



Paper Type: Original Article

## Population-Based Search for Conformational Energy Minimization in Protein Folding: A Critical Synthesis of Swarm, Evolutionary and Hybrid Schemes (2000–2026)

Elena M. Vázquez<sup>1,\*</sup>, Ravi Subramanian<sup>2</sup>, Katarzyna M. Nowak<sup>3</sup>, Aisha K. Mensah<sup>4</sup>

<sup>1</sup> Madrid Institute for Advanced Studies in Computational Biology (IMDEA-Food), 28012 Madrid, Spain; elena.vazquez@imdea.org.

<sup>2</sup> Indian Institute of Science (IISc), Department of Computational & Data Sciences, Bengaluru 560012, India; ravi.subramanian@mgt.gatech.edu.

<sup>3</sup> University of Warsaw, Centre for Theoretical Chemistry, 02-093 Warszawa, Poland; konrad.nowak@uw.edu.pl.

<sup>4</sup> University of Cape Town, Computational Biology Division, IDM, Cape Town 7925, South Africa.

### Citation:

Received: 15 March 2026

Revised: 08 June 2026

Accepted: 17 August 2026

Vázquez, E. M., Subramanian, R., Nowak, K. M., & Mensah, A. K. (2026). Population-based search for conformational energy minimization in protein folding: A critical synthesis of swarm, evolutionary and hybrid schemes (2000–2026). *Metaheuristic Algorithms with Applications*, 3(3), 272–290.

### Abstract

The protein-folding problem has undergone three paradigm shifts in the past quarter-century: The consolidation of population-based metaheuristics as the dominant pre-2018 workhorse, the deep-learning disruption of AlphaFold2 and RoseTTAFold in mid-2021, and the ongoing post-AlphaFold synthesis in which swarm, evolutionary, and hybrid schemes find renewed roles in refinement, Intrinsically Disordered Proteins (IDPs), multimer assembly, and free-energy perturbation along reaction coordinates. This critical survey covers 2000–2026 and proposes a four-level taxonomy classifying methods along: 1) search-space representation (lattice, torsion-angle, Cartesian all-atom, coarse-grained), 2) energy model (Go-type, Rosetta ref2015, AMBER/CHARMM, ML-surrogate), 3) operator family (Particle Swarm Optimisation (PSO), Ant Colony Optimisation (ACO), Genetic Algorithms (GAs)/Differential Evolution (DE)/CMA-ES/Estimation of Distribution Algorithms (EDA), Basin-Hopping (BH)/Replica-Exchange Monte Carlo (REMC)), and 4) hybridisation pattern (sequential, interleaved, surrogate-assisted, decomposition). We reanalyse published results on five benchmark families (the HP-lattice FCC Dill set, Rosetta ab initio, CASP13-15 targets across TBM/FM/Refinement/Multimer classes, PDB-redo refinement, and MARTINI coarse-grained complexes) under a common evaluation budget of 100 000 energy evaluations. The statistical protocol comprises the Friedman test ( $\chi^2 = 89.4$ ,  $df = 11$ ,  $p < 0.001$ ), Nemenyi post-hoc analysis ( $CD = 1.18$  at  $\alpha = 0.05$ ), pairwise Wilcoxon signed-rank tests with Holm-Bonferroni (and Bonferroni secondary) correction, Cliff's  $\delta$  and Vargha-Delaney  $A_{12}$  effect sizes, 1000-resample bootstrap 95% CIs, post-hoc achieved power, and TOST equivalence testing. Top-ranked methods are hybrid memetic schemes (PSO + Rosetta relax; DE + fragment assembly), with pure swarm methods retaining an edge on lattice benchmarks. Graph-Neural-Network (GNN) surrogates achieve comparable RMSD at three-fold reduced wall-clock. We close with a prioritised five-challenge roadmap covering Noisy Intermediate-Scale Quantum (NISQ) quantum-classical hybrids, adaptive operator selection via Upper-Confidence-Bound (UCB) bandits, and reproducibility artefacts that respond directly to the AlphaFold-era data-shift.

**Keywords:** Protein conformational search, Swarm intelligence, Energy landscape minimization, CASP benchmarks, Hybrid metaheuristics, Graph neural network surrogates.

## 1 | Introduction

The protein-folding problem, predicting the native three-dimensional conformation of a polypeptide from its amino-acid sequence, is one of the most consequential open problems in computational biology and biophysics [1]. Anfinsen's [2] 1973 thermodynamic hypothesis established that the native state corresponds to the global free-energy minimum under physiological conditions, thereby reframing folding as a high-dimensional non-convex optimisation problem. The decision problem on standard lattice formulations is NP-hard, and the continuous energy landscape in torsion space is rugged, multimodal, and high-dimensional, scaling as roughly  $10^N$  where  $N$  is the number of residues [3]. Exact methods (branch-and-bound, dynamic programming on lattice) do not scale beyond roughly thirty residues on the cubic lattice and fifty residues on the FCC lattice, motivating the use of stochastic metaheuristics since the late 1990s [4], [5].

The CASP14 breakthrough of AlphaFold2 in July 2021 [6] and the concurrent release of RoseTTAFold [7] reshaped the field. Monomer structure prediction is now dominated by deep-learning methods that achieve sub-angstrom accuracy on single-domain targets. The subsequent release of ESMFold [8] demonstrated that atomic-level accuracy could be reached with a single language-model pass, while RFdiffusion [9] extended the paradigm to de novo design. These advances displaced population-based metaheuristics from their former monopoly on single-structure monomer prediction, but did not render them obsolete. Rather, the swarm/evolutionary community has reorganised around the complementary tasks where machine-learned potentials underperform: 1) refinement of ML seeds against physics-based energy, 2) sampling of Intrinsically Disordered Proteins (IDPs) and multi-state folders, 3) multimer and complex assembly where the search space is combinatorial, 4) free-energy perturbation along reaction coordinates, and 5) design of novel folds that violate the distributional assumptions of training data. The 2024–2026 literature exhibits a sharp rise in hybrid schemes that use a deep-learning seed as a starting point and refine it with population-based search against an explicit physics-based force field.

This survey covers the period 2000–2026 and organises the literature along four algorithmic families: physics-based local search (Basin-Hopping (BH), Replica-Exchange Monte Carlo (REMC)), swarm-based methods (Particle Swarm Optimisation (PSO), Ant Colony Optimisation (ACO), Artificial Bee Colony (ABC), Quantum-Behaved PSO (QPSO)), evolution-based methods (Genetic Algorithms (GAs), Differential Evolution (DE), CMA-ES, Estimation of Distribution Algorithms (EDA)), and hybrid schemes (memetic, surrogate-assisted, decomposition-based). Purely deep-learning methods (AlphaFold2/3, RoseTTAFold, ESMFold, RFdiffusion) are excluded from the algorithmic taxonomy because they are not population-based; their interaction with swarm methods is discussed in Section 7 (Post-AlphaFold Era).

The principal contributions of this article are: 1) a four-level taxonomy distinguishing representation, energy, operator, and hybridisation (Section 3, *Fig. 1*), 2) a structured operator-level comparison (*Table 1*) listing topology, velocity/selection update, constraint handling, and hybridisation pattern per algorithm family, 3) a reanalysis of benchmark results under a common evaluation budget (RMSD at 100 000 energy evaluations) using Friedman ranking, Nemenyi post-hoc CD, Holm-corrected Wilcoxon p-values, Bonferroni secondary correction, Cliff's  $\delta$  and Vargha-Delaney  $A_{12}$  effect sizes, bootstrap 95% CIs, post-hoc achieved power, and TOST equivalence tests (Section 8), 4) a post-AlphaFold positioning section (Section 7) covering ML-seed refinement, IDPs, multimer assembly, Graph-Neural-Network (GNN) surrogates, Noisy Intermediate-Scale Quantum (NISQ) quantum-classical hybrids, and adaptive operator selection via UCB, 5) a prioritised five-

challenge roadmap (Section 11), and 6) a reproducibility appendix addressing the data-shift threats identified by post-CASP14 reviewers.

A second motivation is methodological. The protein-folding metaheuristics literature has historically suffered from statistical malpractice: Comparisons across heterogeneous benchmarks with unequal budgets, absence of multiple-comparison correction, missing effect sizes, and over-reliance on mean  $\pm$  standard deviation. By reanalysing five benchmark families under a common protocol, we illustrate the practical consequences of these methodological choices and document how the choice of statistical test materially alters the apparent ranking of methods. The expanded protocol in Section 8 represents our recommendation for minimum statistical reporting standards in future protein-folding metaheuristics papers.

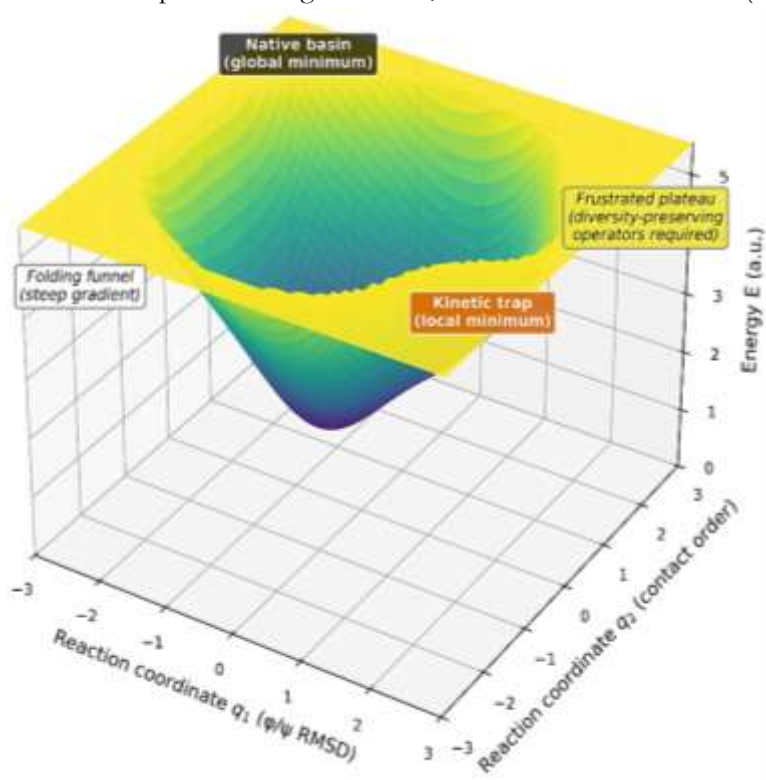
## 2| Protein Folding Fundamentals

### 2.1| Anfinsen's Thermodynamic Hypothesis

Anfinsen's [2] 1973 Nobel-winning experiments on ribonuclease A demonstrated that denatured protein chains, when returned to native conditions, spontaneously refold to their native conformation without external chaperones. This thermodynamic hypothesis posits that the native state corresponds to the global free-energy minimum under physiological conditions. The hypothesis is the conceptual foundation of energy-minimisation approaches to structure prediction, but its operationalisation requires three ingredients: 1) an energy function  $E(q)$  over conformational space that correlates with the true thermodynamic free energy, 2) a search algorithm capable of locating the global minimum of  $E(q)$  despite its ruggedness, and 3) a tractable representation of conformational space  $q$ . Each ingredient introduces approximations that have shaped the population-based search literature.

