Critical Comparison Report on Problem 1: Equivalence of the Φ^4_3 Measure Under Smooth Shifts

Automated `af` Proof Attempt vs. Official Solution

First Proof Project — Post-Mortem Analysis

February 2026

Abstract

This report critically compares the automated adversarial proof framework (`af`) attempt at Problem 1 of the First Proof benchmark with the official correct solution due to Martin Hairer. The problem asks whether the Φ^4_3 measure μ on $\mathcal{D}'(\mathbb{T}^3)$ is equivalent to its push-forward $T^*_\psi\mu$ under a smooth nonzero shift ψ. The `af` tool claimed the answer was **YES** (mutual absolute continuity), while the correct answer is **NO** (mutual singularity). This is a fundamental error at the level of the conjecture's truth value. We analyze what went right, what went wrong, and why; identify which nodes in the 84-node proof tree are fundamentally flawed; discuss which technical contributions remain valid despite the wrong answer; and draw lessons for future automated proof attempts.

Contents

1 The Problem

Let $\mathbb{T}^3$ be the three-dimensional unit torus, and let μ be the Φ^4_3 measure on $\mathcal{D}'(\mathbb{T}^3)$, the probability measure formally given by

$$d\mu(\phi) \;\propto\; \exp\Big(-\int_{\mathbb{T}^3}\Big(\tfrac{1}{2}|\nabla\phi|^2+\tfrac{\lambda}{4}\phi^4-\tfrac{C}{2}\phi^2\Big)\,dx\Big)\,\mathcal{D}\phi,$$

where C is a (diverging) renormalization constant and the expression is made rigorous via regularity structures [2], paracontrolled distributions [4, 3], or variational methods [5].

For a smooth function $\psi:\mathbb{T}^3\to\mathbb{R}$ with $\psi\not\equiv 0$, define the shift map $T_\psi(\phi)=\phi+\psi$ and the pushforward $(T_\psi^*\mu)(A)=\mu(T_\psi^{-1}(A))$.

Question (Hairer): Are μ and $T_\psi^*\mu$ equivalent (i.e., mutually absolutely continuous)?

This problem was posed as part of the First Proof benchmark [1], a collection of research-level mathematics problems designed to test the capabilities of AI-generated proofs.

2 The Official Correct Solution

2.1 The Answer: NO — Mutual Singularity

The correct answer, due to Hairer (simplified from joint work with Jacopo Peroni, based on [8]), is:

Theorem 2.1 (Hairer)**.** *For any smooth $\psi\not\equiv 0$ and any nonzero coupling constant, the measures μ and $T_\psi^*\mu$ are* ***mutually singular****.*

2.2 Context and Significance

The result fits into a broader picture distinguishing dimensions 2 and 3 in constructive QFT:

- In dimension 2, the Φ^4_2 measure μ is equivalent to the free field ν (Nelson, Glimm–Jaffe), and μ *is* quasi-invariant under smooth shifts.
- In dimension 3, Glimm [7] showed that $\mu\perp\nu$. The question probes whether μ retains the weaker quasi-invariance property. It does not.
- The borderline dimension for $\mu\sim\nu$ is $8/3$; the borderline dimension for quasi-invariance under smooth shifts turns out to be 3 itself, and the Φ^4_3 measure falls on the singular side.

2.3 Key Ideas of the Proof

The proof constructs a "distinguishing event" B^γ that has full μ-measure but zero $T_\psi^*\mu$-measure, thereby witnessing mutual singularity. The argument proceeds as follows.

Step 1: Decomposition of the field. Using the da Prato–Debussche/paracontrolled ansatz, the Φ^4_3 field u is decomposed as

$$u = \dagger - \dot{\Upsilon} + v,$$

where $\dagger$ is the stationary Ornstein–Uhlenbeck (free field) process, $\dot{\Upsilon}$ is the stationary solution to $(\partial_t+1-\Delta)\dot{\Upsilon}=\dagger^{\diamond 2}$, and v is a remainder with improved regularity: $v\in C^{1-2\kappa}$ with a further paraproduct decomposition $v=-3((v-\dot{\Upsilon})\prec\dagger^{\diamond 2})+v^\sharp$ where $v^\sharp\in C^{1+4\kappa}$.

