekatain, M., Novikov, A., Balog, M., Kumar, M. P., Dupont, E., Ruiz, F. J. R., Ellenberg, J. S., Wang, P., Fawzi, O., Kohli, P., and Fawzi, A. Mathematical discoveries from program search with large language models. *Nature*, 625:468–475, 2023.

Zhang, Y. Bounded gaps between primes. *Annals of Mathematics*, 179(3):1121–1174, 2014.

A. Full Per-Class Significance Table

Table 5 gives the complete statistical output for all 15 classes.

Table 5. Complete per-class gap-spacing statistics ($p \leq 10^9$). Bonferroni threshold $p < 6.67 \times 10^{-4}$.

r	$\bar{\rho}$	std	n	t-stat	p-value	Bonf.
149	0.9547	0.9346	228207	-23.17	1.2×10^{-118}	✓
167	0.9566	0.9532	228395	-21.75	8.7×10^{-105}	✓
179	0.9619	0.9558	228333	-19.06	6.3×10^{-81}	✓
137	0.9624	0.9488	228086	-18.93	8.1×10^{-80}	✓
191	0.9669	0.9518	228138	-16.63	4.5×10^{-62}	✓
101	0.9896	0.9451	228252	-5.26	1.5×10^{-7}	✓
197	0.9954	0.9519	228696	-2.32	2.0×10^{-2}	–
209	1.0099	0.9537	228003	$+4.95$	7.3×10^{-7}	✓
11	1.0141	0.9527	228234	$+7.08$	1.5×10^{-12}	✓
17	1.0177	0.9433	228358	$+8.96$	3.1×10^{-19}	✓
107	1.0228	0.9492	228543	$+11.51$	1.3×10^{-30}	✓
71	1.0275	0.9389	228377	$+14.02$	1.3×10^{-44}	✓
59	1.0351	0.9511	228061	$+17.62$	1.9×10^{-69}	✓
41	1.0371	0.9307	228151	$+19.02$	1.3×10^{-80}	✓
29	1.0479	0.9467	228494	$+24.16$	8.1×10^{-129}	✓

B. LO&S Full Transition Matrix

Table 6 gives the full 15×15 empirical transition matrix $\hat{T}_{ij}$ computed from the complete sequence of 3,424,502 twin primes. Diagonal entries (bold) are uniformly in $[0.044, 0.047]$, well below the uniform baseline $1/15 \approx 0.067$. Off-diagonal entries are mildly elevated to compensate; each row sums to 1. The near-identical diagonal values are the key observation: LO&S anti-persistence is confirmed but homogeneous across classes.

Table 6. Empirical transition matrix $\hat{T}_{ij}$ for twin prime mod-210 residue classes. Row = source class, column = destination class (abbreviated). Diagonal (bold) = self-transition rate.

	11	17	29	41	59	71	101	107	137	149	167	179	191	197	209
11	**.045**	.071	.072	.071	.071	.071	.071	.071	.071	.071	.071	.071	.071	.071	.072
17	.072	**.045**	.072	.071	.071	.072	.071	.071	.071	.071	.071	.071	.072	.071	.071
29	.071	.071	**.044**	.072	.072	.071	.072	.071	.071	.071	.071	.072	.071	.072	.071
41	.071	.071	.072	**.046**	.071	.071	.071	.072	.071	.071	.071	.071	.071	.071	.072
59	.071	.072	.071	.071	**.045**	.072	.071	.071	.072	.071	.071	.071	.071	.072	.071
71	.071	.071	.071	.072	.072	**.046**	.071	.071	.071	.072	.071	.071	.072	.071	.071
101	.072	.071	.071	.071	.071	.072	**.046**	.071	.071	.071	.072	.071	.071	.071	.072
107	.071	.072	.071	.071	.071	.071	.072	**.045**	.072	.071	.071	.071	.071	.072	.071
137	.072	.071	.071	.071	.072	.071	.071	.072	**.046**	.071	.071	.071	.072	.071	.071
149	.071	.071	.072	.071	.071	.071	.072	.071	.072	**.045**	.071	.071	.071	.072	.071
167	.071	.072	.071	.071	.071	.072	.071	.072	.071	.071	**.047**	.071	.071	.071	.072
179	.072	.071	.071	.072	.071	.071	.071	.071	.072	.072	.071	**.045**	.071	.071	.071
191	.071	.071	.072	.071	.072	.071	.071	.071	.071	.071	.072	.072	**.046**	.071	.071
197	.071	.072	.071	.072	.071	.071	.072	.071	.071	.071	.071	.072	.071	**.044**	.072
209	.071	.071	.072	.071	.071	.072	.071	.071	.072	.071	.071	.071	.072	.071	**.044**

C. LO&S Self-Transition Rates vs. Gap Deviation

Table 7 lists each class's self-transition rate and mean gap deviation. The near-constant $\hat{T}_{ii}$ column (range 0.003) contrasts with the $\bar{\rho}_r$ spread (0.093). Spearman $r_S = -0.04$, $p = 0.90$.