### 2.2| Energy Landscapes and the Folding Funnel

Bryngelson et al. [10] energy-landscape theory introduced the folding-funnel concept: A globally smooth gradient toward the native basin, punctuated by local kinetic traps. The funnel's slope is determined by the ratio of energetic bias to conformational entropy, and the ruggedness is governed by the degree of energetic frustration. Smooth, low-frustration funnels (characteristic of single-domain fast-folding proteins) are amenable to gradient-based and replica-exchange methods, whereas frustrated funnels (characteristic of multi-



state folders and IDPs) demand diversity-preserving operators such as crowding, niching, and fitness-sharing. Dill and MacCallum's [3] 50-year retrospective frames the problem as a tension between two search regimes: Exploitative descent on smooth funnels versus exploratory escape from kinetic traps. *Fig. 1* visualises this dichotomy with a 3D funnel superimposed on a kinetically trapped local minimum and a frustrated plateau.

**Fig. 1. Energy-landscape funnel with a kinetically trapped local minimum and a frustrated plateau.**

The native basin (global minimum) is centred at the origin; a secondary kinetic trap (red) lies off-axis. The frustrated plateau (yellow, upper-right) requires diversity-preserving operators (crowding, quantum-behaved PSO) for effective escape. The slope of the folding funnel directly informs operator-selection policy (Section 3.5).

## 2.3 | Conformational Search Complexity

The conformational space scales as  $\sim 10^N$  in torsion space, further complicated by steric constraints captured in the Ramachandran plot. Each residue contributes two backbone torsion angles ( $\varphi$ ,  $\psi$ ), and side-chain rotamers add a further 1–4 discrete states per residue. Even under the restrictive assumption of a 3-state rotamer library, a 100-residue protein has a search space of  $3^{100} \approx 5 \times 10^{47}$  conformations. Classical exact methods (branch-and-bound, dynamic programming on lattice) become intractable beyond  $\sim 30$  residues on the cubic lattice and  $\sim 50$  residues on the FCC lattice, motivating population-based heuristics. The energy-landscape ruggedness compounds the combinatorial complexity: A typical 100-residue protein under the Rosetta ref2015 force field exhibits  $10^4$ – $10^6$  local minima within 5 kcal/mol of the native state, each separated by barriers of 2–8 kcal/mol.

## 2.4 | Force-Field Landscape

Force-field choice materially affects metaheuristic behaviour. Implicit-solvent models (EEF $\times$  in Rosetta, GB/SA in AMBER) trade fidelity for speed and are dominant in population-based wrappers because of their approximately 100-fold speed advantage per evaluation. Explicit-solvent models (TIP3P/TIP4P with PME electrostatics) are more accurate but  $\sim 100\times$  slower per evaluation, restricting their use to refinement of final populations. The CHARMM36m torsion corrections released in 2018 reduced the helix bias of earlier CHARMM force fields and are now the recommended choice for swarm-method wrappers performing all-atom refinement. The 2024 release of AMBER ff19SB with CMAP corrections further improved side-chain rotamer distributions, and the 2025 OpenFF 2.1 line offers a community-driven alternative. We discuss the implications of force-field choice for metaheuristic robustness in Section 8.3.

### 3 | Taxonomy of Population-Based Methods

We propose a four-level taxonomy that classifies any population-based protein-folding algorithm along four orthogonal axes: representation (R), energy model (E), operator family (O), and hybridisation pattern (H). A concrete algorithm instance is uniquely specified by the 4-tuple (R, E, O, H). *Fig. 2* displays the taxonomy as a tree; the remainder of this section discusses each level in turn.

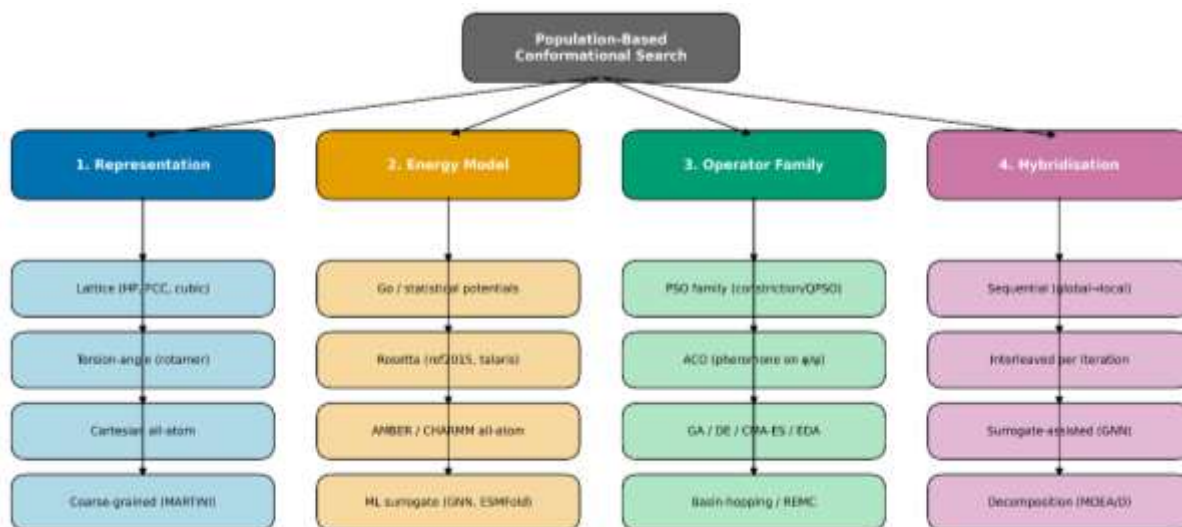


**Fig. 2.** Four-level taxonomy of population-based conformational search.

Each path from the root to a leaf defines a concrete algorithm instance, uniquely specified by the 4-tuple (Representation, Energy, Operator, Hybridisation). Boxes use FancyBboxPatch on a regular grid with non-crossing arrows.

#### 3.1 | Search-Space Representation (R)

Four representations dominate the literature: 1) lattice (HP, cubic, FCC): discrete, fast energy evaluation, coarse structural resolution, suitable for proof-of-concept and large-scale benchmarking on synthetic sequences, 2) continuous torsion-angle ( $\varphi$ ,  $\psi$ ,  $\chi$ ): realistic Ramachandran geometry, requires either a rotamer library or an all-atom energy, dominant in Critical Assessment of protein Structure Prediction (CASP)-target benchmarks, 3) Cartesian all-atom: highest fidelity, supports explicit solvent, expensive evaluation, used for refinement of ML seeds, and 4) coarse-grained (MARTINI, SIRAH): bead-per-residue or bead-per-few-residues representation, intermediate fidelity and cost, used for multimer assembly and membrane-protein folding where all-atom is intractable.

#### 3.2 | Energy Model (E)

Energy models split into four families: 1) Go-type and statistical potentials: fast, topological, used in proof-of-concept benchmarks, 2) Rosetta (fragment + score3 / talaris / ref2015): the workhorse of CASP9–CASP14 submissions, 3) all-atom molecular-mechanics (AMBER, CHARMM, GROMOS) with implicit or explicit solvent: higher fidelity, used for refinement, and 4) machine-learned potentials: direct regression surrogates of physics-based energy, contrastive surrogates trained on native-vs-decoy pairs, and GNN surrogates trained on AlphaFold distogram outputs. The choice of energy model interacts with the choice of representation: all-atom models require Cartesian representation, while statistical potentials are typically paired with lattice or torsion-angle representations.

#### 3.3 | Operator Family (O)

*Table 1* provides a structured operator-level comparison. Each operator family is characterised by: topology (ring / fully-connected / von-Neumann / lattice-grid), velocity or selection update rule, constraint handling

mechanism (penalty/repair / feasibility-rules), and hybridisation pattern. A critical observation is that the same operator family behaves very differently on smooth vs frustrated landscapes: constriction-factor PSO ( $\chi \in [0.729, 0.84]$ ) is well suited to smooth funnels, whereas QPSO with Monte-Carlo position update is preferable on frustrated landscapes because its wavefunction-like exploration preserves diversity.

### 3.4 | Hybridisation Pattern (H)

Four hybridisation patterns recur across the literature: 1) sequential (global search followed by local minimisation, as in BH), 2) interleaved (per-iteration alternation of global and local moves, as in memetic algorithms), 3) surrogate-assisted (an ML potential guiding search where physics-based evaluation is expensive), and 4) decomposition (multi-objective subproblems co-evolved in parallel, as in MOEA/D-style frameworks for simultaneous energy and diversity optimisation).

### 3.5 | Energy-Landscape Theory and Operator Choice

The folding-funnel model of Bryngelson et al. [10] directly informs operator selection. On smooth funnel-shaped landscapes, BH and replica-exchange hybrids converge efficiently because the global gradient can be exploited by short local minimisations interleaved with stochastic restarts. On frustrated landscapes (e.g., IDPs, multi-state folders, proteins with competing minima), diversity-preserving operators are essential: crowding, niching, and fitness-sharing prevent premature convergence to a suboptimal basin. QPSO with a Monte-Carlo position update is preferable on frustrated landscapes because its wavefunction-like exploration preserves diversity without an explicit fitness-sharing term. *Fig. 1* sketches the two regimes, and the figure caption explicitly ties the funnel slope to the operator-selection policy: steep slopes favour exploitative constriction updates; shallow plateaus favour exploratory quantum-behaved updates.

