Supplementary Information for

# Sparse Supervision Turns Polar Pretraining into Transferable Molecular Response Fields

Gal Oren, Boris Fain and Michael Levitt

## Contents

**1 Scope and organization** **3**

**2 Response definition and reference construction** **3**
2.1 Absolute response-ESP scale . . . 5

**3 Site representation and reconstruction error floor** **5**

**4 Architecture, training and development controls** **6**
4.1 What the 48 training examples contain . . . 6

**5 Prospective water panels and prediction chronology** **7**
5.1 Panel-I identity selection . . . 7
5.2 Panel-I qualification gate and later development reuse . . . 9
5.3 Panel-III identity selection . . . 9

**6 Comparator definitions and observable coverage** **10**

**7 Complete prospective water-panel results** **12**

**8 Frozen non-water replacement geometries and contact audit** **16**
8.1 Frozen design and prediction chronology . . . 16
8.2 Response errors on the frozen geometry challenge . . . 16
8.3 Post hoc geometry audit and its implication . . . 17
8.4 Post-audit contact-geometry follow-up . . . 18

**9 Qualitative response-field examples** **19**

**10 Liquid-derived local bridge** **19**

**11 Post hoc model controls and extrapolation tests** **21**
11.1 Same-supervision AIMNet2 representation control . . . 21
11.2 Nested 1/3/6/12-water diagnostic . . . 22
11.3 Controlled removal of the global dipole branch . . . 22
11.4 Fragment-separation diagnostic . . . 23

**12 Post hoc held-out-water downstream response** **24**
12.1 Frozen pilot and confirmatory protocol . . . 24
12.2 Independent-chemistry breadth . . . 25
12.3 Exact global multipole hierarchy . . . 27
12.4 Spatial range and cumulative outer coupling . . . 28
12.5 Response fraction of energy and directionality . . . 29
12.6 Numerical checks and interpretation boundary . . . 29

**13 Numerical and physical-consistency checks** **32**

**14 External-probe map of the fixed response** **32**
14.1 Probe construction and response coupling . . . 32
14.2 Aggregation and uncertainty . . . 33
14.3 What changes with probe distance . . . 34
14.4 Source data . . . 35

**15 Independent post hoc comparison with ARROW** **37**

**16 Reproduction** **39**

# 1 Scope and organization

This Supplementary Information follows the questions in the main text. Sections 2–4 define the QM response, probe surface, site representation and 48-example training set. Sections 5–9 document prospective selection, chronology, comparator coverage and the water and non-water results. Sections 10–12 cover the liquid-derived bridge, controlled ablations, environment-size extension, fragment separation and held-out-water coupling. Sections 13–16 give numerical checks, the post hoc external-probe sweep, an independent polarizable-model comparison and reproduction instructions. Throughout, GLIDER's observable is the fixed-geometry, complex-minus-fragments response potential. The frozen coupling test uses that potential as one contribution to an external-water electrostatic interaction.

The public companion at `https://github.com/Scientific-Computing-Lab/GLIDER_AI` maps experiments to geometries, references where calculated, frozen predictions, summary tables and provenance. Figures here use those archived values. Historical development records do not replace the reported 48-example training manifest.

# 2 Response definition and reference construction

For a fixed complex geometry $X$, the full system $SW$, ghost-basis solute $S^{W\text{-ghost}}$ and ghost-basis water cluster $W^{S\text{-ghost}}$ use the same coordinates and complete complex-centred basis. The response density, potential and dipole are

$$\Delta\rho_{\mathrm{resp}}(\mathbf{r};X) = \rho_{SW}(\mathbf{r};X) - \rho_S^{W\text{-ghost}}(\mathbf{r};X) - \rho_W^{S\text{-ghost}}(\mathbf{r};X), \tag{1}$$

$$\Delta V_{\mathrm{resp}}(\mathbf{r};X) = V_{SW}(\mathbf{r};X) - V_S^{W\text{-ghost}}(\mathbf{r};X) - V_W^{S\text{-ghost}}(\mathbf{r};X), \tag{2}$$

$$\Delta\boldsymbol{\mu}_{\mathrm{resp}}(X) = \boldsymbol{\mu}_{SW}(X) - \boldsymbol{\mu}_S^{W\text{-ghost}}(X) - \boldsymbol{\mu}_W^{S\text{-ghost}}(X). \tag{3}$$

Here $W$ denotes the water environment in the original panels. The prospective non-water test uses the identical construction with the molecular neighbour $P$ in place of $W$. This is a mutual response of the full complex. It is not partitioned into a solute-only response. Because every term is evaluated in the same basis, basis-set superposition is treated consistently in the difference [Boys and Bernardi, 1970]. The nuclear coordinates and charges are identical in the full system and in the union of the fragments, so nuclear contributions cancel at fixed geometry. The remaining density and potential describe electronic redistribution.

This interaction-induced or deformation density is the net self-consistent electronic rearrangement on formation of the fixed complex. It is not a SAPT induction energy, a unique polarization or charge-transfer term, or an EDA partition [Jeziorski et al., 1994, Khaliullin et al., 2007]. GLIDER deliberately models the total spatial electrostatic consequence without assigning the redistribution to those non-unique components.

Reference calculations used restricted ωB97X-D3BJ/def2-TZVPD, numerical grid level 4 and SCF tolerance $10^{-10}$ [Chai and Head-Gordon, 2008, Grimme et al., 2010, Weigend and Ahlrichs, 2005]. Difficult calculations were continued only through the specified same-Hamiltonian numerical ladder. No functional, basis, grid, geometry or target definition was changed for an individual configuration, and no configuration was removed.

The deterministic probe construction used 26 directions per atom at 1.40 and 1.80 times element-specific van der Waals radii. Points within 1.05 radii of another atom were removed, coordinates were rounded to $10^{-5}$ Å and duplicates were deleted (Fig. S1). The response-ESP error for configuration $c$ is

$$e_c^V = \frac{\left[|P_c|^{-1}\sum_{p\in P_c}(\widehat{\Delta V}_{cp} - \Delta V_{cp})^2\right]^{1/2}}{\left[|P_c|^{-1}\sum_{p\in P_c}(\Delta V_{cp})^2\right]^{1/2}}. \tag{4}$$

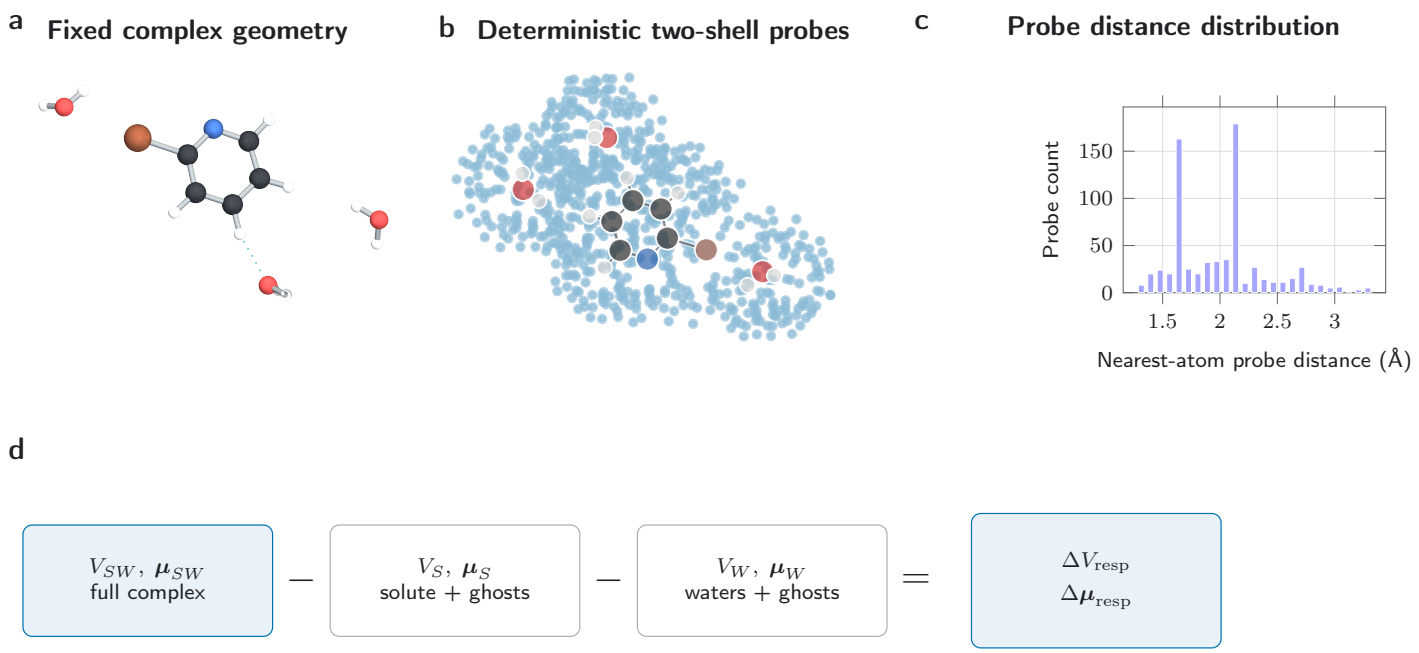



**Figure S1: Reference subtraction and deterministic response-ESP probe surface.** **a**, Actual complex geometry. **b**, The deterministic probe points, projected with the atoms for orientation. **c**, Distribution of nearest-atom probe distances. Probe points are generated on two Fibonacci shells about every atom, culled against neighbouring atoms and deduplicated. **d**, The complete complex and both ghost-basis fragments use identical coordinates and the complete basis. Their difference defines the mutual response. Reference and model predictions use the identical point list.

