Benchmarking structural predictors and pLM binary classifiers for predicting cross-kingdom protein-protein interactions in crop disease pathosystems

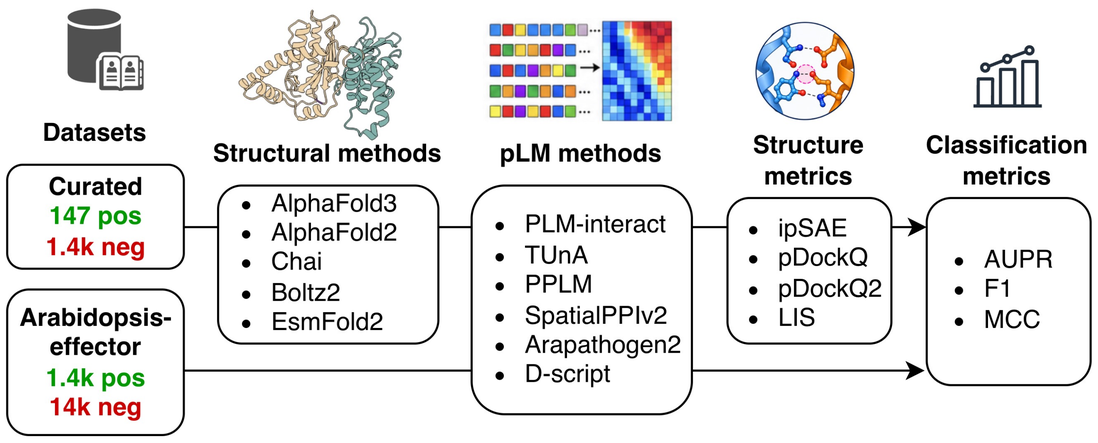
Pavel Misiun[1], Callum Verdonk[1], Mark Derbyshire[1], Kristina Gagalova[1]
[1]Centre for Crop and Disease Management (CCDM), School of Molecular and Life Sciences, Curtin University

ABSTRACT

Crop pathogens secrete **effector proteins** that interact with host proteins to facilitate disease. The discovery of these interactions may be accelerated by emerging methods in structural bioinformatics and AI. We benchmarked five **structure prediction** methods and six **protein language model (pLM)-based classifiers** on experimentally confirmed **effector-host interactions** and non-interacting pairs. Structure prediction methods modelled individual folds with moderate confidence, but interface confidence was low and poorly discriminative. Most frontier pLM classifiers performed near random, whereas **AraPathogen2.0**, trained on plant-pathogen data, performed best and improved further in combination with Boltz2. These results highlight the limitations of current methods and the value of combining complementary approaches.

Benchmark design

All methods were evaluated on a **curated dataset** of experimentally confirmed effector-host interactions (29 pathogen and 16 host species). pLM-based classifiers were additionally evaluated on a larger **Arabidopsis-effector dataset.**

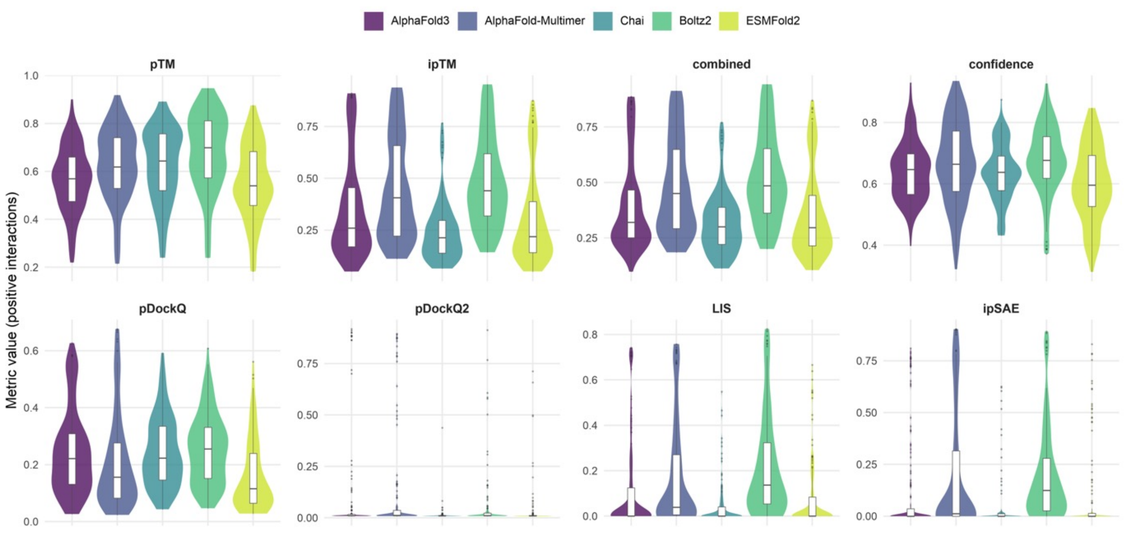


Modelling known interactions

Structure prediction methods (e.g. AlphaFold) report global confidence scores (0-1) at three levels:

Confidence in	Metrics	Observed
1. Individual chain folds	pTM	**Moderate**
2. Relative placement of partners	ipTM, combined, confidence	**Low**
3. Binding interface	pDockQ, pDockQ2, LIS, ipSAE	**Lowest**

**combined = 0.8 × ipTM + 0.2 × pTM; confidence = 0.8 × pLDDT + 0.2 × ipTM (Boltz2).*



Confidence metrics for curated positive interactions across five structure prediction methods.

