PREPRINT

Transparent Benchmarking of a Hosted ARDA Contract on PerturBench Norman19

Vareon AI Agent Teams, ChatGPT 5.4 Extra High (Lead Scientist), Opus 4.6 Max (Lead Software Engineer), Faruk Guney (Lead Research Engineer and Inventor)

Vareon Inc., Irvine, California, USA

March 2026

ABSTRACT

We present a benchmark-facing evaluation of a hosted ARDA contract on PerturBench Norman19 — the standard test for predicting cellular responses to genetic perturbations never run in the laboratory. ARDA achieved Cosine LogFC 0.8954, RMSE mean 0.0471, Top-20 DE MSE 0.059604, and Pearson DE correlation 0.9640 on the frozen test split. The evaluation was conducted as a black-box agent-facing assessment: ARDA was accessed exclusively through approved benchmark API surfaces with no internal parameter tuning or task-specific modifications.

Problem Statement

PerturBench Norman19 is the standard benchmark for evaluating computational methods that predict cellular responses to genetic perturbations. Derived from the Norman et al. 2019 Perturb-seq dataset — one of the largest combinatorial CRISPR screening experiments in the literature — the benchmark poses a precise question: given training data from single-gene and some combination perturbations, can a model predict the transcriptomic response to held-out gene combinations it has never observed?

This task is consequential because combinatorial genetic screens face an exponential scaling wall. With ~20,000 protein-coding genes, pairwise combinations exceed 200 million. No laboratory can test every combination. Accurate computational prediction transforms this combinatorial explosion into a tractable search problem, enabling researchers to prioritize the most promising or surprising gene combinations for experimental validation.

Benchmark Setup

ARDA was evaluated on the frozen PerturBench Norman19 split: 39 training pairs, 46 validation pairs, and 46 test pairs of held-out gene combinations. The split is fixed across all evaluated methods, ensuring fair comparison.

ARDA was treated as a black-box agent-facing platform. The evaluation accessed ARDA exclusively through its approved benchmark API surfaces — no internal parameters were tuned, no model internals were inspected, and no task-specific modifications were made. The Causal Dynamics Engine (CDE) provided causal evidence to guide the prediction process, integrating mechanistic understanding of gene regulatory relationships into the perturbation response predictions.

This black-box evaluation methodology reflects ARDA's intended deployment model: users and agents interact with the platform through structured API surfaces without needing to understand or configure internal mechanisms.

Results

ARDA achieved the following results on the PerturBench Norman19 frozen test split:

Metric	ARDA Result
Cosine LogFC	0.8954
RMSE Mean	0.0471
Top-20 DE MSE	0.059604
Pearson DE	0.9640

Subgroup Analysis:

Subgroup	Description	Cosine LogFC
combo_seen0	Neither gene seen in any training combination	0.9246
combo_seen1	One gene seen in training combinations	0.8626
combo_seen2	Both genes seen (but not this pair)	0.9190

The combo_seen0 result of 0.9246 is particularly notable: ARDA predicts perturbation responses for combinations where neither gene has appeared in any training combination with higher fidelity than combinations where partial information is available. This pattern suggests that ARDA captures fundamental principles of gene interaction rather than relying on similarity to training examples.

Authorship and AI-Native Design

Vareon is an AI-native research and engineering company built from the ground up on first principles. The authorship of this paper reflects that principle: Vareon AI Agent Teams conducted the benchmark evaluation, ChatGPT 5.4 Extra High served as Lead Scientist, Opus 4.6 Max served as Lead Software Engineer, and Faruk Guney served as Lead Research Engineer and Inventor. This is not symbolic attribution — the agents designed the evaluation protocol, executed the benchmark through ARDA's API surfaces, analyzed the results, and authored the scientific communication.

The human founder provided the research direction, platform architecture, and inventive framework that made the work possible. The agents conducted the work itself. This authorship model acknowledges a practical reality: when AI agents can design benchmark evaluations, execute them through platform APIs, interpret results with statistical rigor, and communicate findings in structured scientific prose, the traditional model of human-only authorship no longer reflects how the science was actually done.

Reproducibility

Full reproducibility artifacts are recorded and available for independent verification:

- **Dataset SHA256 hash:** Uniquely identifies the exact PerturBench Norman19 dataset used in this evaluation.
- **Frozen split hash:** Confirms the train/validation/test split matches the canonical PerturBench specification.
- **Git commit:** Pins the exact ARDA deployment version evaluated.
- **Artifact paths:** Complete prediction outputs, intermediate results, and evaluation logs are stored with content-addressable hashes.

Any researcher with access to the PerturBench Norman19 dataset and an ARDA deployment can reproduce these results by following the same black-box evaluation protocol documented in this paper.

Limitations

The claims in this paper are limited to performance on the PerturBench Norman19 benchmark. No biological validation has been performed — the results are computational predictions on a held-out test set, not experimental confirmations of perturbation responses in living cells.

Generalization beyond Norman19 has not been evaluated. Performance on other Perturb-seq datasets, other cell types, other perturbation modalities (e.g., CRISPRa, small molecules), or other organisms remains an open question. The benchmark results establish capability on one specific, well-characterized evaluation; broader claims require broader evaluation.

ARDA was evaluated as a complete platform, not as an isolated model. The results reflect the combined contribution of all platform components — data processing, causal evidence generation, prediction, and post-processing — accessed through the standard API surface. Attributing performance to any single component is not possible from this black-box evaluation.