Table S1: **Panel-wide physical response-ESP scale.** Values are configuration RMS quantities averaged within molecule and then equally across molecules. All values are m$E_h$/e. Full-precision configuration and molecule tables are supplied as source data.

| Dataset | QM response RMS | GLIDER RMSE | MACE-POLAR-1-L RMSE |
|---|---|---|---|
| Panel I | 3.389 | 1.183 | 1.812 |
| Panel II | 3.214 | 1.191 | 1.841 |
| Panel III | 3.031 | 1.156 | 1.732 |
| Liquid-derived bridge | 3.282 | 0.988 | 1.681 |

Configuration errors are averaged within molecule and then equally across molecules. Dipole squared errors are averaged over Cartesian components and configurations within molecule before equal-molecule aggregation and the final square root. A zero-response predictor has $e_c^V = 1$ by construction. $1 - e_c^V$ is therefore not a fraction of the response captured or explained.

### 2.1 Absolute response-ESP scale

Panel-wide absolute quantities were computed directly from every stored QM, GLIDER and MACE-POLAR-1-L probe array, not by multiplying aggregate NRMSE by an average response scale. For each configuration we calculated the RMS QM response and RMS prediction error in m$E_h$/e, averaged configurations within molecule, and then weighted molecules equally. Table S1 therefore uses the same independent-unit convention as the primary endpoint.

The reference response RMS is about 3.0–3.4 m$E_h$/e across the three prospective panels and liquid-derived bridge. GLIDER absolute RMSE is 0.988–1.19 m$E_h$/e, compared with 1.68–1.84 m$E_h$/e for MACE-POLAR-1-L. Configuration- and molecule-level full-precision values, as well as point arrays for the displayed field examples, are included as source data. A response-to-total-ESP ratio is not reported because matched full-complex ESP arrays are not stored for every case. Reconstructing a different total-field endpoint solely for this comparison would not be equivalent to the response calculation.

## 3 Site representation and reconstruction error floor

The site representation contains scalar response charges $\Delta q_i$ and Cartesian response dipoles $\Delta\mathbf{u}_i$. The open-boundary potential is

$$\widehat{\Delta V}(\mathbf{r}) = \sum_i \left[ \frac{\Delta q_i}{|\mathbf{r} - \mathbf{R}_i|} + \frac{\Delta\mathbf{u}_i \cdot (\mathbf{r} - \mathbf{R}_i)}{|\mathbf{r} - \mathbf{R}_i|^3} \right], \tag{5}$$

with zero net response charge and molecular moment

$$\sum_i \Delta q_i = 0, \qquad \Delta\boldsymbol{\mu} = \sum_i \Delta q_i \mathbf{R}_i + \sum_i \Delta\mathbf{u}_i. \tag{6}$$

The atomic variables are not asserted to be unique quantum observables. A regularized constrained least-squares transform defines one reproducible gauge for supervision. The evaluated physical objects remain the response potential and dipole.

The direct site fit used ridge penalties $10^{-6}$ and $10^{-5}$ for charge and dipole columns after column normalization. A Karush–Kuhn–Tucker solve imposed zero net response charge and the three QM dipole components. Its mean development reconstruction error was 0.0309 NRMSE (Fig. S2), nearly an order of magnitude below the chemistry-blocked GLIDER predictive error of 0.2945. Thus the primary predictive discrepancy cannot be attributed to an inability of the chosen atom-centred basis to represent the development fields.

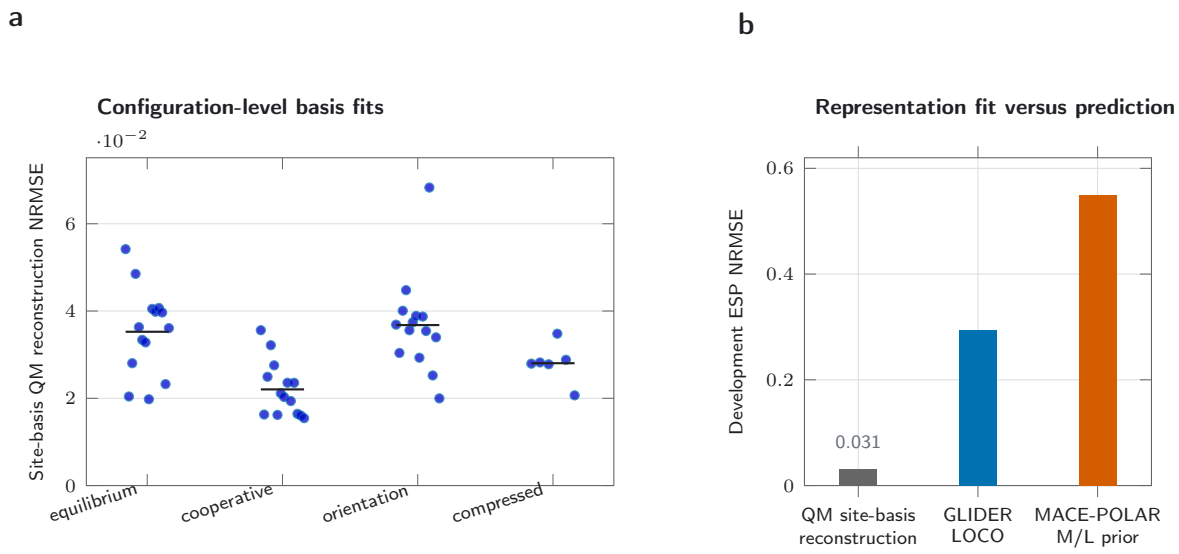



**Figure S2: The atom-centred monopole–dipole basis has a low reconstruction error floor. a**, The same constrained basis used for supervision was fit directly to each QM development response. Points show configuration-level ESP NRMSE. Horizontal lines are regime means. **b**, Mean site-basis reconstruction error compared with the GLIDER and MACE-POLAR prior prediction errors. The 0.0309 value is a representation fit, not a chemistry-transfer prediction. Its much smaller error shows that the site basis is not the dominant bottleneck on the development set.

# 4 Architecture, training and development controls

The frozen encoder is the product-basis representation from MACE-POLAR-1-M [Batatia et al., 2026]. An unfitted response prior is the equal-weight average of released MACE-POLAR-1-M and -1-L full-minus-fragments site responses. A local anisotropic head predicts corrections to response charges and dipoles. A separately trained extensive equivariant readout predicts the molecular response correction. The residual between the global readout and the local-site moment is distributed across atomic dipoles, giving exact local–global consistency without altering charge conservation.

## 4.1 What the 48 training examples contain

The original response-training set contains 14 solutes and 48 configurations (Table S2). Six stress-set solutes each contribute four three-water arrangements: equilibrium, cooperative, orientation and compressed. Eight development solutes each contribute three four-water arrangements: compact, cooperative and orientation. Thus $6 \times 4 + 8 \times 3 = 48$, split 24/24 by water count. Compact is the original motif name mapped to equilibrium in the combined registry. There is no compressed four-water example. Chemistry, motif and water count vary together in this design, which does not isolate their independent effects.

**Table S2:** The original 48 response-training configurations. C: compact, O: orientation, H: cooperative hydrogen bonding, X: compressed.

| Identifier | SMILES | Regimes | Waters | Cases |
|---|---|---|---|---|
| stress_acetamidine | CC(=N)N | C, O, H, X | 3 | 4 |
| stress_dimethyl_sulfone | CS(=O)(=O)C | C, O, H, X | 3 | 4 |
| stress_methanesulfonamide | CS(=O)(=O)N | C, O, H, X | 3 | 4 |
| stress_methyl_sulfoxide | CS(=O)C | C, O, H, X | 3 | 4 |
| stress_nitroethane | CC[N+](=O)[O-] | C, O, H, X | 3 | 4 |
| stress_pyridine_n_oxide | [O-][n+]1ccccc1 | C, O, H, X | 3 | 4 |
| dev_amino_alcohol_halogen | NCC(O)C(F)Cl | C, O, H | 4 | 3 |

**Table S2:** Original response-training configurations (continued).

| Identifier | SMILES | Regimes | Waters | Cases |
|---|---|---|---|---|
| dev_aza_aryl_ether | COc1ncccc1F | C, O, H | 4 | 3 |
| dev_cyano_ether | COCC#N | C, O, H | 4 | 3 |
| dev_cyclic_carbamate | O=C1N(C)CCOC1 | C, O, H | 4 | 3 |
| dev_hydroxy_nitrile | N#CCC(O)C | C, O, H | 4 | 3 |
| dev_oxazolidinone | O=C1OCCN1C | C, O, H | 4 | 3 |
| dev_sulfonyl_amine | CS(=O)(=O)N(C)C | C, O, H | 4 | 3 |
| dev_thio_amide | CSC(=O)N(C)C | C, O, H | 4 | 3 |

Geometries came from the first deterministically generated solute conformer using experimental-torsion knowledge distance geometry (ETKDG) followed by the Merck molecular force field (MMFF). Heteroatom and heavy-atom anchors placed the waters. Equilibrium or compact arrangements placed waters around the solute, cooperative arrangements formed an oriented chain, orientation cases rotated waters about their oxygen atoms, and compressed stress cases shortened an initial contact. These are constructed local configurations, not bulk-liquid frames. Coordinates, motif labels, generator sources and fixed seeds identify every example in the release.

One response label consists of a QM potential on 589–916 retained probe points and a three-component response dipole. The 35,840 points sample 48 fields, rather than forming that many independent examples. Training used a deterministic selection of 384 points per configuration and evaluation used the full surface. The auxiliary atomic site targets derive from the constrained fit to these observables. No energy or force labels trained the response head. A historical development pool also includes the later 48 Panel-I cases, but those cases never entered the published GLIDER training inputs.

All architecture and loss choices used chemistry-blocked development. The selected loss was $L = L_V + 2L_\mu + 0.05L_{\text{site}}$. AdamW used learning rate $2 \times 10^{-3}$, weight decay $2 \times 10^{-5}$, cosine decay, gradient clipping at 5 and 64 fixed epochs. Five ensemble members used seeds 2026081300–2026081304. Encoder weights remained frozen.

