Binding free energies beyond curated benchmarks: what the 2026 OpenFE assessment tells us

A 2026 multi-company benchmark of OpenFE exposes the gap between public relative binding free-energy datasets and blinded industrial chemistry. We examine the numerical results, failure modes, emerging force-field and absolute-free-energy approaches, and practical standards for reliable molecular modelling.

PharmaToxAI · 2026-10-26

Predicting whether a chemical modification improves ligand affinity is one of the most demanding routine tasks in structure-based drug discovery. Molecular docking can propose poses and prioritise compounds quickly, but its scoring functions are not a thermodynamic measurement of binding. Molecular dynamics-based free-energy methods offer a more physically grounded alternative. Their use in medicinal chemistry has expanded as software, GPU hardware and automated preparation pipelines have improved.

A large collaborative assessment published in May 2026 provides an unusually useful reality check. Baumann and colleagues evaluated an open-source relative binding free-energy (RBFE) protocol from the **Open Free Energy (OpenFE)** ecosystem using data contributed by **15 pharmaceutical companies**, including public and blinded private datasets spanning more than **1,700 ligands** [1]. The comparison is particularly informative because the same general method performs substantially differently on familiar benchmark chemistry and on industrial datasets.

The central result is neither a blanket endorsement nor a failure of free-energy calculations. It is evidence that **chemical context, system preparation and the type of transformation being modelled strongly influence whether a calculation is decision-ready**. Recent work on neural-network potentials, absolute binding free energies and coarse-grained models makes that distinction increasingly important.

What a relative binding free energy actually estimates

For two ligands, A and B, binding to the same protein, an RBFE calculation estimates the difference between their binding free energies:

ΔΔGbind(A → B) = ΔGbind(B) − ΔGbind(A).

In a conventional alchemical thermodynamic cycle, A is transformed into B in the protein-bound state and separately in solution. The difference between those two transformations gives the relative binding free energy. The calculation samples a series of intermediate Hamiltonians rather than observing the ligand physically changing its chemical identity.

This construction has a practical advantage: many energetic contributions cancel when A and B are closely related and share a binding mode. It also explains why RBFE is most often applied to congeneric series during lead optimisation rather than to an arbitrary library of unrelated molecules.

At approximately 298 K, **1.36 kcal/mol corresponds to a tenfold difference in an equilibrium dissociation constant**, assuming the compared quantities refer to the same binding equilibrium and standard state. This conversion helps interpret energy errors, but it should not be used to claim that an RMSE directly represents a fixed fold-error for every compound. RMSE summarises a distribution of deviations; individual transformations may perform much better or worse.

For reliable RBFE, several assumptions must be credible: the ligand poses and protonation states are appropriate, the relevant protein conformations are sampled, the force field describes both end states and the intermediate path adequately, and the simulation achieves sufficient overlap and convergence. A well-behaved numerical estimator cannot compensate for an incorrect chemical state.

The 2026 OpenFE assessment: public and industrial performance diverge

The OpenFE study is notable for its scale and for evaluating a shared, openly available protocol across industrial contributors rather than reporting performance on a single target or a carefully selected ligand series [1].

On the **public dataset**, the weighted root-mean-square error across **58 systems** was **1.73 kcal/mol**, with a reported interval of **1.53–1.96 kcal/mol**. Ten systems achieved an RMSE below 1 kcal/mol.

On the **blinded private dataset**, the corresponding weighted RMSE across **37 systems** was **2.44 kcal/mol**, with a reported interval of **1.94–3.06 kcal/mol**. Only two systems achieved sub-kcal/mol accuracy [1].

These are system-aggregated statistics reported by the authors. They should not be confused with the accuracy of every individual perturbation, nor should the reported intervals be interpreted as uncertainty intervals for a new ligand prediction.

The gap is scientifically consequential. Public benchmarks often consist of well-characterised targets and ligand series for which structures, experimental affinities and simulation preparation have already been curated. Industrial programmes may contain less regular chemistry, multiple plausible poses, difficult substituent changes, uncertain experimental reference values or poorly characterised structural waters.