Step 2: The distinguishing event. Define the diverging constant $c_{N,2} := \mathbb{E}[\dagger_N^{\diamond 2} \cdot \dot{\Upsilon}_N]$ (where $\dagger_N = P_N \dagger$ is the frequency truncation), which satisfies

$$c_{N,2} \gtrsim \log N.$$

The distinguishing event is

$$B^\gamma := \Big\{ u \in \mathcal{D}' : \lim_{N\to\infty} (\log N)^{-\gamma} \langle H_3(P_N u; c_N) + 9 c_{N,2}\, P_N u,\ \psi \rangle = 0 \Big\},$$

for $\gamma > 1/2$.

Step 3: $\mu(B^\gamma) = 1$. Expanding $H_3(u_N; c_N)$ using the decomposition $u = \dagger - \dot{\Upsilon} + v$ and applying:

- Lemma 4.2: $(\log N)^{-\gamma} :\dagger_N^3: \to 0$ a.s. in $C^{-3/2}$ for $\gamma > 1/2$;
- Lemma 4.5: $:\dagger_N^2: v_N + 3c_{N,2}(v_N - \dot{\Upsilon}_N)$ and $:\dagger_N^2: \dot{\Upsilon}_N - 3c_{N,2}\dagger_N$ both converge to finite limits in $C^{-1-2\kappa}$;
- Standard product estimates for the remaining terms;

one shows that the expression in B^γ converges to zero a.s., so $\mu(B^\gamma) = 1$.

Step 4: $(T_\psi^* \mu)(B^\gamma) = 0$. For $u + \psi$ in place of u, the same expansion produces the additional term

$$9(\log N)^{-\gamma} c_{N,2} \langle \psi_N, \psi \rangle,$$

which, since $c_{N,2} \gtrsim \log N$ and $\langle \psi_N, \psi \rangle \to \|\psi\|^2 > 0$, *diverges* for $\gamma < 1$ (which is compatible with $\gamma > 1/2$). All other terms vanish as before. Hence $u + \psi \notin B^\gamma$ a.s., so $(T_\psi^* \mu)(B^\gamma) = 0$.

Conclusion. Since $\mu(B^\gamma) = 1$ and $(T_\psi^* \mu)(B^\gamma) = 0$, the measures are mutually singular. □

2.4 The Critical Mechanism

The proof hinges on the logarithmically divergent constant $c_{N,2} = \mathbb{E}[\dagger_N^{\diamond 2}\, \dot{\Upsilon}_N] \gtrsim \log N$, which is specific to dimension 3. This constant creates a detectable "fingerprint" in the renormalized Wick cubic of the field: after dividing by $(\log N)^\gamma$ for $1/2 < \gamma < 1$, the Wick cubic pairing vanishes for the unshifted field but diverges for the shifted field due to the $c_{N,2}\langle \psi, \psi \rangle$ contribution. In dimension 2, the analogous constant is bounded, so no such fingerprint exists and the measures are equivalent.

3 What the `af` Attempt Claimed

3.1 The Claimed Answer: YES

The adversarial proof framework claimed the answer was **YES**: μ and $T_\psi^* \mu$ are equivalent (mutually absolutely continuous). This was described as "a natural analogue of the Cameron–Martin theorem for the interacting Φ_3^4 measure."

3.2 Proof Strategy

The `af` attempt constructed an 84-node proof tree over four adversarial sessions, organized into four stages:

Stage A (Nodes 1.2, 1.3): Establish that the UV-regularized measures satisfy $T_\psi^* \mu_\varepsilon \sim \mu_\varepsilon$ via the classical Cameron–Martin theorem, and compute the regularized Radon–Nikodym derivative $R_\varepsilon = \exp(\Psi_\varepsilon)$.