In leave-one-chemistry-out evaluation, GLIDER gave 0.2945 ESP NRMSE and 0.06547 D dipole RMSE. The unfitted M/L prior gave 0.5494 and 0.0851 D. The minimal and richer local site heads gave 0.5326/0.2368 D and 0.5255/0.2387 D. Identical-feature ridge and the same-data symmetry-kernel control gave 1.173/0.6571 D and 1.157/0.5001 D (Fig. S3). These controls establish that labels plus a generic readout did not reproduce the selected local–global response mapping.

# 5 Prospective water panels and prediction chronology

The three prospective panels use the same four physical regimes: equilibrium, cooperative, orientation and compressed. Panel I contains 12 molecules and 48 configurations. Panel II contains 24 and 96. Panel III contains 20 and 80. All prospective chemistries are absent from GLIDER response supervision. This statement does not imply absence from broad foundation-model pretraining.

For each panel, the model and protocol were fixed before target-free panel construction, prediction recording, reference acquisition and one-time scoring. Some public comparators were added to Panel I after the initial study to complete the comparison set, but used unmodified checkpoints without target fitting. All eligible comparator predictions in Panels II and III preceded reference QM. The complete chronology and content hashes are retained in the release.

## 5.1 Panel-I identity selection

Panel I joins a pre-existing sealed list of seven neutral polyfunctional organic solutes to five identities selected after the GLIDER freeze. The seven came from an earlier many-body study's prospective identity list and span mixed donor/acceptor, cyclic, sulfur-containing and halogenated chemistries. They are a designed set rather than a random sample from a molecular database. The retained overlap audit found

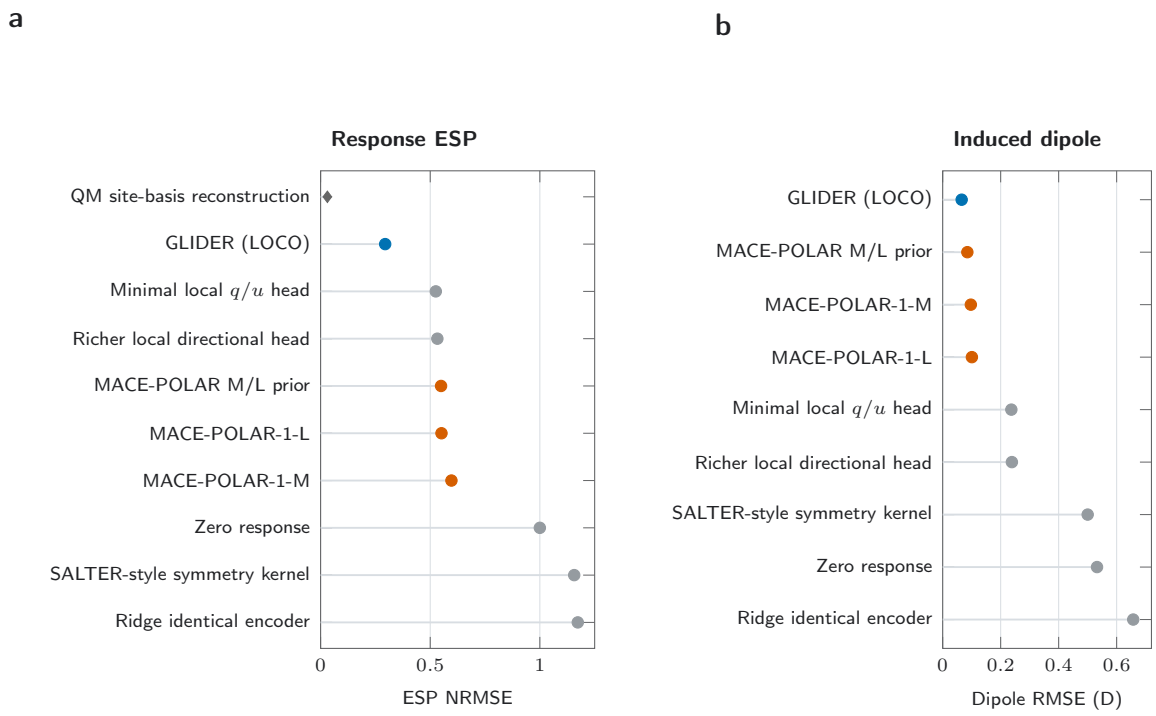



**Figure S3: Complete chemistry-blocked development controls.** ESP NRMSE and induced-dipole RMSE for the unfitted prior, minimal and richer local heads, identical-feature ridge regression, a same-data symmetry-kernel control and the selected local–global model. The QM site-basis reconstruction is shown only as a representation error floor. Every predictive value is from leave-one-chemistry-out evaluation. Lower is better.

no exact-connectivity match with its 40 historical solutes and nearest radius-2 Morgan similarities of 0.143–0.469.

The five further identities came from a fixed candidate pool by max–min distance: 0.6 Morgan/Tanimoto distance plus 0.4 robustly normalized Mordred descriptor distance. The selector took one from each of four predeclared strata and then the most diverse remaining candidate. Original training identities and the sealed seven were novelty anchors. Selection used structure, not response labels or model errors. New three-water geometries and GLIDER predictions for all twelve solutes were recorded before response references were calculated.

### 5.2 Panel-I qualification gate and later development reuse

Before selection, a selector type-casting defect was corrected and recorded while no response labels existed. It did not change a selected or scored panel. The separate original qualification gate required simultaneous improvements in response-potential and induced-dipole accuracy, uncertainty, robustness, physical consistency and runtime. In its archived evaluation, ESP NRMSE 0.3218 exceeded the 0.30 cutoff. Dipole RMSE improved only 3.64% against MACE-M, below the required 15%, and its paired 95% interval $[-0.0608, +0.0296]$ D crossed zero. The broader joint gate therefore failed despite the spatial-field improvement. All twelve solutes, geometries and predictions remained in the benchmark, and the reported GLIDER checkpoint was not refitted.

After labels were open, Panel I entered a separate development pool for a later global-prior model, Family C. That model was evaluated separately on Panel II and never replaced the published GLIDER checkpoint. Panel-I labels did not fit that checkpoint or determine its architecture, loss or stopping rule. Later controls used the panel for evaluation only where stated. The release retains the gate rule, decision and later-pool provenance.

### 5.3 Panel-III identity selection

Panel III used FreeSolv v0.52 commit `6c7d19b4b565537365ffd22006aa2cd4643200c6` only as an identity source [Mobley and Guthrie, 2014]. The identity-selection code consumed a sanitized file containing only identifiers and SMILES. Hydration values were neither copied into that file nor used in selection or response scoring. Eligibility required a neutral, closed-shell, single-fragment molecule with 4–15 heavy atoms and elements H, C, N, O, F, P, S, Cl, Br or I. Exact previous matches and candidates with radius-2 Morgan similarity $\geq 0.70$ to any previous GLIDER chemistry were excluded. A prespecified max–min rule selected 20 molecules across five chemical strata. Table S3 lists the identities in all three panels.

**Table S3:** Prospective molecular identities.

| Panel | Identifier | SMILES |
|---|---|---|
| I | pros_azaaryl_sulfoxide | COc1nc(S(=O)C)cc(F)n1 |
| I | pros_cyano_cyclic_acetal | COC1(C#N)CC(O)C1Cl |
| I | pros_halogen_oxazinanone | O=C1N(CC#N)C(F)(Cl)CO1 |
| I | pros_heteroaryl_amide | O=C(N(C)C(F)Cl)c1nocc1 |
| I | pros_isoxazole_thioether | CSC1=NOC(C(O)F)=C1Cl |
| I | pros_polyfunctional_oxane | N#CC1(O)COC(CCl)C1F |
| I | pros_sulfonyl_piperidinol | CS(=O)(=O)N1CC(F)C(O)C1 |
| I | prospective_diverse_01 | NCC(=O)N |
| I | prospective_diverse_02 | OCCI |
| I | prospective_diverse_03 | O1CCOCC1 |
| I | prospective_diverse_04 | c1nccs1 |
| I | prospective_diverse_05 | O=S1(=O)CCCC1 |
| II | nextgen_prospective_001 | FCCBr |
| II | nextgen_prospective_002 | Oc1ccc(Cl)cc1 |
| II | nextgen_prospective_003 | Brc1ccccn1 |
| II | nextgen_prospective_004 | CCI |
| II | nextgen_prospective_005 | NC(=O)C(F)(F)F |
| II | nextgen_prospective_006 | OCC(F)F |