The private set is not simply a larger copy of the public set, so the difference cannot be attributed to one isolated factor. Nevertheless, it provides direct evidence against extrapolating a public-benchmark headline metric to an unrelated medicinal chemistry campaign.

The paper also distinguishes the performance of **simulated network edges** from that of **all pairwise differences reconstructed from the ligand free-energy network**. For the private data, the authors report an edgewise RMSE of approximately **1.93 kcal/mol** across 1,248 simulated edges, compared with approximately **2.41 kcal/mol** across 16,744 all-to-all pairwise comparisons [1]. These are different evaluation objects and should not be presented as competing estimates of the same statistic. Network inference, edge selection and error propagation matter.

Why transformation design and preparation cannot be delegated blindly

An RBFE network is a set of choices, not merely a set of ligands. Each edge connects a chemical transformation to two simulated environments. Poorly chosen edges can produce large perturbations, inadequate phase-space overlap or unnecessary computational cost.

Transformations involving a change in formal charge, ring closure or opening, large scaffold modifications, different tautomeric states or a substantial binding-pose rearrangement require particular care. The OpenFE study found system-dependent performance rather than a single error source dominating across all cases [1].

A 2025 preprint by Zou and colleagues describes the **Uni-FEP Benchmarks**, assembled from medicinal chemistry examples in ChEMBL. The authors report approximately **1,000 protein–ligand systems and 40,000 ligands**, with examples including scaffold changes and charge-changing transformations [2]. These figures describe the proposed benchmark resource, **not 40,000 independent experimental validations of a particular FEP method**. The work is a preprint, and its dataset construction and performance claims should be evaluated separately from the peer-reviewed OpenFE assessment.

For a practical workflow, a preparation record should preserve more than a SMILES string and a protein structure. It should document ligand stereochemistry, protonation and tautomeric assumptions, the source and quality of the receptor coordinates, missing residues, relevant cofactors and ions, crystallographic or predicted waters, initial poses, atom mapping, the transformation network and the exact force-field/protocol versions.

Experimental affinity data also need curation. An IC50 from a functional assay is not automatically interchangeable with a Kd, and apparent affinity can depend on assay conditions, substrate concentrations and protein construct. A model–experiment comparison is only meaningful when the reference endpoint and its provenance are clear.

Force-field accuracy: neural-network potentials offer progress, not immunity

Sampling is only one source of error. A calculation can converge reproducibly to the wrong free energy if the potential energy function misrepresents ligand strain, torsions or interactions.

In 2025, Sabanés Zariquiey and colleagues reported **QuantumBind-RBFE**, which combines a neural-network potential for ligand intramolecular energetics with molecular mechanics for the surrounding environment [3]. Their AceFF 1.0-based approach improved overall results relative to GAFF2 and an earlier ANI-2x implementation on the benchmark systems studied. Performance relative to OPLS4 varied by target; it was not uniformly superior.

The distinction between intramolecular and intermolecular treatment matters. In the reported NNP/MM scheme, the neural-network model does not replace the full protein–ligand interaction Hamiltonian with a learned quantum model. The surrounding protein, solvent and long-range interactions still depend on the classical description. The authors also identify limitations for uncommon ligand charge states and dependence on the diversity of the potential's training data [3].

A learned potential therefore introduces its own applicability question: **does the ligand chemistry lie within the region where the potential was trained and tested?** Better aggregate accuracy on a benchmark does not eliminate this requirement.

Absolute binding free energies are becoming more scalable

RBFE is efficient when ligands are related and a defensible perturbation network can be built. Absolute binding free-energy (ABFE) methods address a different problem:

estimating a ligand’s binding free energy without requiring a chemically close reference ligand.

Traditional ABFE protocols often use double decoupling and explicit restraints to account for the standard-state and binding-site geometry. Their computational cost and sensitivity to sampling, restraint selection, binding poses and charge treatment have limited large-scale use.