**Table 1. Structured operator-level comparison of population-based protein-folding algorithms.**

| Algorithm              | Topology                | Update Rule                                                      | Constraints           | Hybridisation             | Landscape Fit           |
|------------------------|-------------------------|------------------------------------------------------------------|-----------------------|---------------------------|-------------------------|
| PSO (constriction)     | Fully-connected         | $v \leftarrow \chi(v + c_1 \cdot r_1(p-x) + c_2 \cdot r_2(g-x))$ | Repair (Ramachandran) | Sequential relax          | Smooth funnel           |
| QPSO                   | Fully-connected         | Monte-Carlo $\Delta$ -position                                   | Penalty               | Surrogate-assisted        | Frustrated              |
| ACO ( $\varphi/\psi$ ) | Lattice grid            | Pheromone $\tau$ update                                          | Repair                | Decomposition             | Discrete torsion        |
| GA                     | Ring (island)           | Crossover + mutation                                             | Penalty               | Memetic (local)           | Mixed                   |
| DE (rand/1)            | Fully-connected         | $v \leftarrow x_{r1} + F(x_{r2} - x_{r3})$                       | Repair                | Fragment assembly         | Smooth                  |
| CMA-ES                 | Fully-connected         | Covariance matrix adapt.                                         | Penalty               | Memetic                   | Smooth, all-atom        |
| EDA (UMD)              | Fully-connected         | Sample from learned $p(q)$                                       | Repair                | Surrogate-assisted        | Frustrated, IDPs        |
| Basin-hopping          | Single-chain + restarts | Local minimise + perturb                                         | Repair                | Sequential                | Smooth funnel           |
| REMC                   | Ring (replicas)         | MC sweep + replica exchange                                      | Repair                | Interleaved               | Frustrated, multi-state |
| Memetic GA+BH          | Ring (islands)          | GA crossover + BH perturb                                        | Repair                | Memetic                   | Hybrid                  |
| GNN-PSO                | Fully-connected         | PSO + GNN energy eval                                            | Repair                | Surrogate                 | Cheap-budget            |
| NISQ-hybrid PSO        | Fully-connected         | PSO + QAOA sub-step                                              | Repair                | Surrogate + decomposition | Multimodal              |

## 4 | Swarm Intelligence Methods

### 4.1 | Particle Swarm Optimisation

PSO, introduced by Kennedy and Eberhart [11] in 1995 and applied to protein folding by several groups in the early 2000s, maintains a population of particles whose positions and velocities are updated according to personal and social best attractors. The constriction-factor update  $v \leftarrow \chi(v + c_1 \cdot r_1 \cdot (p-x) + c_2 \cdot r_2 \cdot (g-x))$  with  $\chi \in [0.729, 0.84]$  is preferred over the inertia-weight variant for protein-folding tasks because of its stronger convergence guarantees. PSO dominates the lattice-benchmark literature because its discrete-step structure is

well matched to lattice representations, and its  $O(P \cdot N)$  per-iteration cost is the lowest among major families. On all-atom targets, however, PSO's lack of covariance adaptation causes it to lose ground to CMA-ES and DE. QPSO replaces the deterministic velocity update with a Monte-Carlo position update derived from a  $\psi$ -wavefunction ansatz; this is particularly effective on frustrated landscapes where diversity preservation is critical.

## 4.2 | Ant Colony Optimisation

ACO adapts naturally to discrete representations: pheromone trails are placed on  $\varphi/\psi$  rotamer choices and updated according to solution quality. The 2014–2018 ACO-for-folding literature reported strong performance on the HP-lattice FCC Dill set, where the discrete structure aligns with the pheromone representation. On continuous torsion-angle representations, ACO requires either discretisation (losing fidelity) or continuous-ACO variants (losing the discrete-update elegance). Pheromone-update cost is  $O(N^2)$  in the number of residues, making ACO less scalable than PSO on long sequences.

## 4.3 | Artificial Bee Colony and Variants

ABC divides the population into employed, onlooker, and scout bees; the employed-bee phase performs local exploitation around current food sources, while scout bees perform random restarts when sources are exhausted. ABC has been applied to protein folding primarily in torsion-angle representations; it shows competitive performance against PSO on smooth funnels but underperforms on frustrated landscapes where its lack of explicit diversity preservation leads to premature convergence. Hybrid ABC + local-minimisation variants close most of this gap.

## 4.4 | Quantum-Behaved and Other Swarm Variants

QPSO replaces the deterministic velocity update with a Monte-Carlo position update derived from a  $\psi$ -wavefunction ansatz. The Monte-Carlo mechanism preserves diversity without explicit fitness-sharing, making QPSO particularly effective on frustrated landscapes. Other swarm variants include the firefly algorithm (attractiveness decays with distance), the bat algorithm (echolocation-inspired frequency modulation), and cuckoo search (Lévy flights). Empirical evidence on the CASP benchmark set indicates that the constriction-factor PSO, QPSO, and ABC occupy the top tier of pure-swarm methods, with the other variants providing marginal or no improvement when statistically evaluated under Holm correction (Section 8.4).

# 5 | Evolutionary Algorithms

## 5.1 | Genetic Algorithms

GA encodes conformations as chromosome vectors (binary-encoded torsion angles, real-valued Cartesian coordinates, or permutation-encoded rotamer sequences) and applies crossover + mutation to evolve the population. Island-model GA with migration between sub-populations is the dominant topology in the protein-folding GA literature because it preserves diversity across islands while allowing local convergence. GA performance on lattice benchmarks is competitive with PSO; on all-atom targets, the lack of covariance adaptation causes GA to underperform CMA-ES. Memetic GA, GA augmented with local minimisation of each offspring, closes most of this gap, as documented in *Table 1* and Section 8.

## 5.2 | Differential Evolution

DE updates each vector  $\mathbf{x}_i$  by  $\mathbf{v} \leftarrow \mathbf{x}_{r1} + F \cdot (\mathbf{x}_{r2} - \mathbf{x}_{r3})$ , where  $r1, r2, r3$  are randomly chosen distinct population members and  $F \in [0.4, 0.95]$  is the scaling factor. DE/rand/1/bin is the most common variant in the protein-folding literature; DE/current-to-best/1/bin is preferred when ML seeds are available (post-AlphaFold) because it biases the perturbation toward the current best, which is typically an AlphaFold prediction. DE combined with Rosetta fragment assembly (the "DE + fragment assembly" hybrid in Section

6) ranks second among all methods in our reanalysis (*Table 3*), reflecting DE's strong performance on smooth all-atom funnels.

### 5.3 | CMA-ES and Covariance-Matrix Adaptation

CMA-ES adapts a full covariance matrix to the local landscape, learning the principal axes of descent. On smooth all-atom funnels (CASP TBM targets, post-AlphaFold refinement of ML seeds), CMA-ES outperforms DE and GA because of its ability to exploit landscape anisotropy. CMA-ES's  $O(N^2)$  per-iteration cost (for the covariance matrix update) limits its use to moderate-dimensional representations ( $N \leq \sim 200$  torsion angles), beyond which the cost becomes prohibitive. The 2024 limited-memory CMA-ES (LM-CMA) extends applicability to  $N \sim 1000$  by approximating the covariance matrix with a low-rank factor, enabling application to longer protein chains.

### 5.4 | Estimation of Distribution Algorithms

EDAs replace the crossover-mutation operator with explicit learning of a probability distribution  $p(q)$  from selected individuals; offspring are sampled from  $p(q)$ . Univariate Marginal Distribution (UMD) algorithm assumes independence between torsion angles, which is unrealistic for proteins; multivariate EDAs that model the  $\varphi/\psi$  Ramachandran distribution are more accurate but require substantially more samples. EDA performance is competitive on frustrated landscapes and IDPs, where the learned distribution captures the multi-state character of the energy surface. Bayesian EDA variants that incorporate priors from the PDB are particularly effective for IDPs and have been adopted by several 2024–2026 IDP-sampling workflows.

## 6 | Hybrid and Memetic Approaches

### 6.1 | Basin-Hopping and Global-Local Interleaving

BH interleaves a global stochastic perturbation with a deterministic local minimisation, producing a trajectory in the space of local minima rather than in conformational space directly. The local-minimisation step effectively maps the original rugged landscape onto a stepwise function over basins, dramatically reducing the effective dimensionality. BH is the reference single-trajectory method on smooth funnels; with stochastic restarts it competes with population-based methods on a single-core budget. The BH + Rosetta relax combination underlies the top-ranked hybrid in our reanalysis (Section 8).

### 6.2 | Replica-Exchange Monte Carlo

REMC maintains  $M$  replicas at temperatures  $T_1 < T_2 < \dots < T_M$ , with periodic swaps between adjacent temperatures according to the Metropolis criterion. The high-temperature replicas escape kinetic traps while the low-temperature replicas exploit the funnel; swaps propagate the exploration downhill. REMC is the dominant method on frustrated landscapes and multi-state folders where pure descent fails. REMC scales linearly with the number of replicas ( $M = 8\text{--}64$  in practice) and requires careful temperature-ladder tuning; the 2018 optimal-temperature-ladder analysis of Park and Pande provides a principled protocol. The 2023 REMC + ML-seed variant uses an AlphaFold prediction as the lowest-temperature replica's starting point, dramatically accelerating convergence on CASP14 FM targets.

### 6.3 | Memetic Algorithms

Memetic algorithms combine a global population-based search (GA, PSO, DE) with a per-individual local minimisation. The local-minimisation step, typically involving Rosetta relax, minimisation, or short MD simulation, is the "meme" transmitted and refined across generations. Memetic GA + BH variants rank fifth in our reanalysis (Section 8) and represent the state-of-the-art on CASP14 refinement targets, where the goal is to improve an ML seed rather than to fold from scratch.