| Panel | Identifier | SMILES |
|---|---|---|
| II | nextgen_prospective_007 | O=C1NCCN1 |
| II | nextgen_prospective_008 | CNN |
| II | nextgen_prospective_009 | NCC(O)CO |
| II | nextgen_prospective_010 | Nc1ccccn1 |
| II | nextgen_prospective_011 | O=C1COCCN1 |
| II | nextgen_prospective_012 | CC(O)CN |
| II | nextgen_prospective_013 | COP(=O)(OC)OC |
| II | nextgen_prospective_014 | COC(=O)C(=O)OC |
| II | nextgen_prospective_015 | O=Cc1ccco1 |
| II | nextgen_prospective_016 | CC(=O)COC(C)=O |
| II | nextgen_prospective_017 | CNC=O |
| II | nextgen_prospective_018 | n1ncccc1 |
| II | nextgen_prospective_019 | CSC |
| II | nextgen_prospective_020 | CSc1ccccc1 |
| II | nextgen_prospective_021 | NCCS |
| II | nextgen_prospective_022 | NC(N)=S |
| II | nextgen_prospective_023 | O=S1(=O)CCNCC1 |
| II | nextgen_prospective_024 | CCS(=O)(=O)C |
| III | mobley_1278715 | ClC(Cl)(Cl)C(Cl)(Cl)Cl |
| III | mobley_1352110 | CCCCN(CC)C(=O)SCCC |
| III | mobley_1723043 | FC1(F)C(F)(F)C(F)(F)C1(F)F |
| III | mobley_172879 | c1ccc2c(c1)Oc1ccccc1O2 |
| III | mobley_2261979 | c1cc2c3c(cccc3c1)CC2 |
| III | mobley_2689721 | C1CCCCC1 |
| III | mobley_2727678 | O=c1[nH]cc(I)c(=O)[nH]1 |
| III | mobley_3047364 | CCNc1nc(NC(C)C)nc(SC)n1 |
| III | mobley_3709920 | C=CCCCCCCC |
| III | mobley_3777264 | Nc1cccc([N+](=O)[O-])c1 |
| III | mobley_4479135 | C=C(c1ccccc1)c1ccccc1 |
| III | mobley_4934872 | CCOC(OCC)Oc1ccccc1 |
| III | mobley_5200358 | CNC(=O)O/N=C\C(C)(C)SC |
| III | mobley_6239320 | O=c1[nH]c(=O)[nH]c(=O)[nH]1 |
| III | mobley_7578802 | BrC(Br)Br |
| III | mobley_8337722 | CC(C)[C@@H]1CC[C@@H](C)C[C@H]1O |
| III | mobley_9007496 | CNC(=O)Oc1cccc2ccccc12 |
| III | mobley_9460824 | CCOP(=S)(OCC)SCSCC |
| III | mobley_9534740 | OC[C@H]1O[C@@H](O)[C@H](O)[C@@H](O)[C@@H]1O |
| III | mobley_9741965 | C[C@@H](CCO[N+](=O)[O-])O[N+](=O)[O-] |

## 6 Comparator definitions and observable coverage

A matched comparator had to provide a public release, cover the required chemistry without system-specific fitting, and expose an observable from which the complex-minus-fragments response could be formed faithfully. The comparator review was closed on 16 August 2026. MACE-POLAR-1-S/M/L and AIMNet2/AIMNet2-2025 entered both ESP and dipole rankings. MACE-MDP and GFN2-xTB entered only the dipole ranking. Density, homogeneous-field and system-specific models without a broad applicable checkpoint are discussed as related work. They were not assigned proxy benchmark scores because they do not expose the required observable under the stated comparison conditions.

These public models are zero-target-fit baselines: they test how much of the explicit-environment response is already exposed by broad pretrained models. They are not presented as a same-supervision architecture comparison. The chemistry-blocked ridge, symmetry-kernel and local-head controls instead test whether direct response labels plus a generic small readout are sufficient.

The related-work review included static and total density learning, including total densities of interacting non-covalent complexes [Grisafi et al., 2019, Jørgensen and Bhowmik, 2022, Fabrizio et al., 2019]. Learned environment-dependent multipoles and Wannier centres [Darley et al., 2008, Zhang et al., 2020]. FieldSchNet and electrostatic embedding [Gastegger et al., 2021, Zinovjev, 2023], self-consistent and perturbed response potentials [Gao and Remsing, 2022, Joll et al., 2024], SALTER's transferable real-space density response to homogeneous electric fields and related response models [Lewis et al., 2023,

Shao et al., 2022, Feng et al., 2025], deformation-density descriptors for intermolecular interaction learning [Low et al., 2022, 2023], evidence that interaction-induced density polarization changes intermolecular electrostatic potentials and energetics [Khabibrakhmanov et al., 2025], latent electrical response inferred from energies and forces [Zhong et al., 2025], distilled foundation-model electrostatics [Wang and Cheng, 2026], polarizable atomic multipoles [Park et al., 2026], and the NepoIP/MM polarizable embedding model [Song and Yang, 2025]. These are scientifically relevant precedents, but none provided a broad released checkpoint that both covered the panel chemistry and exposed the counterpoise-consistent complex-minus-fragments response ESP without system-specific fitting. They were therefore retained as conceptual related work rather than converted into proxy benchmark bars.

These precedents overlap with different parts of the present question without sharing its target. A homogeneous-field density response differs from the heterogeneous perturbation created by a finite molecular neighbour. EDDIE-ML uses deformation-density information as a physically informative descriptor for interaction energies, whereas GLIDER predicts the explicit-neighbour-induced response field from geometry. The present observable is specifically the counterpoise-consistent full-complex-minus-ghost-fragments electronic response.

Fabrizio *et al.* is the closest total-density precedent in this list: it learns the electron density of an interacting non-covalent system. The present target instead removes the two ghost-basis fragment densities from the full complex and evaluates only the resulting electronic change. Neither target is a surrogate for the other.

MACE-POLAR ESP used the released Gaussian monopole/dipole density with width 1.5 Å, not an unsmeared point-multipole approximation. The analytic open-boundary decoder agreed with the official Fourier-space implementation, after removal of its boundary-condition constant, within $1.2 \times 10^{-7}$ Hartree/e. AIMNet2 used official environment-dependent charges. Fresh calculator instances prevented cached neighbour-state contamination between complex and fragment calls [Anstine et al., 2025]. AIMNet2-2025 used the official v0.2.0 alias and prespecified default ensemble member 0. MACE-MDP used the released SPICE-$\alpha$ checkpoint [Gonnheimer et al., 2026]. GFN2-xTB used `tblite` 0.7.0 [Bannwarth et al., 2019]. Table S4 records each release, observable and checkpoint provenance.

**Table S4:** Comparator releases, target-fitting status, observables and provenance.

| Method | Release | Class | Observable | Checkpoint SHA-256 |
|---|---|---|---|---|
| GLIDER | reported response ensemble | Response supervised | site response charges/dipoles; ESP and dipole | `a288afa285128e13cb7a05ba9459f1dce07e79b8638c6e272c2509ae68d7a288` |
| MACE-POLAR-1-S | mace_polar_1 | Zero-target-fit ESP/dipole | official Gaussian multipole density; ESP and dipole | `e4495612037b3b3312633182882a38a694ecac9ea0be2b9889ac0b2a84a99510` |
| MACE-POLAR-1-M | mace_polar_1 | Zero-target-fit ESP/dipole | official Gaussian multipole density; ESP and dipole | `fab8b8713c832f31a2a853aaa22fd638be8a369cbf5095e6b3e982a18d10e93a` |
| MACE-POLAR-1-L | mace_polar_1 | Zero-target-fit ESP/dipole | official Gaussian multipole density; ESP and dipole | `9f65f8dc6ddaff1d631e299cb531376a7da5e68d1bef04f34a2d5073d5ef114b` |
| AIMNet2 | aimnetcentral v0.2.0; official alias member 0 | Zero-target-fit ESP/dipole | native atomic charges; point-charge ESP and dipole | `f0f7c054539ad3261bd36f9b11c56d12f87cb723e25bea7521755bbd3ec24e28` |
| AIMNet2-2025 | aimnetcentral v0.2.0; official alias member 0 | Zero-target-fit ESP/dipole | native atomic charges; point-charge ESP and dipole | `043ed5418a104e31f79462f8e5ebeca64a2d24422174f5d29f894d32271981b5` |

| Method | Release | Class | Observable | Checkpoint SHA-256 |
|---|---|---|---|---|
| MACE-MDP | official MACE-MDP.model | Zero-target-fit dipole | official total molecular dipole; dipole only | `126f8d1602549e6fa0df775c701a5119ddeb0e3738202af8e7aa736de6c2b692` |
| GFN2-xTB | tblite 0.7.0 | Zero-target-fit dipole | self-consistent total molecular dipole; dipole only | not applicable; fixed semiempirical method |

# 7 Complete prospective water-panel results

Table S5 gives the complete aggregate leaderboard. The exact per-molecule scores appear in Table S6, and Table S7 separates the four placement regimes.

**Table S5:** Complete aggregate response leaderboard.

| Dataset | Method | ESP NRMSE | Dipole RMSE (D) |
|---|---|---|---|
| Panel I | GLIDER | 0.3218 | 0.1894 |
| Panel I | MACE-POLAR-1-L | 0.5245 | 0.2156 |
| Panel I | MACE-POLAR-1-M | 0.6288 | 0.1965 |
| Panel I | MACE-POLAR-1-S | 0.7621 | 0.2752 |
| Panel I | AIMNet2 | 0.7668 | 0.3629 |
| Panel I | AIMNet2-2025 | 0.8951 | 0.5698 |
| Panel I | MACE-MDP | – | 0.1509 |
| Panel I | GFN2-xTB | – | 0.5438 |
| Panel II | GLIDER | 0.3422 | 0.1125 |
| Panel II | MACE-POLAR-1-L | 0.5710 | 0.1129 |
| Panel II | MACE-POLAR-1-M | 0.6655 | 0.1354 |
| Panel II | MACE-POLAR-1-S | 0.7346 | 0.2450 |
| Panel II | AIMNet2 | 0.8430 | 0.2998 |
| Panel II | AIMNet2-2025 | 0.9045 | 0.3922 |
| Panel II | GFN2-xTB | – | 0.3939 |
| Panel II | MACE-MDP | – | 0.8484 |
| Panel III | GLIDER | 0.4266 | 0.2267 |
| Panel III | MACE-POLAR-1-L | 0.6297 | 0.3560 |
| Panel III | MACE-POLAR-1-M | 0.7091 | 0.3063 |
| Panel III | AIMNet2 | 0.7704 | 0.3140 |
| Panel III | MACE-POLAR-1-S | 0.7857 | 0.2569 |
| Panel III | AIMNet2-2025 | 0.8967 | 0.4823 |
| Panel III | MACE-MDP | – | 0.2463 |
| Panel III | GFN2-xTB | – | 0.4409 |
| Liquid bridge | GLIDER | 0.3449 | 0.0601 |
| Liquid bridge | MACE-POLAR-1-L | 0.5825 | 0.0771 |
| Liquid bridge | MACE-POLAR-1-M | 0.6177 | 0.0777 |
| Liquid bridge | MACE-POLAR-1-S | 0.6828 | 0.1839 |
| Liquid bridge | AIMNet2 | 0.8377 | 0.1548 |
| Liquid bridge | AIMNet2-2025 | 1.0098 | 0.3169 |
| Liquid bridge | GFN2-xTB | – | 0.1924 |
| Liquid bridge | MACE-MDP | – | 1.9090 |

Figure S4 shows the molecule-level errors behind these panel aggregates, including the retained difficult Panel-III case.