A March 2026 **preprint** by Liu and colleagues introduces **Felis**, an automated ABFE toolkit coupled to a data-driven ligand force field [4]. The authors benchmarked **859 ligands across 43 protein targets**, with an additional challenging KRAS(G12D) case study. They report ranking performance comparable to established RBFE approaches in the datasets examined, without target-specific tuning of the force field or alchemical schedule.

This is an important methodological development, but its evidence status is different from that of a peer-reviewed industrial benchmark. It is a developer-reported preprint evaluation, not an independent demonstration of universal ABFE accuracy or throughput.

A second 2026 peer-reviewed study, by Wingbermühle and colleagues, approached scale from a different direction [5]. Their workflows combine explicit-solvent molecular dynamics with **implicit-solvent free-energy estimates**. One workflow generated affinity-related estimates for more than **4,000 protein–ligand complexes**; another evaluated the local stability of more than **800,000 candidate binding poses**. The authors deliberately trade some thermodynamic rigour for throughput and consistent automated processing.

Those implicit-solvent estimates should not be conflated with converged, explicit-solvent double-decoupling ABFE. Their value lies in large-scale ranking and pose assessment under a stated approximation. The paper also illustrates a critical engineering point: structure repair, parameter assignment, automated rejection of unsuitable inputs and reproducible task scheduling are part of the scientific method, not just infrastructure details.

What a very recent coarse-grained study adds

A preprint submitted on **8 October 2026** by Walter and Bereau examines absolute binding free energies using the **Martini 3 coarse-grained force field** for **172 ligands across 15 targets** [6]. The authors report systematic underbinding relative to atomistic references, with target-dependent offsets of **2.7–8.7 kcal/mol**. A simple correction related to the number of charged binding-pocket residues reduced the reported mean absolute error in their leave-one-out analysis.

That numerical improvement needs careful interpretation. The study also reports a **median within-target Spearman correlation of 0.32**, with substantial variation across targets. A correction that reduces a target-level systematic offset does not necessarily solve ligand ranking within a medicinal chemistry series.

The authors attribute part of the behaviour to the coarse-grained representation's limited description of directional hydrogen bonding and the local electrostatic environment. This is a useful counterexample to the idea that inexpensive physical simulations must provide reliable affinity ranking once their systematic bias is corrected. The work remains a preprint and should be treated as a hypothesis-generating methodological comparison.

A defensible validation plan for a medicinal chemistry campaign

Before running hundreds of transformations, the first question should be whether RBFE is appropriate for the series. If the compounds share a binding mode and undergo relatively modest modifications, a connected RBFE network is often a reasonable starting point. If the series includes unrelated scaffolds or poorly defined poses, ABFE, alternative sampling methods or a different prioritisation strategy may be more appropriate.

A useful evaluation has three layers.

First, verify chemical and structural inputs. Confirm identity, stereochemistry, protonation and tautomeric alternatives, structural waters and relevant cofactors. Where multiple states are plausible, their contributions to the binding equilibrium must be addressed rather than silently selecting one.

Second, assess numerical reliability. Use independent replicas where feasible, examine convergence of each alchemical leg, estimator uncertainty, phase-space overlap and sensitivity to initial configurations. Network cycle closure can reveal inconsistencies, although a closed cycle does not prove agreement with experiment. Check whether failures concentrate in specific transformation types.

Third, test decision performance prospectively. Reserve a chemically meaningful subset of compounds or a later medicinal chemistry cycle for evaluation. Report absolute error and ranking metrics, uncertainty calibration where available, coverage and failure rates, and the fraction of proposed synthesis decisions that the calculations would actually have changed. Keep experimental assay uncertainty separate from computational uncertainty.

The computational budget should also be reported. A workflow that reaches the same nominal RMSE at ten times the GPU cost may have a different value in early discovery.

Conversely, a faster method that repeatedly fails on the transformations of greatest chemical interest may save little in practice.

A compact record for each prediction should include the ligand pair, receptor structure, atom mapping, charge/protonation states, force-field versions, alchemical protocol, replicas, sampling length, estimator, convergence diagnostics, predicted ΔΔG, numerical uncertainty, experimental reference and reason for any exclusion. That record makes an apparent success or failure auditable.