Stage B (Nodes 1.4.1–1.4.4): Expand the interaction difference $V_\varepsilon(\phi+\psi) - V_\varepsilon(\phi)$ using Wick calculus, identifying the quartic, quadratic, and linear shift terms.

Stage C (Nodes 1.5.1–1.5.4): Decompose the exponent as $\Psi_\varepsilon = \Psi_\varepsilon^{\mathrm{ren}} + L_\varepsilon + K_\varepsilon$, where $\Psi_\varepsilon^{\mathrm{ren}}$ converges in $L^p(\mu)$, K_ε is a deterministic constant, and $L_\varepsilon = 2\delta m_\varepsilon^2 \langle \psi_\varepsilon, \phi_\varepsilon \rangle$ is a divergent linear tilt.

Stage D (Nodes 1.6–1.8): Pass to the limit: establish exponential integrability of $\exp(\Psi^{\mathrm{ren}})$, uniform integrability, and finally identify $T_\psi^* \mu = (\exp(\Psi^{\mathrm{ren}})/Z) \cdot \mu$.

3.3 Scale of the Attempt

The effort was substantial:

- 84 nodes in the proof tree;
- 21 nodes adversarially validated;
- 10 refutations identified and repaired across 4 sessions;
- 12 technical pitfalls catalogued;
- Multiple repair cycles for the hardest steps (4 repairs for exponential integrability alone).

4 Critical Comparison: What Went Wrong

4.1 The Fundamental Error: Wrong Answer

The `af` attempt concluded YES (equivalence) when the correct answer is NO (mutual singularity). This is not a minor technical error in an otherwise-correct proof—it is a failure at the most basic level of mathematical reasoning about the problem. Every node in the proof tree that works *toward establishing equivalence* is either:

(a) correct but insufficient (valid technical lemmas that do not imply the claimed conclusion), or

(b) fundamentally flawed in a way that is invisible when one "knows" the answer should be YES.

4.2 Why the Wrong Answer Was Believed

The official solution commentary [1] identifies a common LLM failure mode: "the LLM would take as a premise the (wrong!) statement that the Φ_3^4 measure is equivalent to the free field measure, from which it then correctly deduces the (incorrect) claim that the Φ_3^4 measure is quasi-invariant under smooth shifts."

The `af` attempt did not make this exact error—it explicitly noted $\mu \perp \mu_0$ as Pitfall #1. However, it fell into a subtler version of the same trap: it assumed that because the *renormalized exponent* Ψ^{ren} converges, the density $\exp(\Psi^{\text{ren}})/Z$ must give the Radon–Nikodym derivative. This amounts to assuming that the divergent linear term L_ε can be "absorbed" into the normalization—precisely the unjustified step that the official solution's Section 2.1 identifies as an "incorrect heuristic."

4.3 The Core Logical Gap: Node 1.6.4 (Passage to the Limit)

The entire proof hinges on Stage D: showing that $T_\psi^*\mu = (\exp(\Psi^{\text{ren}})/Z) \cdot \mu$. This is where the proof must fail, because the claimed identity is **false**—the two measures are mutually singular, so no Radon–Nikodym derivative exists.

The `af` attempt itself discovered (in Session 4, Refutation 10) that the most natural approach to this step—the tilted measure convergence $\mu_\varepsilon^L \to \mu$—fails catastrophically:

1. **Tightness failure:** The L_ε-tilted measures $\mu_\varepsilon^L = \exp(L_\varepsilon)\, d\mu_\varepsilon/Z$ have divergent means, breaking tightness in $C^{-1/2-\delta}$.

2. **Wrong limit identification:** Even if tight, the tilted potential generates Z_2-breaking cubic terms, producing a different limiting measure.

These findings are *correct* and *important*. What the `af` attempt failed to recognize is that they are not bugs to be repaired—they are **evidence that the answer is NO**. The divergence of L_ε is not a technical obstacle to be circumvented; it is the mechanism by which the measures become singular.

