検証・評価レポート

Kumazawa et al., PAAMS 2023

“Enhancing Safety Checking Coverage with Multi-swarm Particle Swarm Optimization”

独立再実装による再現性・主張検証

2026 年 6 月 12 日

論文の手法（`msPSO-safe-inc` と被覆メトリクス AC_{div} / AC_{int}）を純粋 Python（標準ライブラリのみ）で**ゼロから再実装**し、ランダムグラフベンチマーク上で再現実験を行った。本書は再現可能性と各主張の成否、および「再現しない主張」の根本原因をまとめる。実装・データ一式: https://github.com/yaskodama/kuma2023-repro

1 何を実装したか（論文との対応）

論文の構成要素	実装
オートマトン理論的安全性検査・交差オートマトン（2.1 節）	`benchmark.py`: p 個の ER プロセス＋仕様オートマトン（受理＝デッドロック誤り状態）のインターリーブ積
PSO 位置 → パス符号化・目的関数 式 (3)	`mspso.py: decode, fitness`
動的マルチスウォーム（状態抽象による分割, 2.2 節）	`mspso.py`: 抽象状態（仕様状態）ごとにスウォームを動的生成
`msPSO-safe-inc`（アルゴリズム 1）	`mspso.py: run_incremental`
更新戦略 APUP / APP（3.1 節）	`rdistance` 増加 / `penalty` 付加
AC_{div}, AC_{int}（3.2 節）	iteration ごとに累積計測
ランダムグラフベンチマーク（4.1 節）	ER$(100, 0.05)$、$p = 2 \ldots 6$、各 3 ケース

■論文が明示していない点に置いた仮定（「妥当な仮定で補完」）

- プロセスの合成は**インターリーブ＋仕様モニタの同期**（標準的な安全性検査の積）。
- $firstvisit(x)$ = (パス上の相異なる抽象状態数)$/pathlen$、よって $pathlen \cdot firstvisit$ = パス上の相異なる抽象状態数（「抽象状態の複数回到達を罰する」記述に整合）。
- 位置成分 $x[k]$ は $\lfloor x[k] \rfloor$ mod (後続数) で k 番目の遷移を選択。
- 反例長は「スウォーム開始状態の初期状態からの深さ＋部分パス内オフセット」で計算し、**初期状態からの真のパス長**になるよう補正した。

■実装中に見つけた落とし穴　最初の実装では反例長を各スウォームの代表開始状態から測っており、長さ 1 の偽反例を生んでいた。安全性検査の反例は**初期状態から**の到達経路でなければならない。代表状態の発見深さを追跡して補正した（`_run_and_capture`）。

2　主張ごとの検証結果

#	論文の主張	再現結果	判定
1	実行回数とともに**多様化（被覆）が進む**	累積 AC_{div} が単調増加（visited も単調増加）	✓ 再現
2	多様性は**早期急増し後に飽和**、強度は**単調増加**	多様性ゲイン iters 1 → 10 で +0.264、150 → 200 で +0.004。強度は全 45 run で単調	✓ 再現
3	早期は多様化、後期は強化が支配的	図の Diversity（早期急増 → 平坦）と Intensity（ほぼ線形増加）で確認	✓ 再現
4	APP/APUP は単純再初期化と**同等以上**	8 seed 平均 AUC: none 0.7363, APP 0.7361, APUP 0.7364。差は 4 桁目	△ 部分不一致
5	APP と APUP の優劣は判別困難	AUC で APP 7 勝 / APUP 6 勝 / 2 分	✓ 再現
6	AC_{int} は visited 数にほぼ比例	Pearson 相関 1.000	✓ 再現 *
7	増分化は単一実行より**短い反例**を検出	反例長は単調非増加だが**厳密短縮は 2/45 のみ**	△ 弱く不一致

* ただし §4 の批判（平均化による自明な比例）を参照。

3　評価（査読的コメント）

■強み

- **手法は実装可能で内部整合的**。式 (3)・アルゴリズム 1・メトリクスは曖昧さはあるが再構成でき、定性的トレンド（被覆の早期急増 → 飽和、強度の単調増加）は頑健に再現した。
- ベンチマーク設計（特定問題への偏りを避けるランダムグラフ）は妥当。

■主要な弱点・疑問点

1. **提案戦略の有効性が誇張気味**（主張 4）。APP/APUP は単純再初期化と統計的に区別できない（8 seed 平均で AUC 差は 0.0003）。論文 Fig.1–3 でも各線は近接し、論文自身 “comparable or faster”・“difficult to compare”・p=5 case2 では baseline 最速、と認めている。要旨/結論の “effective as compared to simple random re-initialization” は**過大主張**に見える。
2. AC_{int} **の「visited 数に比例」はほぼ恒等式**（主張 6）。全抽象状態で平均すると $\sum_s AC_{int}(s)/|S_{sp}| =$ (総 visited)$/|S_{sp}|$ となり定義上 visited に厳密比例する。「興味深い観察」は平均化方法から自明に従う。