**Table S6:** Per-molecule prospective response-ESP comparison.

| Panel | Molecule | GLIDER | MACE-POLAR-1-L | Difference |
|---|---|---|---|---|
| I | pros_azaaryl_sulfoxide | 0.3156 | 0.4872 | -0.1716 |
| I | pros_cyano_cyclic_acetal | 0.4000 | 0.4854 | -0.0854 |
| I | pros_halogen_oxazinanone | 0.2451 | 0.4777 | -0.2326 |
| I | pros_heteroaryl_amide | 0.3546 | 0.5052 | -0.1507 |
| I | pros_isoxazole_thioether | 0.3598 | 0.4946 | -0.1348 |
| I | pros_polyfunctional_oxane | 0.2382 | 0.5053 | -0.2671 |
| I | pros_sulfonyl_piperidinol | 0.3249 | 0.4189 | -0.0940 |
| I | prospective_diverse_01 | 0.2618 | 0.5833 | -0.3215 |
| I | prospective_diverse_02 | 0.7442 | 0.7686 | -0.0245 |

| Panel | Molecule | GLIDER | MACE-POLAR-1-L | Difference |
|---|---|---|---|---|
| I | prospective_diverse_03 | 0.2181 | 0.5414 | -0.3233 |
| I | prospective_diverse_04 | 0.2393 | 0.5567 | -0.3174 |
| I | prospective_diverse_05 | 0.1601 | 0.4694 | -0.3093 |
| II | nextgen_prospective_001 | 0.3962 | 0.4515 | -0.0552 |
| II | nextgen_prospective_002 | 0.3778 | 0.5492 | -0.1713 |
| II | nextgen_prospective_003 | 0.5054 | 0.7032 | -0.1978 |
| II | nextgen_prospective_004 | 0.9096 | 0.8583 | 0.0512 |
| II | nextgen_prospective_005 | 0.3379 | 0.5506 | -0.2127 |
| II | nextgen_prospective_006 | 0.3202 | 0.5146 | -0.1944 |
| II | nextgen_prospective_007 | 0.2803 | 0.5839 | -0.3036 |
| II | nextgen_prospective_008 | 0.2607 | 0.5338 | -0.2731 |
| II | nextgen_prospective_009 | 0.2442 | 0.5745 | -0.3304 |
| II | nextgen_prospective_010 | 0.1563 | 0.4959 | -0.3396 |
| II | nextgen_prospective_011 | 0.2255 | 0.4721 | -0.2466 |
| II | nextgen_prospective_012 | 0.2807 | 0.5242 | -0.2436 |
| II | nextgen_prospective_013 | 0.3125 | 0.6310 | -0.3185 |
| II | nextgen_prospective_014 | 0.2894 | 0.6000 | -0.3105 |
| II | nextgen_prospective_015 | 0.2733 | 0.5424 | -0.2691 |
| II | nextgen_prospective_016 | 0.3019 | 0.5187 | -0.2168 |
| II | nextgen_prospective_017 | 0.3322 | 0.6463 | -0.3141 |
| II | nextgen_prospective_018 | 0.3239 | 0.4381 | -0.1143 |
| II | nextgen_prospective_019 | 0.3448 | 0.5320 | -0.1872 |
| II | nextgen_prospective_020 | 0.3351 | 0.5893 | -0.2541 |
| II | nextgen_prospective_021 | 0.3679 | 0.5849 | -0.2170 |
| II | nextgen_prospective_022 | 0.6094 | 0.7721 | -0.1626 |
| II | nextgen_prospective_023 | 0.2605 | 0.5682 | -0.3077 |
| II | nextgen_prospective_024 | 0.1666 | 0.4694 | -0.3027 |
| III | mobley_1278715 | 0.7184 | 0.8394 | -0.1210 |
| III | mobley_1352110 | 0.2469 | 0.5201 | -0.2732 |
| III | mobley_1723043 | 0.4427 | 0.4759 | -0.0332 |
| III | mobley_172879 | 0.2855 | 0.5308 | -0.2453 |
| III | mobley_2261979 | 0.1991 | 0.5056 | -0.3064 |
| III | mobley_2689721 | 0.1796 | 0.5673 | -0.3876 |
| III | mobley_2727678 | 0.6678 | 0.7839 | -0.1161 |
| III | mobley_3047364 | 2.3079 | 2.2039 | 0.1040 |
| III | mobley_3709920 | 0.2525 | 0.5266 | -0.2741 |
| III | mobley_3777264 | 0.2294 | 0.4883 | -0.2589 |
| III | mobley_4479135 | 0.3349 | 0.5207 | -0.1858 |
| III | mobley_4934872 | 0.2308 | 0.4827 | -0.2519 |
| III | mobley_5200358 | 0.3717 | 0.5362 | -0.1645 |
| III | mobley_6239320 | 0.2831 | 0.5645 | -0.2814 |
| III | mobley_7578802 | 0.4134 | 0.5547 | -0.1414 |
| III | mobley_8337722 | 0.1800 | 0.4980 | -0.3180 |
| III | mobley_9007496 | 0.2237 | 0.4419 | -0.2182 |
| III | mobley_9460824 | 0.3577 | 0.5298 | -0.1721 |
| III | mobley_9534740 | 0.2424 | 0.4988 | -0.2564 |
| III | mobley_9741965 | 0.3640 | 0.5252 | -0.1612 |

For the primary comparison, paired 95% molecule-blocked intervals for GLIDER minus MACE-POLAR-1-L were $[-0.259, -0.145]$ in Panel I, $[-0.263, -0.190]$ in Panel II and $[-0.247, -0.153]$ in Panel III. The corresponding molecule wins were 12/12, 23/24 and 19/20. The observed ESP ranking and paired intervals therefore agree across all three prospective panels. The dipole ranking varies and its closest comparisons remain unresolved in each prospective panel. Figure S5 shows the resampling distributions and leave-one-molecule sensitivity behind the spatial comparison.

Every Panel-III leave-one-molecule-out ESP difference remained favourable, from $-0.219$ to $-0.193$. The conclusion is therefore not contingent on omitting the difficult `mobley_3047364` case. Chemistry groups overlap and serve only as diagnostics. All examined group means (oxygen-centred, N-rich/heteroaromatic, sulfur/phosphorus, halogenated, Br/I, weakly polar/nonpolar and strong hydrogen-bonding) favoured GLIDER. The same was true in all four physical regimes in each panel.