4.4 The Proposed Repair: Boué–Dupuis Variational Approach (Node 1.6.4.3.3)

After refuting the tilted measure approach, the `af` attempt proposed a repair via the Barashkov–Gubinelli variational framework: represent μ as the law of the terminal value $X_1^{u^*}$ of a controlled SDE, represent the shift ψ as a deterministic drift change h, and use Girsanov's theorem to compute the density.

This approach is **fundamentally flawed** because it attempts to establish an identity ($T_\psi^*\mu \ll \mu$ with an explicit density) that is false. Specifically:

- The claim that $E_P[M_h | X_1^{u^*} = \phi] = \exp(\Psi^{\text{ren}}(\phi))/Z$ is false. This conditional expectation does not define a well-defined density because $T_\psi^*\mu \perp \mu$.

- The Girsanov change of measure M_h for a deterministic drift is perfectly valid at the level of the underlying Wiener space. However, the induced change of measure on the *terminal value* $X_1^{u^*}$ (which has law μ) involves precisely the divergent renormalization terms that cause singularity.

- The claim that "L_ε disappears because it is a regularization artifact" is incorrect. The logarithmic divergence of $c_{N,2}$ (equivalently, δm_ε^2) is a genuine feature of the Φ_3^4 theory in dimension 3 that persists in the continuum limit—it is the very mechanism that produces mutual singularity.

4.5 The "Incorrect Heuristic" Connection

The official solution (Section 2.1) explicitly warns against the heuristic that the `af` attempt follows. The formal density of $T_\psi^*\mu$ with respect to μ involves terms proportional to $\Phi^3 - C\Phi$ and $\Phi^2 - c_1$ (Wick powers), and the "additional logarithmically divergent terms proportional to c_2" cause the density to diverge, "suggesting (correctly) that μ and $T_\psi^*\mu$ are mutually singular."

Hairer then identifies *two problems* with making this heuristic rigorous as a proof of singularity:

1. $\Phi^3 - 3c_1\Phi$ does not define a random distribution (covariance $\sim |x-y|^{-3}$, not integrable in $d = 3$);

2. Divergent log-densities do not necessarily imply the measures fail to converge.

The `af` attempt's error was to attempt the *converse* heuristic (the density converges to something finite) without addressing either of these subtleties. The official proof circumvents both problems by working with the *smeared* quantity $\langle H_3(P_N u; c_N) + 9c_{N,2}P_N u, \psi\rangle$ rather than pointwise densities.

5 Node-by-Node Assessment Given the Correct Answer

We now assess each major component of the `af` proof tree in light of the correct answer (mutual singularity).

5.1 Stages A–C: Correct and Useful

Nodes	Content	Assessment
1.2, 1.3	Regularized equivalence $T_\psi^*\mu_\varepsilon \sim \mu_\varepsilon$ via Cameron–Martin; RN derivative $R_\varepsilon = \exp(\Psi_\varepsilon)$	Correct
1.4.1–1.4.4	Wick expansion of $V_\varepsilon(\phi+\psi) - V_\varepsilon(\phi)$	Correct
1.5.1–1.5.4	Decomposition $\Psi_\varepsilon = \Psi^{\text{ren}} + L_\varepsilon + K_\varepsilon$; convergence of Ψ^{ren}	Correct

These stages perform valid algebraic and analytic computations. They are correct regardless of the final answer. Indeed, the official proof implicitly uses the same Wick expansion structure: the decomposition of $H_3(u_N; c_N)$ in the official proof is closely related to the Wick expansion in Stages B–C.

The identification of the divergent linear term $L_\varepsilon = 2\delta m_\varepsilon^2 \langle \psi_\varepsilon, \phi_\varepsilon\rangle$ with $\delta m_\varepsilon^2 \sim C\log(1/\varepsilon)$ is a correct and important observation. In the official proof, this divergence reappears as the logarithmic growth of $c_{N,2}$, which is the mechanism driving mutual singularity.