3. **「より短い反例」の主張（7）は本設定では実証しにくい**。反例が短い（長さ 1–5）と初回 run でほぼ最短に到達し増分短縮の余地がない。論文はベンチマーク難易度（最短反例長）と効果の関係を分析していない。
4. **再現性に響く記述不足**。スウォーム動的生成（[16] 依存）、初期状態の選び方、`firstvisit`/`numblocked` の定義、`stag_rounds`/`itupert`/`DL` の役割、位置 → パス符号化が論文単体では復元不能。これらは結果の挙動を大きく左右する。

4 追試: Table-1 予算のまま積空間を拡大（n=1000）

当初は Table-1 の per-run 予算（particle 20, generations 30）をそのまま使うと本再現の抽象空間（仕様 101 状態）が **1 run で飽和**したため予算を縮小していた。これは**抽象空間が小さすぎた**ことが原因なので、論文のグラフ構造（平均出次数 ≈ 2.5、ER(100, 0.05) 相当）を保ったまま**状態数を** n=1000 **に拡大**（$p_{\text{edge}} = 2 \cdot 2.475/(n-1)$）し、**Table-1 予算をそのまま**再実行した（45 通り、約 4 分）。

- **飽和は解消**。1 run の被覆は AC_{div}@1 ≈ 0.25–0.40。以後 @10 ≈ 0.6 → @50 ≈ 0.78 → end 0.26–0.89 と**早期急増 → 平坦**（多様性ゲイン iters 1 → 10 = +0.326、150 → 200 = +0.006）。強度は全 45 run で単調増加。論文 Fig.1–3 と定性的に一致。✓
- 反例長は 4–7 と論文の範囲に接近（全 45 run で反例検出、長さ単調非増加）。
- **しかし提案戦略の優位は依然として現れない**。3 seed 平均の多様性 AUC（被覆速度）は **none 0.6739 / APP 0.6737 / APUP 0.6739**（4 桁目のみの差）。

⇒ **予算縮小版でも、Table-1 予算のまま積空間を拡大した本追試でも結論は同じ**: 多様化・強度のトレンドは再現するが、**提案更新戦略が単純再初期化より有効という核心主張は再現しない**。主張 4 の頑健な反証である。

5 根本原因の診断（なぜ APP/APUP は効かないのか）

「効かない」を主張で終わらせず、結合強度を掃引して原因を特定した（Table-1 予算 + n=1000、4 インスタンス × 3 seed、baseline AUC = 0.7386）。

戦略	多様性 AUC	vs none
none (baseline)	0.7386	—
APUP（論文の増分）	0.7387	+0.0001
APP weight = 1（論文相当）	0.7383	−0.0003
APP weight = 5	0.7386	±0.0000
APP weight = 20	0.7386	±0.0000
APP weight = 100	0.7386	±0.0000
APP weight = **500**	0.7388	+0.0002

■原因 1: APUP は原理的に効かない（構造的上限） $rdistance$ は式 (3) で $\dfrac{1}{1 + pathlen + rdistance}$ という**寄与が高々 1 の項**にしか入らない。一方 $DL \cdot deadlock = 10^4$、$pathlen \cdot firstvisit \leq maxlen = 5$。よって $rdistance$ をいくら増やしても（論文の増分 $+|S_{sp}|$ でも）探索方向を変える力がない。**増分**