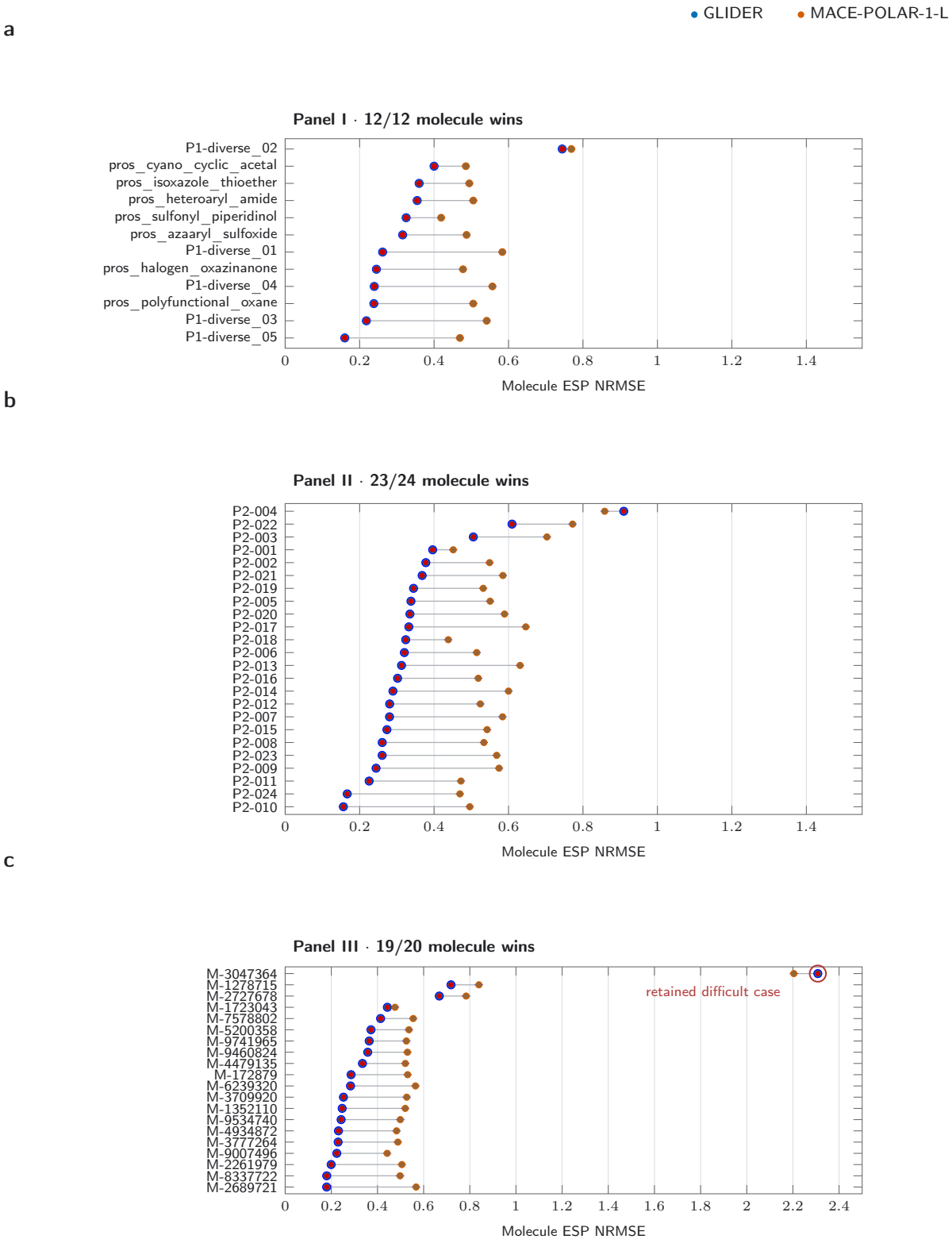



**Figure S4: Per-molecule response-ESP comparison on all prospective panels.** Each row is a molecule mean over four environments. Connected blue and orange points show GLIDER and MACE-POLAR-1-L, respectively. Panel titles report the number of molecules for which GLIDER has lower error. All 56 molecules are shown and no molecule was removed. The circled Panel-III point `mobley_3047364` is the largest GLIDER error and remains in the aggregate.

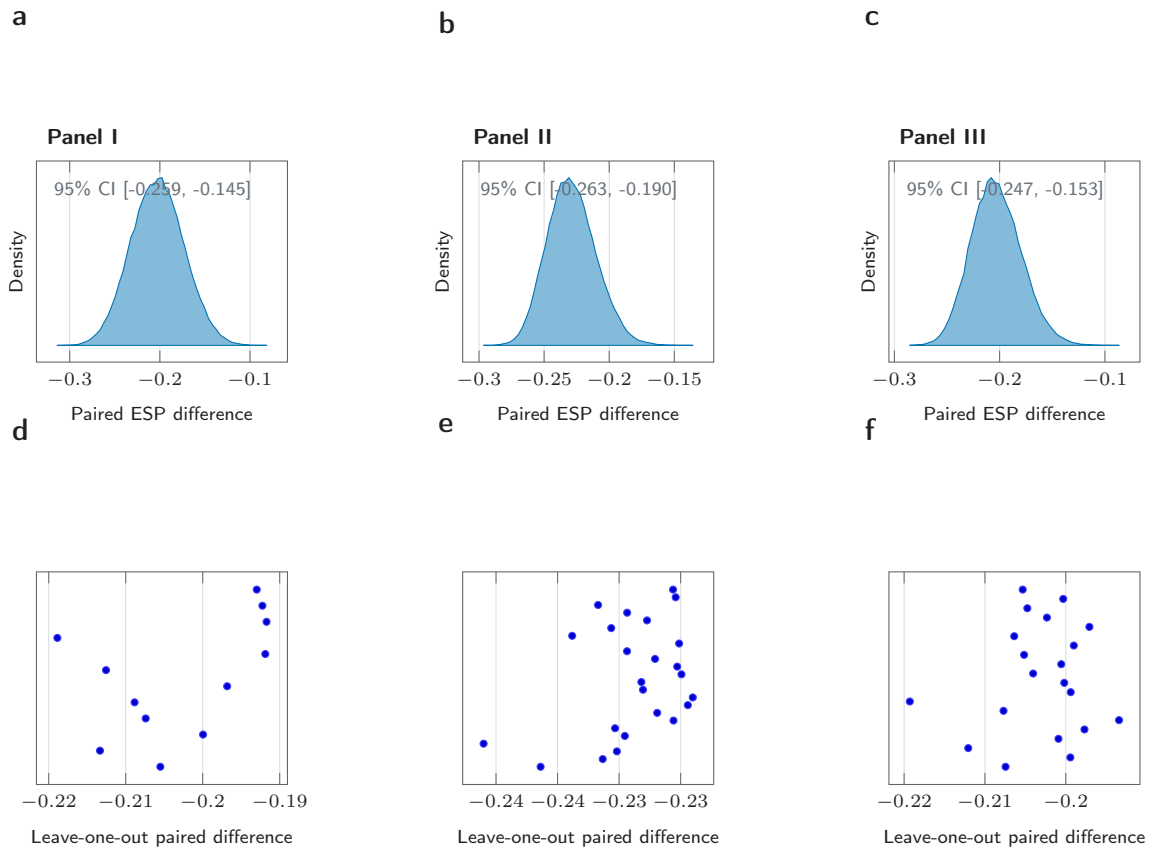



**Figure S5: Bootstrap and leave-one-molecule sensitivity of the primary prospective comparison.** **a**–**c**, Distributions of GLIDER minus MACE-POLAR-1-L aggregate response-ESP NRMSE from 100,000 molecule-blocked resamples. Negative values favour GLIDER. **d**, Leave-one-molecule-out differences for Panel I. **e,f**, the corresponding Panel-II and Panel-III diagnostics. These values diagnose dependence on individual molecules and do not replace the full-panel endpoint.

**Table S7:** Prospective response-ESP results by physical regime.

| Dataset | Regime | GLIDER | MACE-POLAR-1-L | Difference |
|---|---|---|---|---|
| prospective_1 | compressed | 0.2527 | 0.4978 | -0.2451 |
| prospective_1 | cooperative | 0.2797 | 0.5264 | -0.2467 |
| prospective_1 | equilibrium | 0.3144 | 0.5120 | -0.1976 |
| prospective_1 | orientation | 0.4403 | 0.5616 | -0.1214 |
| prospective_2 | compressed | 0.3512 | 0.5820 | -0.2308 |
| prospective_2 | cooperative | 0.2925 | 0.5539 | -0.2614 |
| prospective_2 | equilibrium | 0.3483 | 0.5560 | -0.2077 |
| prospective_2 | orientation | 0.3767 | 0.5920 | -0.2154 |
| prospective_3_canonical | compressed | 0.2892 | 0.5047 | -0.2155 |
| prospective_3_canonical | cooperative | 0.7045 | 0.8875 | -0.1830 |
| prospective_3_canonical | equilibrium | 0.3345 | 0.5112 | -0.1768 |
| prospective_3_canonical | orientation | 0.3781 | 0.6154 | -0.2373 |
| liquid_derived_bridge | liquid | 0.3449 | 0.5825 | -0.2377 |

# 8 Frozen non-water replacement geometries and contact audit

## 8.1 Frozen design and prediction chronology

This predeclared challenge applies the final checkpoint, response-specialized on solute–water environments, to three other neighbour species without retraining. All 12 Panel-I solutes were retained, avoiding a performance-informed solute subset. Each was paired with $NH_3$, $CH_3OH$ and $CH_3CN$ in two deterministic placements, giving 72 complexes. The panel tests the model on those fixed replacement geometries; a later contact audit showed that most should not be interpreted as ordinary attractive complexes.

The source Panel-I equilibrium and orientation configurations supplied the two geometries. In each source frame, the water oxygen nearest any solute atom was selected with atom index as the tie-break. The perturbant anchor (nitrogen for $NH_3$, oxygen for $CH_3OH$ and nitrile nitrogen for $CH_3CN$) was placed at that oxygen and aligned to the ordered local O–H frame. All source waters were then removed. A fixed list of 24 proper rotations provided a geometry-only clash fallback. Sixty-nine cases used the first orientation and three used the second. Its permissive clearance limits were 0.72 Å for the shortest interfragment distance and 0.62 for the shortest covalent-radius ratio. The new neighbour was not relaxed, and no interaction-energy screen was applied. No energy, response label or model output entered placement.

The complete design, geometries, metrics, bootstrap, failure rules and interpretation categories were committed at `103f0599ce796cc002ea7dbc88d9a4b944a7d12e`. GLIDER and every eligible comparator were evaluated next, and 617 prediction/probe files were hashed and committed at `d85de17d2a627508edf85a46defeb865b5332e14`. A remote attestation at `d7ba1f76124c5858a3b29100f4a2510992d9c43b` recorded that the prediction commit was present on GitHub and that no response-reference directory existed. Only then were the three counterpoise-consistent reference calculations run for each complex. All 72 cases converged. Six iodine cases used the fixed same-Hamiltonian recovery ladder. No case was excluded, replaced or altered.

## 8.2 Response errors on the frozen geometry challenge

Table S8 gives all eligible methods on the non-water test. Table S9 separates its three neighbour identities.

**Table S8:** Aggregate response errors on the frozen non-water replacement-geometry challenge. Most contacts are repulsive at the stored QM geometry; see Supplementary Fig. S16.

| Method | ESP NRMSE | ESP RMSE (m$E_h$/e) | Dipole RMSE (D) |
|---|---|---|---|
| AIMNet2 | 0.8753 | 3.1776 | 0.4146 |
| AIMNet2-2025 | 1.0951 | 4.2893 | 0.8893 |
| GFN2-xTB | – | – | 0.3442 |
| GLIDER | 0.4504 | 1.6553 | 0.1925 |
| MACE-MDP | – | – | 0.7512 |
| MACE-POLAR-1-L | 0.6172 | 2.2923 | 0.2062 |
| MACE-POLAR-1-M | 0.6061 | 2.1676 | 0.2221 |
| MACE-POLAR-1-S | 0.9347 | 3.7062 | 0.3312 |
| Zero response | 1.0000 | 4.0064 | 0.4123 |

The predeclared primary comparison was MACE-POLAR-1-L. GLIDER response-ESP NRMSE was 0.4504 versus 0.6172, a 27.0% reduction. The paired difference was −0.1668 (95% solute-blocked interval $[-0.2209, -0.1271]$), with 12/12 solute wins and 69/72 configuration wins. MACE-POLAR-1-M was numerically the strongest public field comparator after scoring at 0.6061. GLIDER remained 25.7% lower, with interval $[-0.2137, -0.1051]$, 12/12 solute wins and 64/72 configuration wins. The mean reference response scale was 4.006 m$E_h$/e. Absolute RMSE was 1.655 for GLIDER, 2.168 for MACE-POLAR-1-M and 2.292 m$E_h$/e for MACE-POLAR-1-L.