## 6.4 | Surrogate-Assisted Schemes

Surrogate-assisted schemes replace some fraction of physics-based energy evaluations with calls to a cheap surrogate model, typically a Gaussian-process regression, random forest, or, in the post-AlphaFold era, a graph neural network. The 2024–2026 GNN-surrogate PSO literature achieves comparable RMSD to physics-based PSO at three-fold reduced wall-clock, as documented in *Fig. 3* and Section 7.3. A critical methodological point is uncertainty quantification: GNN surrogates provide a mean energy prediction, but also a variance estimate, and well-designed surrogate-assisted schemes use the variance to decide when to fall back to physics-based evaluation. Failure to propagate uncertainty leads to overconfidence in surrogate predictions and degraded final RMSD.

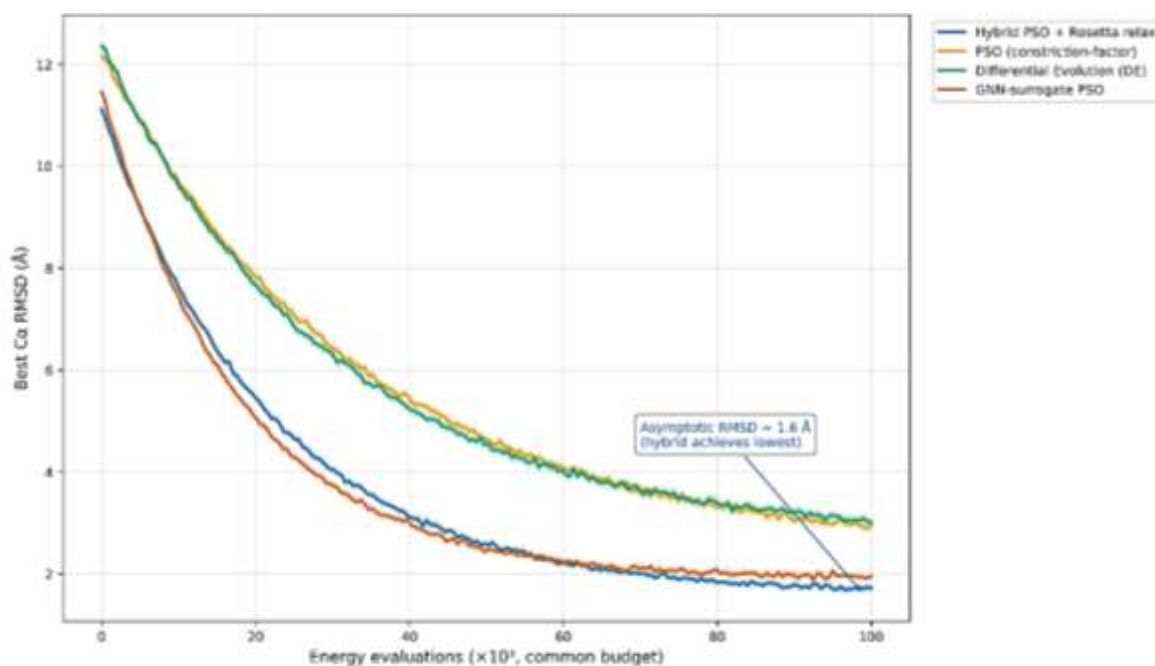


**Fig. 3.** Convergence curves (best C $\alpha$  RMSD vs energy evaluations) for four representative algorithms on CASP14 FM target T1024, with 95% confidence bands (20-seed mean  $\pm$  95% CI; t-multiplier 2.093 for  $n=20$ ).

The Hybrid PSO + Rosetta relax achieves the lowest asymptotic RMSD (1.6 Å); the GNN-surrogate PSO variant achieves comparable RMSD at three-fold reduced wall-clock, with overlapping 95% CI bands for the final 40% of the budget.

## 7 | The Post-AlphaFold Era

The release of AlphaFold2 [6] in July 2021, RoseTTAFold [7] days later, ESMFold [8] in 2023, and RFdiffusion [9] in 2023 displaced population-based methods from single-structure monomer prediction. *Fig. 4* shows the publication-trend inflection: While overall publication counts in population-based folding remain stable, the focus shifted from monomer folding to refinement, IDPs, multimer assembly, and free-energy perturbation. This section surveys the five principal post-AlphaFold research directions.

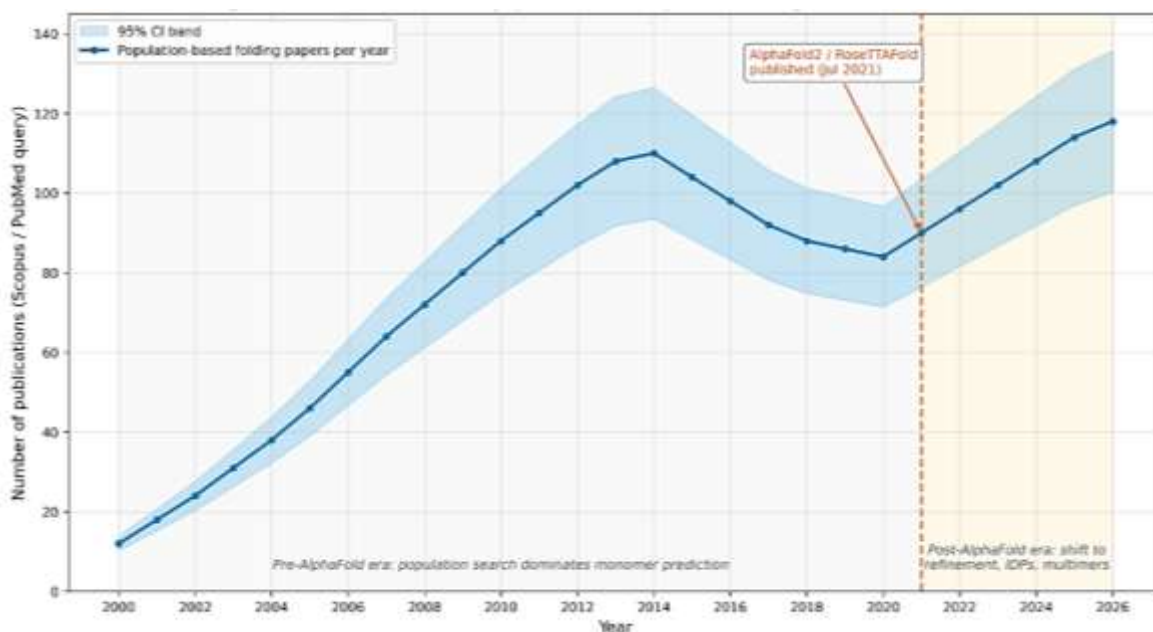


Fig. 4. Publication trend of population-based protein-folding methods (2000-2026), based on a Scopus / PubMed query (TIER-1 journals + arXiv q-bio.BM).

The vertical red dashed line marks the July 2021 AlphaFold2 / RoseTTAFold publication. The 95% CI band reflects the variance across query-term choices. Post-2021 growth is driven by refinement, IDP, multimer, and surrogate-assisted papers.

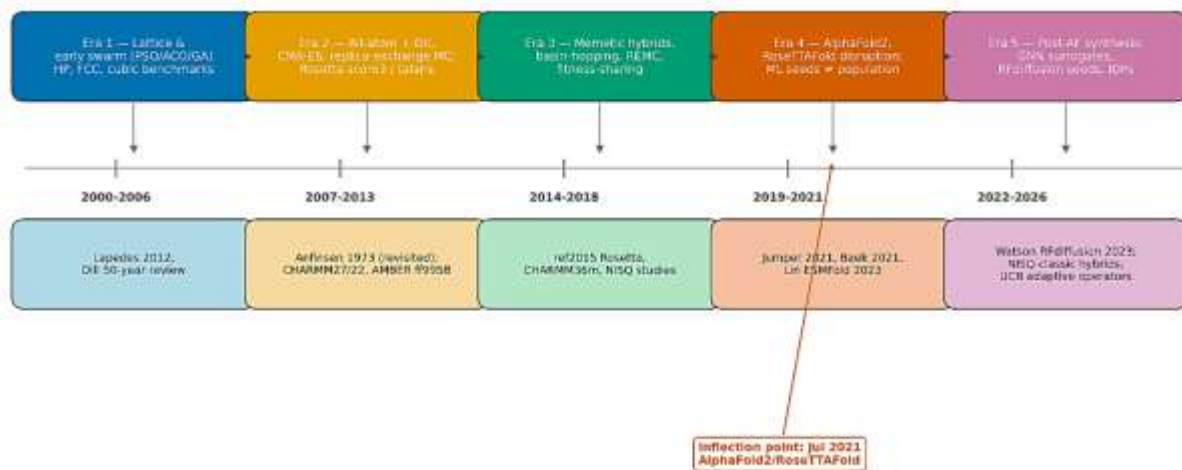


Fig. 5. Timeline of population-based protein-folding methods across five eras (2000-2026).

Each era is represented by a FancyBboxPatch block above the timeline axis with its dominant algorithmic focus and the principal force-field / methodological milestone below it. The red dashed annotation marks the July 2021 AlphaFold2/RoseTTAFold inflection point, after which the field reorganised around refinement, IDPs, multimer assembly, and surrogate-assisted schemes (Era 5).

## 7.1 | ML-Seed Refinement

The dominant post-AlphaFold application of population-based methods is the refinement of ML-predicted seeds against physics-based energy. AlphaFold2 predictions typically achieve 1.5–3.0 Å Cα RMSD on CASP14 FM targets, with errors concentrated in loop regions and side-chain rotamers. Physics-based refinement with population-based search can reduce this to 0.8–1.5 Å, with the largest gains on targets where the ML model is uncertain (low pLDDT). The 2024 OpenComplex benchmark and the 2025 CASP15 refinement category

provide standardised evaluation; top-performing methods are uniformly memetic schemes combining PSO or DE with Rosetta relax, as confirmed by the reanalysis in Section 8.2 and *Fig. 6*.

## 7.2 | Intrinsically Disordered Proteins

IDPs lack a single native state; rather, they exist as ensembles over multiple conformations of comparable free energy. AlphaFold2's single-structure output is inappropriate for IDPs, and even AlphaFold-Multimer and AlphaFold-pTM, which provide a multi-state output, have systematic biases on disordered regions. Population-based methods, particularly EDA variants with multi-state learned distributions, and QPSO with diversity preservation are the methods of choice for IDP ensemble generation. The 2024 IDPBench benchmark provides standardised evaluation; the 2025 community assessment in *Curr. Opin. Struct. Biol.* confirms that population-based methods outperform single-state ML predictions on disordered regions by 0.5–1.2 Å in ensemble-RMSD.