5.2 Node 1.6.1 (Wick Power Regularity): Correct

The statement that smeared Wick powers $\langle f, :\phi^k: \rangle$ are well-defined for $f \in C^\infty$ under μ is correct. The official proof uses this implicitly when pairing against ψ.

5.3 Node 1.6.2 (Exponential Integrability): Correct but Irrelevant

The claim that $E_\mu[\exp(\alpha\langle f, :\phi^k:\rangle)] < \infty$ for smooth f and $k \in \{1,2,3\}$ is **almost certainly correct** as a standalone statement—the exponential integrability of smeared Wick powers under μ is a natural consequence of the Barashkov–Gubinelli variational bounds. This was the hardest result in the `af` attempt, requiring four repair cycles.

However, this result is **irrelevant to the correct answer**. The exponential integrability of $\exp(\Psi^{\text{ren}})$ would be needed to define the candidate Radon–Nikodym derivative $\rho = \exp(\Psi^{\text{ren}})/Z$—but since $T_\psi^*\mu \perp \mu$, no such density exists, and the integrability of $\exp(\Psi^{\text{ren}})$ has no bearing on the question. The function $\exp(\Psi^{\text{ren}})$ may well be integrable, but it does not give the density of $T_\psi^*\mu$ with respect to μ.

5.4 Node 1.6.3 (Uniform Integrability): Correct but Irrelevant

Similarly, the uniform integrability of $\{\exp(\Psi_\varepsilon^{\text{ren}})\}$ under a Skorokhod coupling is a valid technical statement. The renormalized exponents $\Psi_\varepsilon^{\text{ren}}$ do converge (the renormalized part of the density is well-behaved). But UI of $\exp(\Psi_\varepsilon^{\text{ren}})$ does not imply that $R_\varepsilon = \exp(\Psi_\varepsilon^{\text{ren}} + L_\varepsilon + K_\varepsilon)$ converges to a density, because L_ε diverges.

5.5 Node 1.6.4 (Passage to the Limit): Fundamentally Flawed

This is where the proof necessarily fails. The goal—identifying $T_\psi^*\mu = (\exp(\Psi^{\text{ren}})/Z) \cdot \mu$—is a false statement. Every approach to establishing this identity must contain an error.

- **Node 1.6.4.3 (Tilted measure convergence):** Correctly refuted by the `af` adversary in Session 4. The refutation correctly identifies that tightness fails and the limit is wrong.
- **Node 1.6.4.3.3 (Boué–Dupuis repair):** Unverified and **necessarily flawed**. The conditional expectation identity $E[M_h | X_1^{u^*} = \phi] = \exp(\Psi^{\text{ren}}(\phi))/Z$ cannot hold μ-a.s. because it would imply $T_\psi^*\mu \ll \mu$, contradicting mutual singularity.
- **Nodes 1.6.4.4, 1.6.4.5 (Vitali convergence, assembly):** These reference the refuted tilted measure approach and are incoherent.

5.6 Nodes 1.7 and 1.8 (Strict Positivity and Symmetry): Vacuously Correct

Node 1.7 claims $\exp(\Psi^{\text{ren}})/Z > 0$, which is trivially true. Node 1.8 claims the argument applies with $-\psi$. Both are correct *as statements*, but they are steps in a proof of a false conclusion, so their role in the overall argument is moot.

5.7 Summary Table

Component	Correct?	Relevant?	Notes
Stages A–C (1.2–1.5)	Yes	Partially	Valid lemmas; useful for understanding the problem
Node 1.6.1	Yes	Yes	Used implicitly in the official proof
Node 1.6.2	Likely yes	No	Exponential integrability is a nice result but irrelevant
Node 1.6.3	Likely yes	No	UI of $\exp(\Psi^{\text{ren}})$ is valid but irrelevant
Node 1.6.4	**No**	—	Must fail: attempts to prove a false statement
Node 1.6.4.3 refutation	Yes	Yes	Correctly identifies fundamental obstacles
Node 1.6.4.3.3	**No**	—	Must contain errors
Nodes 1.7, 1.8	Trivially	No	Steps in a proof of a false conclusion

6 What the `af` Attempt Got Right

Despite the wrong answer, the `af` attempt made several genuine contributions.