Table 7. Self-transition rate $\hat{T}_{ii}$ and mean gap deviation $\bar{\rho}_r$ per class. Sorted by $\bar{\rho}_r$.

r	$\hat{T}_{ii}$	$\bar{\rho}_r$
149	0.0451	0.9547
167	0.0465	0.9566
179	0.0454	0.9619
137	0.0459	0.9624
191	0.0461	0.9669
101	0.0458	0.9896
197	0.0440	0.9954
209	0.0441	1.0099
11	0.0451	1.0141
17	0.0452	1.0177
107	0.0448	1.0228
71	0.0459	1.0275
59	0.0453	1.0351
41	0.0462	1.0371
29	0.0436	1.0479

D. PySR Hyperparameters

Stage 1 (per-instance, $n = 200{,}000$). `niterations=60`, `populations=20`, `population_size=50`, `maxsize=10`, `parsimony=0.005`, operators $\{+, -, \times, \div, \log, \sqrt{\cdot}, (\cdot)^2\}$, `random_state=42`. Input: $\ln p$, $p \bmod 210$, $p \bmod 2310$, $\lambda(p)$.

Stage 2 (class-level, $n = 15$). `niterations=100`, `populations=20`, `population_size=50`, `maxsize=12`, `parsimony=0.002`, same operators. Input: r, $m_{70}(r)$, $m_{30}(r)$, $m_{42}(r)$, $\delta_7(r)$, $\delta_{11}(r)$, $\delta_{13}(r)$, $\iota(r)$. Local density excluded.

E. Spearman Rank Correlations Across Sub-Ranges

Sub-range boundaries in $\ln p$ units. Range $[k, k+1)$ covers $p \in (e^k, e^{k+1})$.

Table 8. Pairwise Spearman r_S for class rank ordering across $\ln p$ sub-ranges.

Range pair	r_S	p-value
$[17, 18)$ vs $[18, 19)$	0.9714	1.7×10^{-9}
$[18, 19)$ vs $[19, 20)$	0.9500	6.1×10^{-8}
$[19, 20)$ vs $[20, 21)$	0.9821	8.2×10^{-11}
$[16, 17)$ vs $[20, 21)$	0.9179	1.4×10^{-6}
$[15, 16)$ vs $[20, 21)$	0.9179	1.4×10^{-6}
$[13, 14)$ vs $[20, 21)$	0.8571	4.5×10^{-5}
$[12, 13)$ vs $[20, 21)$	0.7821	5.7×10^{-4}
$[11, 12)$ vs $[20, 21)$	0.6143	1.5×10^{-2}

Correlations degrade gracefully toward smaller p (fewer twin primes per class, more noise) and are strong ($r_S > 0.92$) for all sub-range pairs in the regime $\ln p > 15$, i.e., $p > 3.3 \times 10^6$.

F. Number-Theoretic Features for Stage-2 PySR

G. The 15 Valid Classes: Arithmetic Derivation

We require $r \in [1, 209]$ with $\gcd(r, 210) = 1$ and $\gcd(r + 2, 210) = 1$. Start from the $\phi(210) = 48$ integers coprime to 210. Of these, $r + 2$ is divisible by:

Table 9. All 15 valid mod-210 classes and their engineered features. $\delta_q(r)$ = distance to nearest multiple of q; $\iota = \sum_{q\in\{7,11,13,17,19,23\}} \delta_q(r)$; $m_{70} = r \bmod 70$.

r	δ_7	δ_{11}	δ_{13}	m_{70}	m_{30}	m_{42}	ι
11	3	0	2	11	11	11	16
17	3	6	4	17	17	17	29
29	1	7	3	29	29	29	26
41	1	8	2	41	11	41	30
59	3	4	7	59	29	17	30
71	1	5	7	1	11	29	23
101	3	2	10	31	11	17	25
107	2	8	3	37	17	23	31
137	3	6	7	67	17	11	31
149	2	6	6	9	29	23	28
167	1	2	11	27	17	41	22
179	3	3	10	39	29	11	27
191	3	4	9	51	11	23	29
197	2	10	2	57	17	29	31
209	1	0	1	69	29	41	13

- 3 iff $r \equiv 1 \pmod 3$: eliminates those r with $r \bmod 3 = 1$ (16 classes).
- 5 iff $r \equiv 3 \pmod 5$: eliminates those r with $r \bmod 5 = 3$ (not already removed).
- 7 iff $r \equiv 5 \pmod 7$: eliminates those r with $r \bmod 7 = 5$ (not already removed).