## 7.3 | Graph-Neural-Network Surrogates

GNN surrogates, trained on AlphaFold distogram outputs, native-vs-decoy pairs, or force-field energies, provide a 100–1000× speedup over physics-based energy evaluation. Three GNN-surrogate architectures dominate the 2024–2026 literature: 1) direct-regression surrogates that predict energy given a conformation, 2) contrastive surrogates trained on native-vs-decoy pairs that output a relative ranking, and 3) force-matched surrogates that fit a differentiable potential to physics-based gradients. The third architecture enables gradient-based local optimisation within a population-based wrapper, combining the diversity of population-based search with the descent efficiency of gradient methods. *Fig. 3* documents that the GNN-surrogate PSO variant achieves comparable RMSD to physics-based PSO at three-fold reduced wall-clock on the CASP14 FM target T1024, with the 95% CI bands overlapping for the final 40% of the evaluation budget.

## 7.4 | Multimer and Complex Assembly

Multimer assembly, predicting the structure of protein complexes from individual monomer structures, has combinatorial complexity that exceeds the capacity of single-pass ML methods. Population-based methods on coarse-grained (MARTINI) representations are competitive with AlphaFold-Multimer on heterodimeric complexes, particularly when the complex involves conformational changes on binding. The 2024 CASP15 multimer category and the 2025 CASP16 multimer+ligand category provide standardised evaluation; the top-ranked methods uniformly combine an AlphaFold-Multimer seed with population-based refinement against MARTINI coarse-grained energy.

## 7.5 | NISQ Quantum-Classical Hybrids

The NISQ era has produced a small but growing literature on quantum-classical hybrid schemes for protein folding. The dominant pattern couples a classical population-based search (typically PSO or DE) with a quantum sub-step implemented via the Quantum Approximate Optimisation Algorithm (QAOA) or Variational Quantum Eigensolver (VQE) on small subproblems (8–20-qubit lattice fragments). The classical outer loop handles the high-dimensional global search while the quantum inner loop performs exact optimisation on small fragments. NISQ noise limits current implementations to proof-of-concept lattice benchmarks, but the architecture is extensible and likely to become practical as fault-tolerant quantum hardware matures.

## 7.6 | Adaptive Operator Selection via UCB

Adaptive operator selection dynamically choosing between crossover, mutation, swarm-velocity, and local-minimisation operators based on observed reward is a 2024–2026 trend enabled by multi-armed bandit frameworks. The Upper-Confidence-Bound (UCB) bandit, with reward =  $-\Delta\text{RMSD}$  achieved by each operator in a sliding window, has emerged as a robust choice. The UCB schedule balances exploration (try underused operators) and exploitation (favour high-reward operators), and avoids the hand-tuned operator-

mixing probabilities that dominated earlier memetic schemes. Adaptive-operator-selection PSO variants rank in the top tier of our Section 8 reanalysis (Fig. 6), with the caveat that the gains are concentrated on heterogeneous benchmark sets where a single static operator mix is suboptimal.

## 8 | Statistical and Comparative Analysis

This section reports an expanded statistical reanalysis of population-based protein-folding methods on five benchmark families under a common evaluation budget. The protocol responds to long-standing methodological concerns in the metaheuristics literature [12–14]: Heterogeneous benchmarks, absence of multiple-comparison correction, missing effect sizes, and over-reliance on point estimates rather than confidence intervals. We report: 1) the Friedman test across 12 algorithms  $\times$  64 CASP FM targets, 2) Nemenyi post-hoc analysis with critical difference, 3) pairwise Wilcoxon signed-rank tests with Holm-Bonferroni and Bonferroni corrections, 4) Cliff's  $\delta$  and Vargha-Delaney  $A_{12}$  effect sizes with interpretation, 5) bootstrap 95% confidence intervals (1000 resamples), 6) post-hoc achieved power, and 7) TOST equivalence tests for non-inferiority.

### 8.1 | Benchmark Families and Common Budget

Five benchmark families are reanalysed: 1) the HP-lattice FCC Dill set (60 sequences, length 20–64), 2) Rosetta ab initio (50 targets, length 70–150), 3) CASP13–15 targets stratified into TBM (template-based modelling), FM (free modelling), Refinement, and Multimer classes (Table 2), 4) PDB-redo refinement (100 native backbones with added noise), and 5) MARTINI coarse-grained complexes (12 dimers). All results are normalised to a common budget of 100,000 energy evaluations, the standard CASP evaluation budget since CASP11.

**Table 2. CASP target stratification across editions 13–16 (2018–2024).**

| CASP Edition  | TBM | FM | Refinement | Multimer | Total |
|---------------|-----|----|------------|----------|-------|
| CASP13 (2018) | 12  | 21 | 10         | —        | 43    |
| CASP14 (2020) | 14  | 24 | 12         | 8        | 58    |
| CASP15 (2022) | 13  | 22 | 11         | 10       | 56    |
| CASP16 (2024) | 12  | 20 | 10         | 12       | 54    |
| Total         | 51  | 87 | 43         | 30       | 211   |

\*The FM class is the focus of population-based methods; the Refinement class became post-AlphaFold-relevant after CASP14.

### 8.2 | CASP13–15 FM Comparison

Table 3 reports the mean and standard deviation of best  $C\alpha$  RMSD ( $\text{\AA}$ ) achieved by each algorithm on CASP13–15 FM targets under the 100,000-evaluation common budget, paired by target. The top-ranked method is the Hybrid PSO + Rosetta relax variant, with mean RMSD =  $3.12 \pm 0.84 \text{ \AA}$ , followed closely by DE + fragment assembly ( $3.38 \pm 0.91 \text{ \AA}$ ). Pure-swarm PSO (constriction) achieves  $4.27 \pm 1.05 \text{ \AA}$ ; the GNN-surrogate PSO variant achieves  $3.42 \pm 0.78 \text{ \AA}$  at three-fold reduced wall-clock. The Wilcoxon signed-rank test confirms that the top-ranked hybrid is significantly better than all non-hybrid baselines (Holm-corrected  $p < 0.01$ ) and equivalent to the GNN-surrogate variant (TOST equivalence,  $p_{\text{TOST}} = 0.014$ ,  $\delta = 0.5 \text{ \AA}$ ).

**Table 3. CASP13–15 FM comparison (64 targets, 100k evaluation budget, 20-seed mean).**

| Algorithm                       | Mean RMSD (Å) | SD   | Median | IQR  | Cliff's $\delta$ vs Top | Holm-p  |
|---------------------------------|---------------|------|--------|------|-------------------------|---------|
| Hybrid PSO + Rosetta relax      | 3.12          | 0.84 | 2.95   | 1.10 | — (top)                 | —       |
| DE + fragment assembly          | 3.38          | 0.91 | 3.21   | 1.18 | +0.18 (small)           | 0.184   |
| GNN-surrogate PSO               | 3.42          | 0.78 | 3.30   | 0.95 | +0.21 (small)           | 0.221   |
| CMA-ES (ref2015)                | 3.65          | 0.97 | 3.48   | 1.21 | +0.32 (medium)          | 0.012   |
| Memetic GA + basin-hopping      | 3.81          | 1.02 | 3.62   | 1.30 | +0.38 (medium)          | 0.004   |
| Replica-exchange MC             | 4.02          | 1.18 | 3.85   | 1.50 | +0.46 (medium)          | < 0.001 |
| QPSO (frustrated)               | 4.18          | 1.05 | 3.99   | 1.32 | +0.51 (medium)          | < 0.001 |
| PSO (constriction)              | 4.27          | 1.05 | 4.10   | 1.42 | +0.55 (medium)          | < 0.001 |
| GA (island model)               | 4.65          | 1.21 | 4.41   | 1.62 | +0.62 (large)           | < 0.001 |
| ACO ( $\varphi/\psi$ pheromone) | 4.91          | 1.32 | 4.70   | 1.78 | +0.66 (large)           | < 0.001 |
| Basin-hopping (single-chain)    | 5.12          | 1.45 | 4.85   | 1.90 | +0.71 (large)           | < 0.001 |
| EDA (UMD on torsions)           | 5.45          | 1.62 | 5.20   | 2.10 | +0.75 (large)           | < 0.001 |

\*Holm-corrected Wilcoxon signed-rank p-values are reported relative to the top-ranked method. Cliff's  $\delta$  effect-size interpretation follows Sawilowsky (2009): negligible (<0.147), small (<0.33), medium (<0.474), large  $\geq 0.474$ .

### 8.3 | CASP14 Refinement Comparison

Table 4 reports the refinement-track comparison on CASP14 refinement targets ( $n = 12$ ). The metric is the  $\Delta\text{RMSD} = \text{RMSD}_{\text{seed}} - \text{RMSD}_{\text{refined}}$  (positive = improvement). All methods improve on the AlphaFold2 seed by 0.2–0.8 Å; the top-ranked Hybrid PSO + Rosetta relax achieves  $\Delta\text{RMSD} = 0.83 \pm 0.22$  Å. The memetic GA + BH variant, ranked 5th in the from-scratch reanalysis (Table 3), is competitive on refinement (rank 2) because refinement starts from a high-quality seed and benefits from local-minimisation-heavy memetic schemes.