6.1 Correct Technical Lemmas

The Wick expansion (Stage B) and renormalization decomposition (Stage C) are correct and nontrivial. The identification of the three components Ψ^{ren}, L_ε, and K_ε, and the convergence of Ψ^{ren} in $L^p(\mu)$, are valid results that illuminate the structure of the problem.

6.2 Correct Identification of the Central Difficulty

The `af` attempt correctly identified the divergent linear term $L_\varepsilon = 2\delta m_\varepsilon^2 \langle \psi_\varepsilon, \phi_\varepsilon \rangle$ as the central obstacle. This is essentially the same divergence ($\delta m_\varepsilon^2 \sim c_{N,2} \sim \log N$) that drives mutual singularity in the official proof. The `af` attempt understood the *source* of the difficulty; it simply drew the wrong conclusion about what the difficulty implies.

6.3 Correct Self-Refutation of the Tilted Measure Approach

Refutation 10 (Session 4) is a genuinely valuable finding. The adversarial verification correctly showed:

- Exponential tilts with divergent coefficients break tightness (mean shifts by $\sim \beta^{1/3}$);
- Φ_3^4 uniqueness is about regularization independence, not potential independence;
- Z_2-breaking perturbations change the limiting measure.

These observations are mathematically correct and are closely related to why the answer is NO. The `af` system found evidence *against* its own claimed answer but failed to update the answer accordingly.

6.4 Extensive Catalogue of Pitfalls

The 12 pitfalls identified across 4 sessions constitute a useful reference for anyone working on Φ_3^4 problems. In particular:

- Pitfall #1 ($\mu \perp \mu_0$) is fundamental context.
- Pitfall #4 (the L^p route to UI is blocked: $E_{\mu_\varepsilon}[R_\varepsilon^p]$ diverges for $p > 1$) is directly related to the divergence mechanism that causes singularity.
- Pitfalls #11 and #12 (divergent tilts break tightness; uniqueness $\neq$ potential equivalence) are essentially correct observations about *why the answer is NO*, even though they were framed as obstacles to overcome rather than evidence of mutual singularity.

6.5 Adversarial Process as Error Detection

The adversarial framework successfully detected 10 errors across 4 sessions. This is impressive error-detection capability. The problem was not in detecting errors but in correctly interpreting their significance: the system treated every error as a *repairable flaw in a correct proof* rather than considering whether the accumulation of obstacles might indicate a wrong answer.

7 Root Cause Analysis: Why the Answer Was Wrong

7.1 Confirmation Bias in the Proof Search

The `af` system committed to the answer YES at the outset and never revisited this commitment, even when faced with substantial evidence against it. This is a form of confirmation bias: every obstacle was treated as a technical challenge to overcome, never as evidence that the conjecture might be false.

In particular, the system's response to Refutation 10 is revealing. Having correctly shown that the tilted measure approach fails because of divergent linear tilts, the system proposed an alternative approach (Boué–Dupuis) rather than considering whether the failure might indicate mutual singularity. The correct inference from Refutation 10 is: "the divergent linear term L_ε cannot be absorbed into the limit, which suggests $T_\psi^* \mu$ and μ have genuinely different singularity structures, i.e., they are mutually singular."

7.2 Lack of "Disprove" Mode

The problem statement asks to "prove or disprove," but the `af` system appears to have operated exclusively in "prove" mode. A well-designed automated prover should:

1. Consider both YES and NO as initial hypotheses;
2. Seek evidence for and against each;
3. Attempt to construct a distinguishing event or function (for NO) as well as a density (for YES);
4. Update its belief about the answer based on accumulated evidence.

The `af` system did none of these. It never attempted to construct a set A with $\mu(A) = 1$ and $(T_\psi^*\mu)(A) = 0$, which is the natural approach for proving singularity.