**Table S9:** Response errors by non-water neighbour on the frozen replacement-geometry challenge. The species groups do not establish performance at relaxed attractive contacts.

| Neighbour | Method | ESP NRMSE | ESP RMSE (m$E_h$/e) | Dipole RMSE (D) |
|---|---|---|---|---|
| NH3 | GLIDER | 0.4667 | 1.5549 | 0.1769 |
| NH3 | MACE-POLAR-1-L | 0.6154 | 2.0744 | 0.2031 |
| NH3 | MACE-POLAR-1-M | 0.6479 | 2.0706 | 0.2331 |
| CH3OH | GLIDER | 0.3668 | 1.3422 | 0.1389 |
| CH3OH | MACE-POLAR-1-L | 0.5512 | 2.0328 | 0.1771 |
| CH3OH | MACE-POLAR-1-M | 0.5433 | 1.9104 | 0.1416 |
| CH3CN | GLIDER | 0.5176 | 2.0686 | 0.2461 |
| CH3CN | MACE-POLAR-1-L | 0.6849 | 2.7698 | 0.2344 |
| CH3CN | MACE-POLAR-1-M | 0.6273 | 2.5217 | 0.2712 |

The primary spatial comparison favoured GLIDER separately for $NH_3$, $CH_3OH$ and $CH_3CN$. GLIDER-minus-MACE-POLAR-1-L intervals were $[-0.2222, -0.0874]$, $[-0.2718, -0.1170]$ and $[-0.2079, -0.1266]$, respectively. The same conclusion held against MACE-POLAR-1-M for every neighbour. Both geometry regimes also favoured GLIDER: NRMSE was 0.436 versus 0.628 for the altered-orientation cases and 0.465 versus 0.606 for the equilibrium-like cases against MACE-POLAR-1-L.

The aggregate dipole RMSE was 0.1925 D for GLIDER and 0.2062 D for MACE-POLAR-1-L, but the paired interval $[-0.0340, 0.00244]$ D included zero. $NH_3$ and $CH_3OH$ favoured GLIDER, whereas $CH_3CN$ was unresolved and had a slightly larger GLIDER point estimate. These are exact results for the frozen 72 configurations, not evidence of a general advantage at attractive non-water contacts.

For context, the same 12 solutes averaged 0.3774 GLIDER NRMSE in their existing three-water equilibrium/orientation cases and 0.4504 in the new one-neighbour panel, an increase of 0.0730. This is descriptive rather than an identity-only effect because environment count and geometry change with identity. Complete per-case and per-solute values, the difficult retained polyfunctional-oxane case, prediction and reference hashes, and the independent raw-array verification are in `prospective_nonwater_environment/`.

## 8.3 Post hoc geometry audit and its implication

The stored component energies permit a contact check without new QM work. We computed the counterpoise interaction energy for every case as the full complex energy minus the solute-with-ghosts and neighbour-with-ghosts energies at the same coordinates and basis. Sixty-two of 72 energies are positive,

52 exceed +10 kcal mol$^{-1}$, the median is +23.7 kcal mol$^{-1}$ and the maximum is +166.2 kcal mol$^{-1}$. The shortest interfragment atomic distance is 0.91 Å (Supplementary Fig. S16). Positive total interaction energy is a fixed-geometry sign of a repulsive contact, although it does not separate exchange, induction, charge transfer and other contributions to the response field. Substituting a larger molecule at a water-oxygen anchor can place its extra atoms very near the solute; the permissive clash rule did not prevent this outcome.

The original aggregate includes all 72 cases and remains reproducible. A post hoc subset with negative counterpoise energy contains only ten cases from two solutes. GLIDER has lower NRMSE in all ten, but that small, selection-conditioned subset cannot establish general transfer across the three species. Restricting the threshold to +10 kcal mol$^{-1}$ retains 20 cases from five solutes, still too little to replace the predeclared panel. The released contact-audit CSV, figure, and script contain all case energies, minimum separations and paired response errors. A separate geometry-selected follow-up was frozen before acquisition of its new QM response labels; it is described next.

### 8.4 Post-audit contact-geometry follow-up

The follow-up was designed after the contact problem was found, so its question and chronology differ from the original predeclared challenge. It retained all twelve Panel-I solutes and all three neighbour species. For each solute–species pair, the two archived replacement geometries supplied starting poses. Every solute atom was fixed while the neighbour was optimized under MMFF94s with interfragment interactions enabled. We compared the two converged poses by MMFF interfragment energy at their optimized coordinates and selected the lower one, breaking ties by source identifier. Each selected pose had to have negative MMFF interfragment energy and a shortest cross-fragment heavy-atom separation of at least 2.4 Å. All 36 pairs passed those geometry-only conditions; selected minimum heavy-atom separations ranged from 2.68 to 3.55 Å. The selected geometries, both candidate energies and the acceptance rule are released. MMFF supplied a contact construction, not a QM response label or a physical decomposition.

The unchanged published GLIDER checkpoint and unfitted MACE-POLAR-1-L were then evaluated on the same solute-centred probe points. A freeze manifest hashed the 36 geometries and all 72 prediction arrays before the first new QM reference calculation. New full-complex and ghost-fragment references used the original $\omega$B97X-D3(BJ)/def2-TZVPD Hamiltonian and probe definition. Case scores use the same response-ESP NRMSE as the original panel; means average the three species within each solute before averaging the twelve solutes. Paired uncertainty resamples whole solutes, 100,000 times. A counterpoise interaction energy is reported for every optimized geometry as a contact sanity check. Selection never used these QM energies, new response labels or prediction errors. Because the original Panel-I reference labels were already known when this follow-up was conceived, the follow-up should not be read as a second independent preregistered test.

All 36 new counterpoise interaction energies were negative, with median −4.86 and range −8.36 to −1.98 kcal mol$^{-1}$ (Supplementary Fig. S17a). On the complete follow-up, equal-solute response-ESP NRMSE was 0.538 for GLIDER and 0.827 for MACE-L, a 34.9% reduction. The paired GLIDER-minus-MACE-L difference was −0.288 (95% solute-blocked bootstrap interval $[-0.375, -0.205]$). GLIDER had lower error for all twelve solutes after averaging over neighbours and in 34 of 36 individual configurations (Supplementary Fig. S17b). The improvement was present in the means for all three neighbour species. All full-complex and ghost-fragment reference calculations converged with the stated Hamiltonian; one iodine-containing case required a second-order SCF recovery and two used projected atomic density as an initial guess. These solver choices and the complete component records are released. The negative interaction energies make these geometries suitable for a bounded attractive-contact check, but do not decompose the response mechanism, establish thermal populations or test arbitrary non-water contacts.