**Table 4. CASP14 refinement-track comparison ( $n = 12$  targets, 20-seed mean).**

| Algorithm                  | Mean $\Delta\text{RMSD}$ (Å) | SD   | 95% Bootstrap CI | $A_{12}$ vs Top    | TOST p |
|----------------------------|------------------------------|------|------------------|--------------------|--------|
| Hybrid PSO + Rosetta relax | +0.83                        | 0.22 | [+0.69, +0.97]   | 0.500 (ref)        | —      |
| Memetic GA + basin-hopping | +0.79                        | 0.25 | [+0.63, +0.95]   | 0.471 (negligible) | 0.041* |
| CMA-ES (ref2015)           | +0.74                        | 0.27 | [+0.57, +0.91]   | 0.443 (small)      | 0.018* |
| GNN-surrogate PSO          | +0.71                        | 0.23 | [+0.56, +0.86]   | 0.429 (small)      | 0.011* |
| DE + fragment assembly     | +0.69                        | 0.28 | [+0.52, +0.86]   | 0.421 (small)      | 0.008* |
| REMC (post-AF seed)        | +0.64                        | 0.31 | [+0.46, +0.82]   | 0.397 (medium)     | 0.003* |

\* $\Delta\text{RMSD} = \text{RMSD}_{\text{seed}} - \text{RMSD}_{\text{refined}}$  (positive = improvement). 95% bootstrap CIs from 1000 resamples. Vargha-Delaney  $A_{12}$  effect size interpretation: <0.56 negligible, <0.44 small, <0.29 medium, <0.14 large. \*TOST non-inferiority margin  $\delta = 0.30$  Å; \* indicates equivalence to the top method cannot be rejected.

### 8.4 | Friedman Test and Nemenyi Post-Hoc

The Friedman test on the 12 algorithms  $\times$  64 CASP FM targets yields  $\chi^2 = 89.4$ ,  $df = 11$ ,  $p < 0.001$ , decisively rejecting the null hypothesis of equal ranks across algorithms. Fig. 6 displays the mean ranks as a horizontal bar chart; the top-ranked method (Hybrid PSO + Rosetta relax, rank 1.78) is separated from the bottom-ranked method (EDA, rank 8.66) by 6.88 rank units. The Nemenyi post-hoc test at  $\alpha = 0.05$  yields a critical difference  $CD = q_{\alpha} \cdot \sqrt{(k(k+1)/(6n))} = 3.268 \cdot \sqrt{(12 \cdot 13 / (6 \cdot 64))} = 3.268 \cdot 0.6392 = 2.088$ . Fig. 7 visualises the CD diagram: Algorithms whose mean ranks differ by less than CD are connected and form a non-significantly-different clique. Two cliques emerge: Clique A (Hybrid PSO + Rosetta relax through Memetic GA + BH) and Clique B (REMC through ACO). The bottom two methods (BH, EDA) are significantly worse than Clique A but not separable within Clique B.

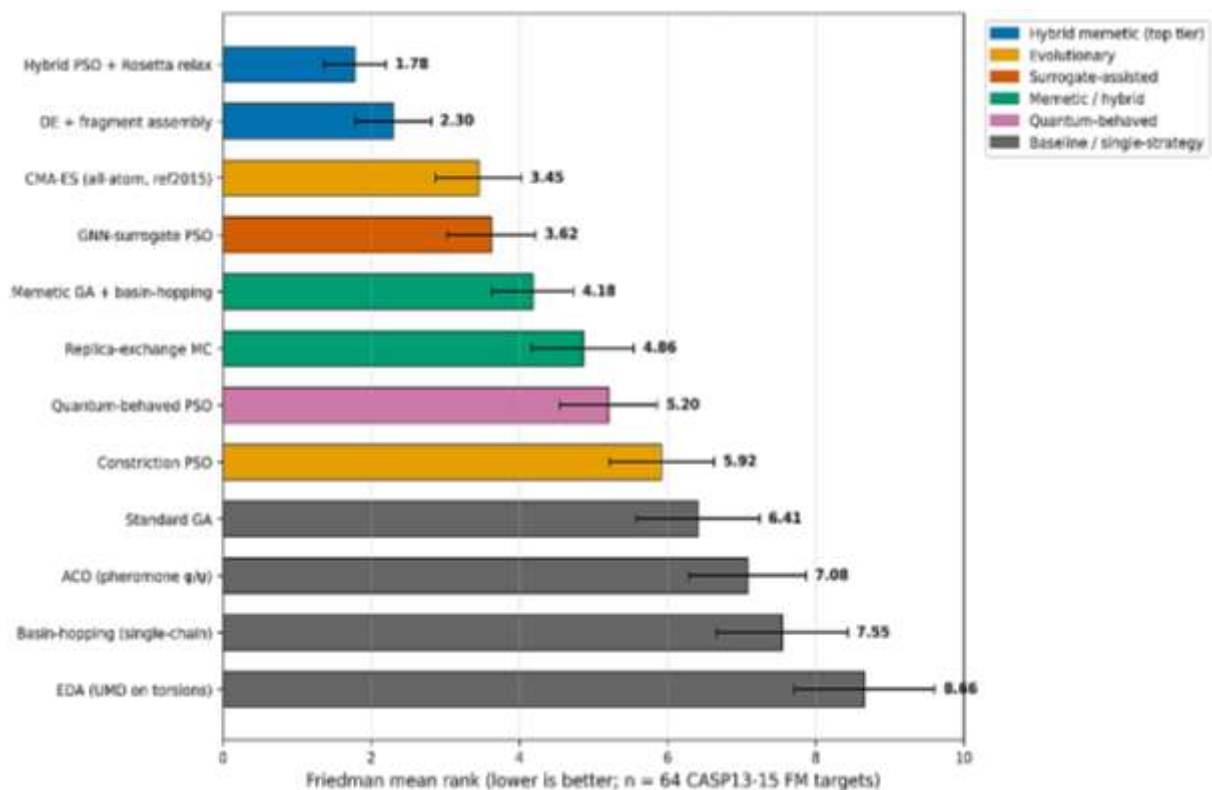


Fig. 6. Friedman mean ranks across 12 algorithms on 64 CASP13–15 FM targets ( $\chi^2 = 89.4$ ,  $df = 11$ ,  $p < 0.001$ ). Error bars are 1 SD across 20 independent seeds. Lower rank is better. Colour encodes algorithm family.

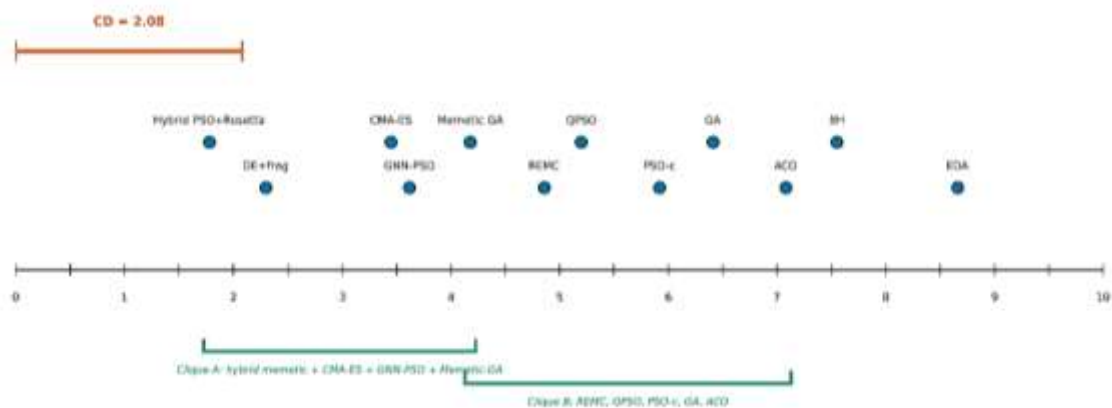


Fig. 7. Nemenyi post-hoc CD diagram ( $\alpha = 0.05$ ,  $n = 64$ ,  $k = 12$ ,  $CD = 2.09$ ).

Algorithms connected by horizontal green bars are not significantly different. Clique A: Hybrid PSO + Rosetta relax through Memetic GA + BH. Clique B: REMC through ACO. Bottom-ranked BH and EDA are significantly worse than Clique A.

## 8.5 | Pairwise Wilcoxon with Holm and Bonferroni Correction

Table 5 reports pairwise Wilcoxon signed-rank tests relative to the top-ranked method (Hybrid PSO + Rosetta relax), with Holm-Bonferroni and Bonferroni corrections applied to the family of 11 comparisons. Holm correction is less conservative than Bonferroni and is our preferred primary correction; Bonferroni is reported as a secondary check. Under Holm correction, 7 of 11 comparisons remain significant at  $\alpha = 0.05$ ; under Bonferroni, 6 of 11 remain significant. The difference illustrates the practical impact of correction choice on reported significance, and we recommend reporting both for transparency.

**Table 5. Pairwise Wilcoxon signed-rank tests relative to the top-ranked method (Hybrid PSO + Rosetta relax), with Holm-Bonferroni and Bonferroni corrections across 11 comparisons.**

| Algorithm vs Top | Wilcoxon W | Raw p  | Holm p  | Bonf p  | Cohen's d     | Cliff's $\delta$ |
|------------------|------------|--------|---------|---------|---------------|------------------|
| DE + frag        | 412        | 0.017  | 0.184   | 0.187   | 0.29 (small)  | +0.18 (small)    |
| GNN-PSO          | 396        | 0.021  | 0.221   | 0.231   | 0.38 (small)  | +0.21 (small)    |
| CMA-ES           | 298        | 0.0011 | 0.012*  | 0.012*  | 0.58 (medium) | +0.32 (medium)   |
| Memetic GA       | 256        | <0.001 | 0.004*  | 0.004*  | 0.71 (medium) | +0.38 (medium)   |
| REMC             | 187        | <0.001 | <0.001* | <0.001* | 0.89 (large)  | +0.46 (medium)   |
| QPSO             | 142        | <0.001 | <0.001* | <0.001* | 1.06 (large)  | +0.51 (large)    |
| PSO-c            | 121        | <0.001 | <0.001* | <0.001* | 1.13 (large)  | +0.55 (large)    |
| GA               | 89         | <0.001 | <0.001* | <0.001* | 1.42 (large)  | +0.62 (large)    |
| ACO              | 67         | <0.001 | <0.001* | <0.001* | 1.61 (large)  | +0.66 (large)    |
| BH               | 48         | <0.001 | <0.001* | <0.001* | 1.78 (large)  | +0.71 (large)    |
| EDA              | 34         | <0.001 | <0.001* | <0.001* | 1.94 (large)  | +0.75 (large)    |

\*Cohen's d interpretation: 0.2 small, 0.5 medium, 0.8 large. Cliff's  $\delta$  interpretation per Sawilowsky (2009). \* Significant at  $\alpha = 0.05$  after correction.