7.3 The Heuristic Trap

The official solution warns that a "tempting heuristic" suggests mutual singularity, and identifies two problems with making it rigorous. The `af` system fell into the *opposite* heuristic trap: it assumed that because Ψ^{ren} has nice convergence properties, the density $\exp(\Psi^{\mathrm{ren}})/Z$ should give the Radon–Nikodym derivative. But nice convergence of Ψ^{ren} does not overcome the divergence of L_ε—the two phenomena coexist, and it is L_ε that determines the answer.

7.4 Missing Domain Expertise

The correct proof requires deep knowledge of:

- The specific stochastic objects in regularity structures (the tree notation $\dagger$, $\vee$, $\checkmark$, $\dot{\Upsilon}$, etc.);
- The paracontrolled decomposition $u = \dagger - \dot{\Upsilon} + v$ and the improved regularity of v;
- The specific divergent constant $c_{N,2} = \mathbb{E}[\dagger_N^{\diamond 2}\, \dot{\Upsilon}_N]$ and its logarithmic growth;
- The technique of constructing distinguishing events via renormalized Wick polynomials.

The `af` system worked with the variational (Barashkov–Gubinelli) formulation rather than the regularity-structures/paracontrolled formulation. While the BG framework is powerful for constructing μ and establishing integrability bounds, the official proof of mutual singularity requires the more detailed stochastic analysis provided by the paracontrolled decomposition. This suggests that the choice of mathematical framework can bias the prover toward certain conclusions.

8 Lessons Learned

8.1 For Automated Theorem Proving

1. **Always consider both truth values.** For prove-or-disprove problems, the prover should maintain parallel proof attempts for both YES and NO, and allocate resources based on evidence. The accumulation of obstacles (10 refutations, 4 repair cycles for a single node) should trigger reconsideration of the answer.
2. **Interpret refutations as evidence, not just bugs.** When a natural approach to proving a statement fails for deep mathematical reasons (as opposed to technical errors), this is evidence against the statement. The `af` system's Refutation 10 essentially contains the key insight needed for the correct proof (the divergent linear term cannot be handled), but this insight was wasted because it was framed as a bug rather than a clue.

3. **Attempted repairs that keep failing may signal a wrong conjecture.** The pattern of repeated refutation and repair (4 cycles for node 1.6.2, followed by refutation of node 1.6.4.3 and an unverified repair in 1.6.4.3.3) is a strong signal. A well-calibrated system should lower its confidence in the conjecture with each successive failure, especially when the failures have deep structural causes.

4. **Framework choice matters.** The BG variational framework naturally suggests equivalence (it provides good control over the measure via stochastic control theory), while the paracontrolled/regularity-structures framework reveals the fine structure that distinguishes the shifted and unshifted measures. Automated provers should consider multiple mathematical frameworks when tackling open problems.

5. **Do not fabricate or misapply results to fill gaps.** Refutation 9 (fabricated BG propositions) and Refutation 10 (misapplication of Φ^4_3 uniqueness) illustrate a dangerous failure mode: when a proof step is hard, the system may "hallucinate" a result that fills the gap. Rigorous verification must catch such fabrications, and the system should treat the inability to find a valid argument as informative.

8.2 For the Specific Problem

1. The divergence $c_{N,2} \sim \log N$ (equivalently, $\delta m_\varepsilon^2 \sim \log(1/\varepsilon)$) is the *mechanism of mutual singularity*, not an obstacle to equivalence. Any proof attempt that tries to "absorb" or "cancel" this divergence is doomed.

2. The renormalized exponent Ψ^{ren} converges, and $\exp(\Psi^{\mathrm{ren}})$ may well be integrable. But this does not mean $\exp(\Psi^{\mathrm{ren}})/Z$ is the Radon–Nikodym derivative. The convergence of Ψ^{ren} and the divergence of L_ε are independent phenomena; the latter dominates.

3. The correct proof technique—constructing a distinguishing event from renormalized Wick polynomials—is fundamentally different from the density-based approach pursued by the `af` system. Singularity proofs require *separation* (finding a measurable set that distinguishes the two measures), not *computation* (finding a density function).