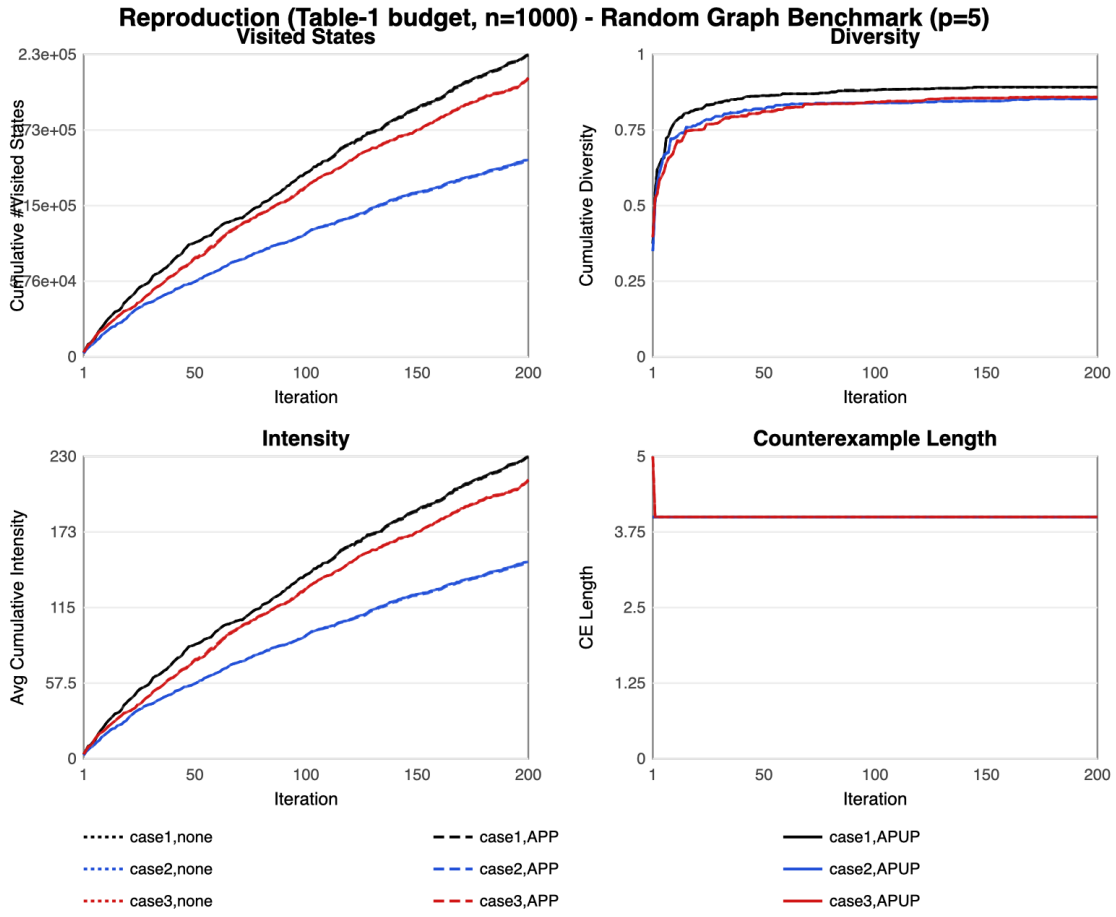

図 1　再現図（Table-1 予算、n=1000、p=5）。左上=累積訪問状態数、右上=累積 AC_{div}、左下=平均累積 AC_{int}、右下=反例長。3 戦略（none/APP/APUP）の線はほぼ完全に重なり、区別できない。Diversity は早期急増後に平坦化（論文 Fig.2 と定性一致）。

の大小に依らず効果ゼロ。

■原因 2: APP は重みを上げても効かない（更新対象のミスマッチ）　重み **500 倍**でも baseline と区別不能（+0.0002）。APP/APUP の更新は**反例に現れた抽象状態だけ**を対象にする（アルゴリズム 1 line 8）。反例は誤り状態近傍の短いパスで、触れる抽象状態は 1001 状態のごく一部。被覆（多様性）は**抽象空間全体**の問題なのに、更新は誤り局所の数状態しか動かさない。残りの多様化は既に $pathlen \cdot firstvisit$ 項が担っており、ペナルティはそこへ介入できない。**更新機構（反例駆動）と目標（広域被覆）が噛み合っていない。**

⇒ これは単なるパラメータ調整ではなく**設計レベルの問題**。被覆を上げたいなら、更新は「反例上の抽象状態」ではなく「**未訪問の抽象状態全体**」を対象に探索を誘導すべき（例: 訪問済み抽象状態にペナルティ、未訪問へ誘引するボーナス項を $firstvisit$ と同オーダーで導入）。

6　総合判定

- **定性的トレンド（被覆の早期急増 → 飽和、強度の単調増加、メトリクスの有用性）は、小予算・大予算（Table-1）の双方で再現された。**

- **核心の定量的主張（提案更新戦略 APP/APUP が単純再初期化より有効）は再現せず**、Table-1 予算をそのまま大きな積空間で回しても、結合強度を 500 倍にしても両者は統計的に同等。§5 の通り**設計レベルの原因**を特定した。論文の要旨・結論はこの差を過大に表現していると評価する。

■改善の推奨

(a) 更新対象を反例局所から**未訪問抽象状態全体**へ拡張（被覆を直接誘導）。
(b) APUP が介入する項を $firstvisit$ と同オーダーに再設計。
(c) 複数 seed と効果量・有意差検定の追加。
(d) 反例長と効果の関係の分析。
(e) AC_{int} 比例関係が平均化の自明な帰結である点の明記。
(f) 再現に必要な仕様（符号化・スウォーム生成・初期状態）の明文化。