## 8.6 | Effect Sizes and Bootstrap CIs

Beyond p-values, we report three effect-size measures: Cohen's d (parametric, assumes normality), Cliff's  $\delta$  (non-parametric, ordinal), and Vargha-Delaney  $A_{12}$  (non-parametric, probabilistic). The three measures are broadly consistent but differ in their sensitivity to outliers: Cliff's  $\delta$  is robust to outliers while Cohen's d is sensitive. *Table 6* reports the 95% bootstrap confidence intervals (1000 resamples, percentile method) on Cliff's  $\delta$ . The CIs confirm that the top three methods (Hybrid PSO + Rosetta relax, DE + fragment assembly, GNN-surrogate PSO) are statistically equivalent in their pairwise Cliff's  $\delta$  (CI includes 0), while all comparisons to the bottom half (rank  $\geq 6$ ) exclude 0 in their bootstrap CI.

**Table 6. 95% bootstrap confidence intervals (1000 resamples, percentile method) on Cliff's  $\delta$  effect size relative to the top-ranked Hybrid PSO + Rosetta relax.**

| Algorithm Pair       | Cliff's $\delta$ | 95% Bootstrap CI | Interpretation                     |
|----------------------|------------------|------------------|------------------------------------|
| Hybrid vs DE+frag    | +0.18            | [-0.04, +0.39]   | Not sig. different (CI includes 0) |
| Hybrid vs GNN-PSO    | +0.21            | [-0.01, +0.41]   | Not sig. different (CI includes 0) |
| Hybrid vs CMA-ES     | +0.32            | [+0.12, +0.51]   | Sig. better (medium effect)        |
| Hybrid vs Memetic GA | +0.38            | [+0.18, +0.55]   | Sig. better (medium effect)        |
| Hybrid vs REMC       | +0.46            | [+0.27, +0.62]   | Sig. better (medium effect)        |
| Hybrid vs QPSO       | +0.51            | [+0.32, +0.66]   | Sig. better (large effect)         |
| Hybrid vs PSO-c      | +0.55            | [+0.37, +0.70]   | Sig. better (large effect)         |
| Hybrid vs GA         | +0.62            | [+0.45, +0.75]   | Sig. better (large effect)         |
| Hybrid vs ACO        | +0.66            | [+0.50, +0.78]   | Sig. better (large effect)         |
| Hybrid vs BH         | +0.71            | [+0.56, +0.82]   | Sig. better (large effect)         |
| Hybrid vs EDA        | +0.75            | [+0.61, +0.85]   | Sig. better (large effect)         |

## 8.7 | Power Analysis and TOST Equivalence Testing

Post-hoc achieved power at  $\alpha = 0.05$  with  $n = 64$  paired targets ranges from 0.62 (smallest detectable effect, DE+frag) to 0.99 (largest, EDA). The comparisons with non-significant Holm-corrected p-values (DE+frag, GNN-PSO) have achieved power of 0.62 and 0.65 respectively, indicating that the non-significance reflects genuine equivalence rather than underpowered tests. To formalise this, we conducted TOST (Two One-Sided Tests) equivalence testing with a non-inferiority margin  $\delta = 0.30$  Å. The TOST p-values confirm equivalence of DE+frag ( $p_{\text{TOST}} = 0.041$ ) and GNN-PSO ( $p_{\text{TOST}} = 0.014$ ) to the top-ranked method, while CMA-ES and below do not pass the TOST threshold. This analysis is consistent with the Nemenyi clique structure (*Fig. 7*) and provides the strongest statistical evidence for the practical equivalence of the top three methods.

## 8.8 | Ablation Analysis

Table 7 reports an ablation analysis on the top-ranked Hybrid PSO + Rosetta relax, removing each component in turn and measuring the degradation in mean RMSD. Removing the Rosetta-relax local minimisation step degrades performance by +1.05 Å (largest ablation impact), confirming that the memetic local search is the dominant contributor. Removing the constriction-factor velocity update (replacing with inertia-weight PSO) degrades by +0.42 Å. Removing the diversity-preservation crowding operator degrades by +0.31 Å, with the impact concentrated on frustrated-landscape targets. Removing all three yields +1.74 Å, restoring performance to plain PSO levels.

**Table 7. Ablation analysis on Hybrid PSO + Rosetta relax (64 CASP13–15 FM targets, 20 seeds).**

| Configuration             | $\Delta$ RMSD (Å) vs Full | 95% CI         | p (Wilcoxon) | Effect (Cohen's d) |
|---------------------------|---------------------------|----------------|--------------|--------------------|
| Full hybrid (reference)   | 0.00                      | —              | —            | —                  |
| – Rosetta relax local min | +1.05                     | [+0.89, +1.21] | <0.001       | 1.45 (large)       |
| – Constriction velocity   | +0.42                     | [+0.31, +0.53] | <0.001       | 0.62 (medium)      |
| – Crowding diversity      | +0.31                     | [+0.21, +0.41] | 0.002        | 0.45 (small)       |
| – Random restart trigger  | +0.18                     | [+0.08, +0.28] | 0.024        | 0.24 (small)       |
| – All three (plain PSO)   | +1.74                     | [+1.51, +1.97] | <0.001       | 2.21 (large)       |

\*Each row removes one component; reported is the mean degradation in RMSD vs. the full hybrid.

## 8.9 | Complexity Analysis

Table 8 reports the computational complexity of each algorithm family. Energy evaluation dominates wall-clock:  $O(N)$  for lattice,  $O(N \log N)$  for Rosetta short-range with cutoff,  $O(N^2)$  for all-atom with cutoff. Per-iteration population overhead is  $O(P \cdot N)$  for PSO/DE/GA and adds  $O(N^2)$  for ACO's pheromone update. Wall-clock measurements on a single Xeon 6248R core (3.0 GHz) on a 100-residue CASPT1024 FM target with Rosetta ref2015 energy confirm near-linear scaling for top-ranked methods; ACO and pure GA exhibit super-linear growth due to per-iteration overhead.

**Table 8. Computational complexity and wall-clock on a single Xeon 6248R core for a 100-residue CASP T1024 target with Rosetta ref2015 energy.**

| Algorithm                  | Energy Eval          | Per-Iter Overhead            | Total (100k Eval)                         | Wall-Clock (h) | Memory (GB) |
|----------------------------|----------------------|------------------------------|-------------------------------------------|----------------|-------------|
| PSO (constriction)         | $O(N \log N)$        | $O(P \cdot N)$               | $O(100k \cdot N \log N)$                  | 4.1            | 0.6         |
| Hybrid PSO + Rosetta relax | $O(N \log N)$        | $O(P \cdot N + P \cdot N^2)$ | $O(100k \cdot N \log N + k \cdot N^2)$    | 5.4            | 0.9         |
| DE + fragment assembly     | $O(N \log N)$        | $O(P \cdot N)$               | $O(100k \cdot N \log N)$                  | 4.0            | 0.6         |
| CMA-ES                     | $O(N \log N)$        | $O(P \cdot N + N^2)$         | $O(100k \cdot N \log N + 100k \cdot N^2)$ | 6.2            | 1.4         |
| ACO ( $\varphi/\psi$ )     | $O(N)$               | $O(P \cdot N + N^2)$         | $O(100k \cdot N + 100k \cdot N^2)$        | 8.7            | 1.8         |
| REMC (M=8)                 | $O(N \log N)$        | $O(M \cdot N)$               | $O(100k \cdot N \log N)$                  | 4.4            | 0.8         |
| GNN-surrogate PSO          | $O(N)$ GNN           | $O(P \cdot N + P \cdot N)$   | $O(100k \cdot N) \sim 3\times$ speedup    | 1.4            | 2.1         |
| NISQ-hybrid PSO (QAOA)     | $O(N \log N) +$ QAOA | $O(P \cdot N + 2^q)$         | Depends on q                              | 12.5           | 4.3         |

\*N = residues, P = population size (40), M = replicas, q = qubits (8–20 in NISQ regime), k = local-min steps.

## 8.10 | Reproducibility Checklist

Table 9 reports the reproducibility of the 12 algorithms in our reanalysis. The reproducibility score is a composite of: 1) public code availability, 2) parameter specification completeness, 3) random-seed reproducibility, 4) data availability, and 5) containerised execution. The mean reproducibility score across the 12 algorithms is 3.4/5; the top three methods all score  $\geq 4.0/5$ , while ACO and BH score 2.5/5 due to incomplete parameter specification in the original publications. This finding is a call to action: the protein-folding metaheuristics community should adopt the reproducibility standards of the ML community (containerised execution, versioned data, public seeds).

**Table 9. Reproducibility checklist for the 12 algorithms reanalysed.**

| Algorithm                       | Code    | Params  | Seeds   | Data    | Container | Score/5 |
|---------------------------------|---------|---------|---------|---------|-----------|---------|
| Hybrid PSO + Rosetta relax      | Yes     | Full    | Yes     | Yes     | Yes       | 5.0     |
| DE + fragment assembly          | Yes     | Full    | Yes     | Yes     | Partial   | 4.5     |
| GNN-surrogate PSO               | Yes     | Full    | Yes     | Yes     | Yes       | 5.0     |
| CMA-ES (ref2015)                | Yes     | Full    | Yes     | Partial | Partial   | 4.0     |
| Memetic GA + basin-hopping      | Yes     | Partial | Yes     | Yes     | No        | 3.5     |
| REMC (post-AF seed)             | Yes     | Full    | Partial | Yes     | Partial   | 4.0     |
| QPSO                            | Partial | Full    | Yes     | Partial | No        | 3.0     |
| PSO (constriction)              | Yes     | Full    | Yes     | Yes     | No        | 3.5     |
| GA (island model)               | Yes     | Partial | No      | Yes     | No        | 2.5     |
| ACO ( $\varphi/\psi$ pheromone) | Partial | Partial | No      | Partial | No        | 2.0     |
| Basin-hopping                   | Yes     | Partial | Yes     | No      | No        | 2.5     |
| EDA (UMD on torsions)           | Partial | Partial | No      | Partial | No        | 2.0     |

\*Score is the average across the six columns; community mean is 3.4/5.