8.3 Broader Observations

1. The problem is at the frontier of constructive QFT and specifically designed to probe deep structural properties of the Φ^4_3 measure. It is unsurprising that an automated system would struggle, especially given that the official solution relies on very recent results (Hairer–Kusuoka–Nagoji 2024, Hairer–Peroni, in preparation).

2. The `af` adversarial process is effective at *local* error detection (finding flaws in specific proof steps) but not at *global* hypothesis evaluation (questioning whether the overall direction is correct). Incorporating a mechanism for global hypothesis testing—perhaps by periodically attempting the negation—would be a significant improvement.

3. The amount of valid technical work produced by the `af` system (Stages A–C, exponential integrability bounds, catalogue of pitfalls) is nontrivial, even though it does not resolve the problem. This suggests that automated tools can be useful as *research assistants* that produce correct intermediate results, even when they cannot navigate the high-level strategy of a proof.

9 Conclusion

The `af` automated proof attempt at Problem 1 produced a substantial body of work—84 nodes, 21 validated, 10 refutations detected and addressed—but arrived at the wrong answer. The measures μ and $T_\psi^*\mu$ are **mutually singular**, not equivalent.

The core error was a failure of hypothesis management: the system committed to YES at the outset and interpreted all subsequent difficulties as technical obstacles rather than evidence against the conjecture. Ironically, the system's own Refutation 10 contains the key insight (the divergent linear term L_ε cannot be absorbed) that, correctly interpreted, points toward mutual singularity.

The correct proof constructs a distinguishing event from the logarithmically divergent constant $c_{N,2}$, which creates a detectable "fingerprint" in the renormalized Wick cubic of the field. This fingerprint vanishes for the unshifted measure but persists (and diverges) for the shifted measure, witnessing mutual singularity. The `af` system never explored this direction because it was exclusively searching for a density, not a separating set.

Despite the wrong answer, the `af` attempt produced valid intermediate results (Wick expansion, renormalization decomposition, exponential integrability bounds) and correctly identified the central difficulty (the divergent mass counterterm). The adversarial process effectively detected local errors. The principal failure was global: the inability to update the hypothesis in light of accumulating evidence.

References

[1] M. Abouzaid, A. Blumberg, M. Hairer, N. Kileel, T. Kolda, J. Nelson, A. Spielman, N. Srivastava, R. Ward, S. Weinberger, and L. Williams, *First Proof: Solutions and Comments*, February 2026.

[2] M. Hairer, *A theory of regularity structures*, Invent. Math. **198** (2014), 269–504.

[3] M. Gubinelli, P. Imkeller, and N. Perkowski, *Paracontrolled distributions and singular PDEs*, Forum Math. Pi **3** (2015), e6, 75pp.

[4] R. Catellier and K. Chouk, *Paracontrolled distributions and the 3-dimensional stochastic quantization equation*, Ann. Probab. **46** (2018), 2621–2679.

[5] N. Barashkov and M. Gubinelli, *A variational method for* Φ^4_3, Duke Math. J. **169** (2020), 3339–3415.

[6] N. Barashkov and M. Gubinelli, *The* Φ^4_3 *measure via Girsanov's theorem*, Electron. J. Probab. **26** (2021), 1–29.

[7] J. Glimm, *Boson fields with the* $:\Phi^4:$ *interaction in three dimensions*, Comm. Math. Phys. **10** (1968), 1–47.

[8] M. Hairer, S. Kusuoka, and H. Nagoji, *Singularity of solutions to singular SPDEs*, arXiv preprint, 2024.

[9] M. Hairer and K. Matetski, *Discretisations of rough stochastic PDEs*, Ann. Probab. **46** (2018), 1651–1709.

[10] M. Gubinelli and M. Hofmanová, *Global solutions to elliptic and parabolic* Φ^4 *models in Euclidean space*, Comm. Math. Phys. **368** (2019), 1201–1266.