Implications for predictive toxicology and ADMET

Binding free energy is not a toxicity endpoint. A favourable ΔG for a putative off-target does not establish target engagement at an achievable exposure, and even confirmed engagement does not by itself establish an adverse outcome.

The relevant bridge to ADMET is mechanistic and conditional. When a particular receptor, enzyme or transporter has a defensible relationship to a safety question, free-energy calculations may help prioritise analogues for experimental testing. That interpretation still requires information about unbound concentration, tissue exposure, target biology, competing interactions and the quality of the structural model.

In this setting, a useful computational result is a **ranked, uncertainty-qualified hypothesis** linked to a defined biological question. It should not be converted into a generic toxicity score or presented as experimental validation.

Where the field stands in late 2026

The OpenFE collaboration changes the standard against which claims of free-energy accuracy should be read. Public datasets remain valuable for method development, but **blinded industrial chemistry exposes a larger and more variable error distribution**. The observed difference is substantial enough to affect how confidently a project can rank compounds whose predicted affinities differ by only a small amount.

Neural-network potentials are improving some force-field limitations. Automated ABFE workflows are expanding the chemical range that can be evaluated without congeneric reference ligands. Coarse-grained and implicit-solvent approaches are exploring the throughput–accuracy trade-off. None of these developments removes the need to inspect chemical states, sampling quality, applicability and experimental reference data.

For practical molecular modelling, the most useful advance may be the growing availability of **open protocols tested on chemically difficult, independently held datasets**. That makes it easier to ask a precise question: not whether a method is accurate in general, but whether its demonstrated performance is adequate for the transformation, target and decision at hand.

References

1. Baumann HM, Horton JT, Henry MM, Travitz A, Ries B, Gowers RJ, Swenson DWH, et al. **Large-Scale Collaborative Assessment of Binding Free Energy Calculations for Drug Discovery Using OpenFE.** *Journal of Chemical Information and Modeling*. 2026;66(11):6429–6452. DOI: 10.1021/acs.jcim.6c00089. Full text.

2. Zou R, Wang L, Wang X, Ding Y, Zheng H. **Breaking Barriers in FEP Benchmarking: A Large-Scale Dataset Reflecting Real-World Drug Discovery Challenges.** *ChemRxiv*. 2025. **Preprint.** DOI: 10.26434/chemrxiv-2025-rf8mf. Source dataset.

3. Sabanés Zariquiey F, Farr SE, Doerr S, De Fabritiis G. **QuantumBind-RBFE: Accurate Relative Binding Free Energy Calculations Using Neural Network Potentials.** *Journal of Chemical Information and Modeling*. 2025;65(8):4081–4089. DOI: 10.1021/acs.jcim.5c00033. Original article.

4. Liu Y, Wang A, Xia Y, Wang Z, Yan W. **Development and large-scale benchmarks of a protein–ligand absolute binding free energy toolkit.** *arXiv*. 2026;2603.22274. **Preprint.** Original manuscript.

5. Wingbermühle S, Biswas AD, Bonanni D, Shugaeva T, Gadioli D, Beránek J, et al. **Molecular Dynamics Workflows to Compute Large-Scale Sets of Absolute Binding Free Energies Aiding Drug Candidate and Binding Pose Selection.** *Journal of Chemical Theory and Computation*. 2026;22(10):5122–5142. DOI: 10.1021/acs.jctc.5c02127. Original article.

6. Walter LJ, Bereau T. **Predictable Accuracy of Martini 3 Absolute Protein-Ligand Binding Free Energies across 15 Targets.** *arXiv*. 2026;2610.11714. **Preprint submitted 8 October 2026.** Original manuscript.

7. Open Free Energy consortium. **OpenFE: open-source software for alchemical free-energy calculations.** Project repository and technical documentation. GitHub. **Software documentation, not an independent validation study.**

Online edition: https://pharmatoxai.ar/whitepapers/binding-free-energy-benchmarks-openfe-2026/