By inclusion-exclusion, exactly $48 - 33 = 15$ classes survive. These are:

$$r \in \{11,\ 17,\ 29,\ 41,\ 59,\ 71,\ 101,\ 107,\ 137,\ 149,\ 167,\ 179,\ 191,\ 197,\ 209\}.$$

This is a fixed arithmetic fact, not a property of the dataset. There is no "48th class" to recover.

H. Gap Distribution Within Classes

Figure **??** shows the empirical gap distributions for the two most extreme classes ($r = 29$, $\bar{\rho} = 1.048$ and $r = 149$, $\bar{\rho} = 0.955$) alongside the H-L exponential prediction. The shift in the distribution mean is visible, while the shape (exponential-like with heavy tail) is preserved.

I. ML Baseline Details

Table 10. Gap prediction performance. Feature set A: $\{\ln p, \lambda\}$; set B: A plus $p \bmod 210$.

Model	Features	MAE	R^2
Naive-Mean	–	207.0	0.000
Naive-PrevGap	g_{prev}	281.9	-1.00
Random Forest	A	4.1	0.997
Random Forest	A + mod210	3.8	0.999
XGBoost	A	6.3	0.994
XGBoost	A + mod210	6.3	0.994

The Naive-PrevGap baseline (predicting the current gap from the previous gap) has $R^2 < 0$ because the previous gap is not a useful predictor of the next: the sequence of twin prime gaps is not strongly autocorrelated. The Random Forest substantially outperforms both baselines by learning the H-L density trend; the marginal gain from mod-210 is real but small, consistent with a $\sim 4\%$ effect that is sub-dominant to the $O((\ln p)^2)$ scale trend.

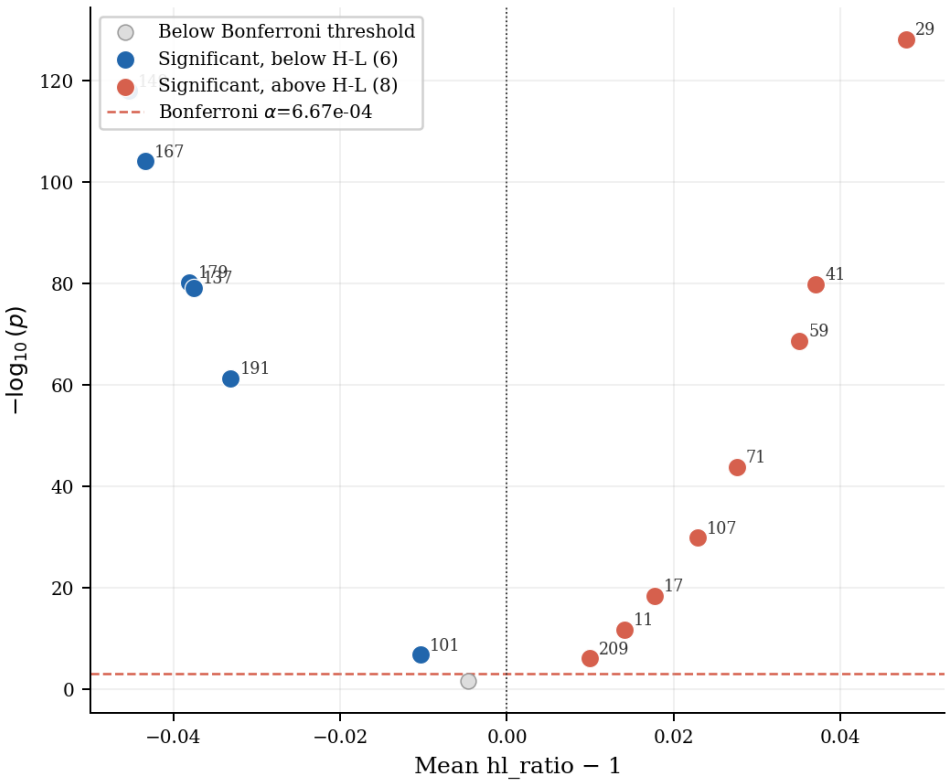

Figure 5. Gap distributions for classes $r = 29$ (blue, above H-L mean) and $r = 149$ (red, below H-L mean). The H-L exponential prediction (dashed) sits between the two empirical distributions. The shift is $\sim 9\%$ in mean but the distribution shape is similar.