These results suggest that the benchmarked methods can reliably fold individual partners in effector-host complexes, but are markedly less confident about their co-folding into a bound complex.

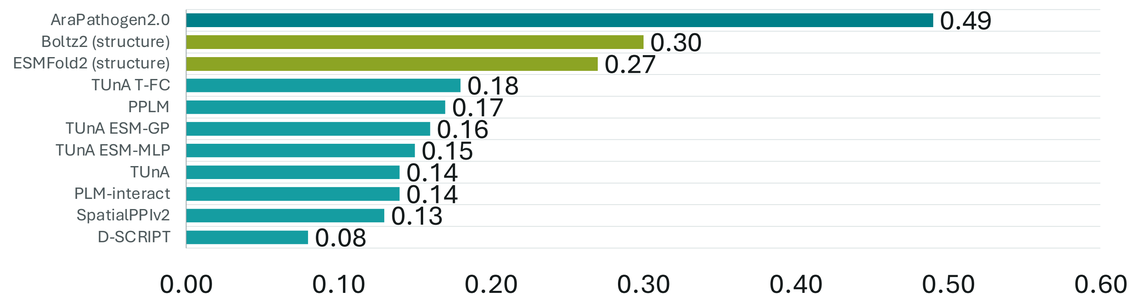
Sources of low interface confidence

Interface prediction relies on inferring **co-evolution** (correlated variation of contacting residues), derived from **multiple sequence alignments (MSAs)** [1] or learned implicitly by protein language models [2]. Effector-host pairs offer little such signal:

- **Shallow effector alignments:** Effectors evolve rapidly and are often restricted to one pathogen lineage [3, 4]; few sequence homologues compared to host proteins.
- **No cross-kingdom pairing:** species-paired MSAs carry little inter-chain signal for pathogen-host pairs [5, 6].
- **Intrinsic disorder:** effectors had low per-residue confidence.

Discriminating true interactions

Confidence scores were evaluated as binary classifiers by **AUPR** (area under the precision-recall curve; random ≈ 0.09).



AUPR on the curated benchmark. Teal, pLM-based classifiers; green, structure prediction methods (Boltz2 confidence; ESMFold2 combined score). Random baseline ≈ 0.09.

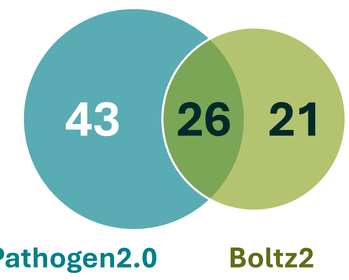
- **AraPathogen2.0**, trained on plant-pathogen data, was the most performant, even when accounting for data leakage.
- **Boltz2 and ESMFold2** best discriminating confidence metrics outperformed all other pLMs, though required more computation.
- **General pLM classifiers** were near random on both datasets.

An ensemble improves discrimination

No single method was consistently accurate, yet 20 of 21 experimentally determined complexes were modelled acceptably (DockQ ≥ 0.23) by at least one. We therefore combined the best structural (Boltz2) and pLM (AraPathogen2.0) methods.

They detect different interactions: of 90 true interactions in either method's top 10%, only 26 were captured by both.

Combing them improves accuracy: a noisy-OR combination of percentile ranks increased AUPR from **0.49 to 0.57**.



True interactions in each method's top 10%; 56 of 146 were recovered by neither.

Key takeaways

- Most existing frontier methods are not reliable predictors of effector-host interactions.
- These methods rely on either explicit (MSA) or implicit (learned) representations of co-evolutionary information from protein sequences [1, 3, 4].
- Effectors are often evolutionarily obscure [5, 6], thus coevolutionary information is sparse and poorly represented.
- The complementarity of methods can be exploited in an ensemble approach to increase discriminatory power, but this comes with a trade off of computational cost.

References

1. Kim AR et al. (2024) bioRxiv 2024.02.19.580970.
2. Luo Y et al. (2026) bioRxiv 2026.04.14.718427.
3. Jumper J et al. (2021) Nature 596(7873):583–89.
4. Lin Z et al. (2023) Science 379(6637):1123–30.
5. Müller MC et al. (2026) PLoS Pathog 22(1):e1013870.
6. Seong K & Krasileva KV. (2023) Nat Microbiol 8(1):174–87.

The support of the Grains Research and Development Corporation (GRDC) provided through the Analytics for the Australian Grains Industry (AAGI) project is acknowledged.