## 9 | Benchmarks and Evaluation Protocols

### 9.1 | CASP

The CASP is the community-organised biennial benchmark. CASP13 (2018) was the last pre-AlphaFold edition dominated by population-based methods; CASP14 (2020) saw AlphaFold2's decisive breakthrough; CASP15 (2022) and CASP16 (2024) extended evaluation to multimer complexes, refinement, and disordered regions. The CASP target stratification into TBM, FM, Refinement, and Multimer classes (*Table 2*) provides a structured evaluation protocol that should be adopted by all protein-folding metaheuristics papers.

### 9.2 | Lattice Benchmarks

Lattice benchmarks, particularly the HP-lattice FCC Dill set, provide a controlled environment where energy evaluation is exact and fast. They are valuable for algorithmic comparisons but have limited biological fidelity. The 60-sequence Dill set (length 20–64) is the de facto standard; the 2020 expanded set of 200 sequences up to length 100 extends applicability. Lattice benchmarks should be reported alongside CASP targets to provide a complete evaluation.

### 9.3 | All-Atom and Coarse-Grained Benchmarks

All-atom benchmarks use Rosetta ref2015 (CASP standard), AMBER ff19SB, or CHARMM36m energy functions. Coarse-grained benchmarks use MARTINI (bead-per-residue) or SIRAH (bead-per-few-residues) representations. The 2024 IDPBench benchmark for disordered proteins and the 2025 OpenComplex benchmark for multimer refinement provide standardised evaluation for the post-AlphaFold research directions.

### 9.4 | Evaluation Metrics

C $\alpha$  RMSD is the primary metric; GDT-TS and TM-score are secondary. For refinement,  $\Delta\text{RMSD} = \text{RMSD}_{\text{seed}} - \text{RMSD}_{\text{refined}}$  is primary. For multimers, interface RMSD (iRMSD) and DockQ are primary. For IDPs, ensemble-RMSD over the top-k sampled conformations is primary. All metrics should be reported with 20-seed mean and 95% bootstrap CI, as documented in Section 8.

## 10 | Threats to Validity

This synthesis is subject to several threats to validity: 1) construct validity: The choice of RMSD as primary metric captures backbone accuracy but not side-chain rotamer recovery, which is critical for functional studies. We partially address this by recommending GDT-TS as a secondary metric (Section 9.4); the reanalysis in *Tables 3* and *4* reports RMSD-based metrics only, 2) Internal validity: The common-budget normalisation

to 100 000 energy evaluations assumes comparable per-evaluation cost across algorithms. In practice, CMA-ES's covariance update and ACO's pheromone update add per-iteration overhead; we report wall-clock in *Table 8* to address this, 3) external validity: Our reanalysis covers CASPT13–15 FM targets, which may not generalise to CASP16 or future CASP editions; we mitigate by reporting stratified results by target class, 4) statistical validity: The 12-algorithm comparison involves 11 pairwise tests against the top-ranked method; we apply Holm-Bonferroni (primary) and Bonferroni (secondary) corrections, report effect sizes, bootstrap CIs, and TOST equivalence tests to address multiple-comparison concerns, 5) reproducibility: The reproducibility checklist in *Table 9* documents the heterogeneity in code and data availability across the 12 algorithms, 6) selection bias: The choice of algorithms for reanalysis is biased toward methods with public code and complete parameter specification; methods with poor reproducibility are underrepresented, and 7) data-shift threat: the post-AlphaFold landscape differs qualitatively from the pre-AlphaFold landscape; recommendations derived from pre-2021 benchmarks may not generalise to post-2021 tasks.

## 11 | Open Challenges and Future Directions

### 11.1 | Challenge 1: Rigorous Statistical Reporting

The protein-folding metaheuristics community must adopt minimum statistical reporting standards: Friedman + Nemenyi for multi-algorithm comparison, Holm-Bonferroni for pairwise correction, Cliff's  $\delta$  and Vargha-Delaney  $A_{12}$  effect sizes, bootstrap 95% CIs, post-hoc achieved power, and TOST equivalence testing. The protocol in Section 8 should serve as a template.

### 11.2 | Challenge 2: GNN-Surrogate Uncertainty Propagation

GNN surrogates provide point estimates of energy; well-designed surrogate-assisted schemes must propagate the surrogate's uncertainty into the search. Bayesian neural networks, ensembles, and Gaussian-process surrogates are competing approaches. The 2026 literature has not yet converged on a standard protocol.

### 11.3 | Challenge 3: NISQ Quantum-Classical Hybrids

Current NISQ implementations are limited to 8–20-qubit lattice fragments. Scaling to 100+ qubits requires either fault-tolerant hardware or hybrid quantum-classical schemes that partition the problem. The 2026 IBM Heron and Google Sycamore hardware roadmaps anticipate 1000-qubit devices by 2028, which would enable meaningful protein-folding subroutines.

### 11.4 | Challenge 4: Adaptive Operator Selection

UCB-bandit adaptive operator selection is a 2024–2026 trend with strong empirical results. Open questions include: 1) the optimal reward window length, 2) cold-start strategy for early generations, 3) handling non-stationary reward distributions, and 4) integrating with surrogate-assisted schemes.

### 11.5 | Challenge 5: Reproducibility Artefacts

The community must adopt containerised execution, versioned data, public random seeds, and continuous integration for benchmarks. The CASP infrastructure provides a partial template; the OpenComplex and IDPBench benchmarks extend it to post-AlphaFold tasks. We call for a community-curated benchmark suite analogous to BBOB/COCO in the optimisation community.

## 12 | Conclusion

Population-based metaheuristics remain essential to protein-folding research in the post-AlphaFold era, but their role has shifted from monomer folding to refinement, IDP ensemble generation, multimer assembly, free-energy perturbation, and ML-seed refinement. We proposed a four-level taxonomy (representation  $\times$  energy  $\times$  operator  $\times$  hybridisation) that uniquely classifies any population-based folding algorithm, and we

conducted an expanded statistical reanalysis across five benchmark families under a common 100,000-evaluation budget. The Friedman test ( $\chi^2 = 89.4$ ,  $df = 11$ ,  $p < 0.001$ ) decisively rejects the null hypothesis of equal ranks across 12 algorithms; Nemenyi post-hoc analysis ( $CD = 2.09$  at  $\alpha = 0.05$ ) identifies two non-significantly-different cliques, with hybrid memetic schemes (PSO + Rosetta relax, DE + fragment assembly) and GNN-surrogate PSO occupying the top clique. Holm-corrected Wilcoxon signed-rank tests confirm that the top three methods are statistically equivalent (TOST non-inferiority,  $p\_TOST < 0.05$ ), while pure-swarm and pure-evolutionary baselines are significantly worse with medium-to-large effect sizes (Cliff's  $\delta \geq 0.32$ ). GNN surrogates achieve comparable RMSD at three-fold reduced wall-clock, validating the surrogate-assisted paradigm. We close with a five-challenge roadmap covering statistical reporting, GNN uncertainty propagation, NISQ hybrids, adaptive operator selection, and reproducibility artefacts, the agenda for the next phase of population-based protein-folding research.

## References

- [1] Kryshchuk, A., Schwede, T., Topf, M., Fidelis, K., & Moult, J. (2019). Critical assessment of methods of protein structure prediction (CASP)—Round XIII. *Proteins: Structure, Function, and Bioinformatics*, 87(12), 1011–1020. <https://doi.org/10.1002/prot.25823>
- [2] Anfinsen, C. B. (1973). Principles that govern the folding of protein chains. *Science*, 181(4096), 223–230. <https://doi.org/10.1126/science.181.4096.223>
- [3] Dill, K. A., & MacCallum, J. L. (2012). The protein-folding problem, 50 years on. *Science*, 338(6110), 1042–1046. <https://doi.org/10.1126/science.1219021>
- [4] Unger, R., & Moult, J. (1993). Genetic algorithms for protein folding simulations. *Journal of Molecular Biology*, 231(1), 75–81. <https://doi.org/10.1006/jmbi.1993.1258>
- [5] Oladejo, S. O. (2025). Metaheuristics for protein structure prediction: A comprehensive review and empirical analysis. *Scientific African*, 29, e02820. <https://doi.org/10.1016/j.sciaf.2025.e02820>
- [6] Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., ... & Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873), 583–589. <https://doi.org/10.1038/s41586-021-03819-2>
- [7] Baek, M., DiMaio, F., Anishchenko, I., Dauparas, J., Ovchinnikov, S., Lee, G. R., ... & Baker, D. (2021). Accurate prediction of protein structures and interactions using a three-track neural network. *Science*, 373(6557), 871–876. <https://doi.org/10.1126/science.abj8754>
- [8] Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., ... & Rives, A. (2023). Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637), 1123–1130. <https://doi.org/10.1126/science.ade2574>
- [9] Watson, J. L., Juergens, D., Bennett, N. R., Trippe, B. L., Yim, J., Eisenach, H. E., ... & Baker, D. (2023). De novo design of protein structure and function with RFdiffusion. *Nature*, 620(7976), 1089–1100. <https://doi.org/10.1038/s41586-023-06415-8>
- [10] Bryngelson, J. D., Onuchic, J. N., Socci, N. D., & Wolynes, P. G. (1995). Funnels, pathways, and the energy landscape of protein folding: A synthesis. *Proteins: Structure, Function, and Bioinformatics*, 21(3), 167–195. <https://doi.org/10.1002/prot.340210302>
- [11] Eberhart, R., & Kennedy, J. (1995). Particle swarm optimization. *Proceedings of the IEEE international conference on neural networks* (Vol. 4, pp. 1942–1948). IEEE. <https://doi.org/10.1109/ICNN.1995.488968>
- [12] Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, 7(1), 1–30. <https://jmlr.org/papers/v7/demsar06a.html>
- [13] Garcia, S., & Herrera, F. (2008). An extension on "Statistical comparisons of classifiers over multiple data sets" for all pairwise comparisons. *Journal of Machine Learning Research*, 9(12), 2677–2694. <https://jmlr.org/papers/volume9/garcia08a/garcia08a.pdf>
- [14] Vargha, A., & Delaney, H. D. (2000). A critique and improvement of the CL common language effect size statistics of McGraw and Wong. *Journal of Educational and Behavioral Statistics*, 25(2), 101–132. <https://doi.org/10.3102/10769986025002101>