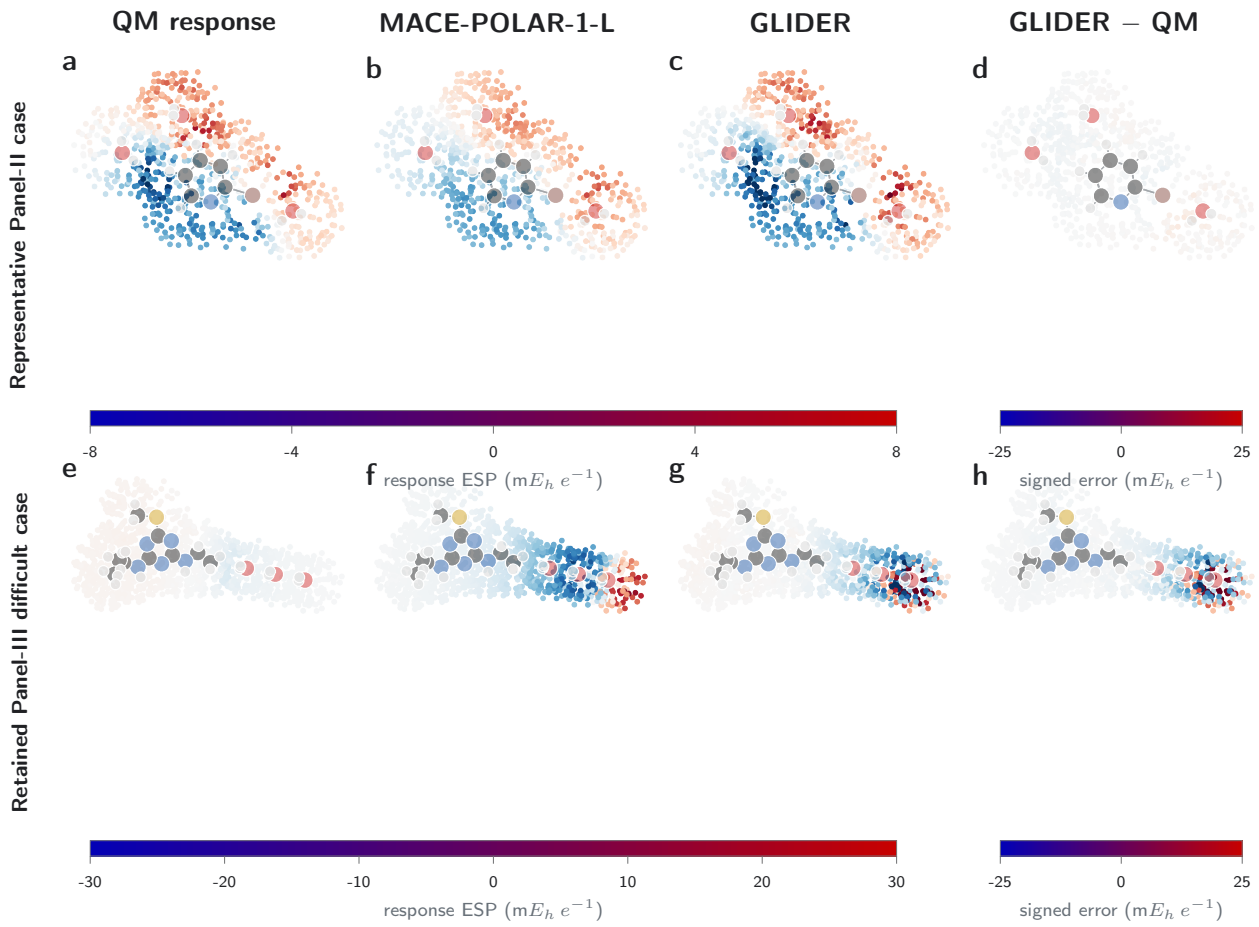



**Figure S6: Reference, prediction and error fields for representative and difficult configurations.** Rows show an ordinary prospective configuration and the retained difficult Panel-III case. Columns show the QM response ESP, MACE-POLAR-1-L, GLIDER and signed GLIDER minus QM error. Values are evaluated on identical stored probe coordinates. The three response panels within each row share a response scale, and both error panels share one error scale. These are actual predictions and references, not illustrative heat maps.

## 9 Qualitative response-field examples

Figure S6 visualizes the observable underlying the scalar NRMSE. The typical example shows that the comparator often recovers the broad sign of the response but misses near-surface anisotropy. The difficult case is included to show the limit of the predictor rather than only a favourable example. Its high GLIDER error remains part of the molecule-equal Panel-III score.

Figure S7 summarizes which public methods expose a faithful spatial response, which expose only a dipole, and which are included as conceptual comparators.

## 10 Liquid-derived local bridge

The constructed training and prospective panels do not test configurations drawn from liquid simulation. This bridge samples local structures from existing explicit-water states. Six solutes outside the original 14-solute response-training set supplied four clusters each (Table S10). Sampling used OpenFF Sage 2.2.1, AM1-BCC charges and TIP3P water at 298.15 K and 1 bar, with a 1 fs integration step, rigid water, PME electrostatics and a 1 nm nonbonded cutoff.

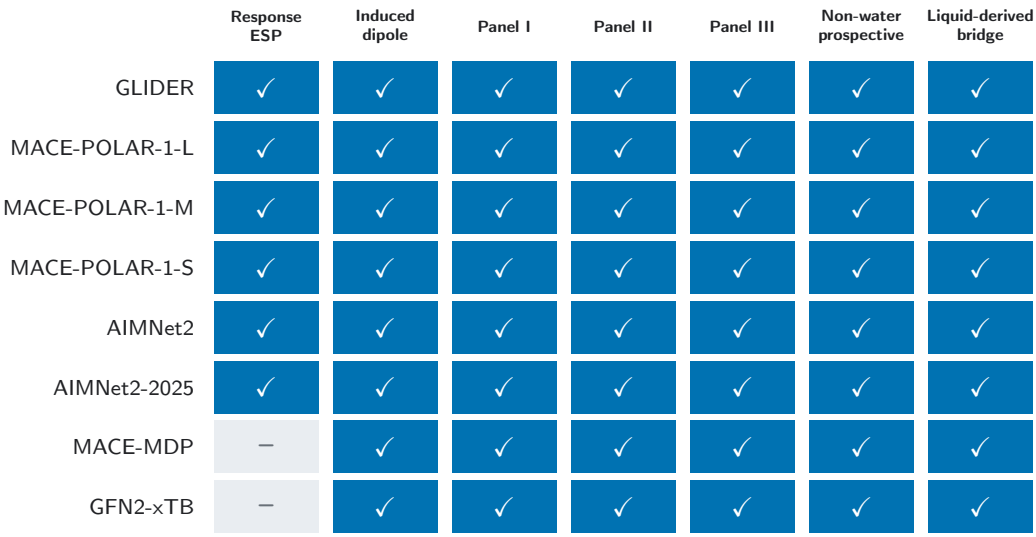



**Figure S7: Comparator coverage and benchmark role.** Rows show public methods and columns show whether the released implementation supplies a faithful full-minus-fragments response ESP, an induced dipole and broad panel coverage without fitting. Filled cells indicate supported use. Methods lacking the spatial observable are retained as dipole-only or conceptual comparators rather than assigned a proxy response ESP.

Four solutes came from fully coupled endpoint states of earlier Hamiltonian replica-exchange runs. Following 50 equilibration iterations, production lasted 200 iterations for acetylene and urea, 300 for hexafluoroisopropanol and 400 for 2-butanone, with 500 integration steps per iteration. The respective production lengths were 100, 150 and 200 ps per replica. The extractor follows the fully coupled state across replicas, not one continuous replica trajectory. Four evenly indexed frames came from endpoint archives saved every ten iterations.

The other two solutes came from Langevin-middle trajectories with 100 ps equilibration followed by four 50 ps blocks, retaining each block endpoint. We call all four selected configurations temporally spaced. Spacing alone does not demonstrate statistical decorrelation. Waters were ranked by minimum-image oxygen distance to the nearest solute atom. The solute and three nearest intact waters defined each of the 24 local response complexes. Periodic wrapping was removed consistently. Source trajectories, frame indices, topology and extraction code are identified in the release.

GLIDER predictions were recorded before response QM. Other eligible public comparators were evaluated once on the same configurations with their fixed decoders. The six identities had been development solutes in a separate interaction-energy investigation, although they were absent from the original GLIDER response-training set. Thus the bridge tests a different geometry source and new response labels with a fixed checkpoint. Solute identities and geometry source change together, so their contributions cannot be separated here.

**Table S10:** Solutes in the liquid-derived local bridge.

| Identifier | SMILES | Chemical stratum |
|---|---|---|
| MNSOL-CONN-HSFWRNGVRCDJHI | C#C | weakly_polar_nonpolar |
| MNSOL-CONN-ZWEHNKRNPOVVGH | CCC(C)=O | oxygen_carbonyl |
| MNSOL-CONN-XSQUKJJJFZCRTK | NC(N)=O | nitrogen_strong_hydrogen_bonding |

Table S11: **Post hoc same-supervision representation control.** Molecule-equal response-ESP NRMSE. The AIMNet2 response head used only the original 14-chemistry/48-environment development set; no panel score entered selection. The comparison is post hoc because panel labels already existed.

| Dataset | GLIDER | AIMNet2 zero fit | AIMNet2 + response head |
|---|---|---|---|
| Panel I | 0.3218 | 0.7668 | 0.7285 |
| Panel II | 0.3422 | 0.8430 | 0.7999 |
| Panel III | 0.4266 | 0.7704 | 1.2942 |

| Identifier | SMILES | Chemical stratum |
|---|---|---|
| MNSOL-CONN-BYEAHWXPCBROCE | OC(C(F)(F)F)C(F)(F)F | halogenated_strong_hydrogen_bonding |
| MNSOL-CONN-QKSKPIVNLNLAAV | ClCCSCCCl | sulfur_and_halogen |
| MNSOL-CONN-WRXCXOUDSPTXNX | Cn1cnc2c(N)ncnc21 | nitrogen_rich_heteroaromatic |

The aggregate GLIDER response-ESP error was 0.345 NRMSE versus 0.583 for MACE-POLAR-1-L. The paired difference interval was $[-0.336, -0.155]$, with six of six solute wins. Dipole RMSE was 0.0601 versus 0.0771 D, with paired interval $[-0.0241, -0.0107]$. These results describe the local extracted clusters. They do not estimate bulk-liquid response or shell convergence.

# 11 Post hoc model controls and extrapolation tests

These analyses were designed and performed after the prospective labels were known. They are separated from the prospective evidence throughout and are not described as preregistered or prediction-before-label tests.

## 11.1 Same-supervision AIMNet2 representation control

The cleanest alternative frozen representation available in the repository and installed public implementation was the final 256-channel per-atom `aim` tensor from official AIMNet2 v0.2.0 ensemble member 0. It is a stable forward output rather than an internal hook. The checkpoint SHA-256 was `f0f7c054bd6464982f6aa35244952e6ba68b06a7a362ddda384514b11208563d`. The base response was the model's native full-complex-minus-fragments atomic charge response. Since the final AIMNet2 tensor is rotationally invariant and does not expose vector channels, the head also used a fixed 32-component geometry-derived directional basis to produce atomic dipoles.

Only the original 14-chemistry/48-environment response-development set entered training. The control used the same response targets, normalized ESP/dipole/site loss, 64-epoch AdamW schedule, five seeds and leave-one-chemistry-out logic as GLIDER. Per-atom layer normalization was fixed from development-only numerical testing because element-wise standardization becomes undefined in a fold whose held chemistry contains the only atom of an element. This label-independent normalization remains defined for such element-unseen folds. No prospective score was calculated until the architecture, checkpoint and reporting rule were fixed.

The final development LOCO response-ESP NRMSE was 0.684, compared with 0.295 for GLIDER. On Panels I and II the trained AIMNet2 control improved modestly over AIMNet2 zero fit. On Panel III it worsened to 1.29 (Table S11). The result shows that matched direct response supervision did not reproduce GLIDER transfer with this final invariant representation. It does not imply that AIMNet2 contains no useful response information, nor does it compare every possible internal layer or equivariant architecture.

Table S12: **Post hoc nested shell-size diagnostic.** Three solutes and three fixed parent snapshots per solute were evaluated at every water count. Scale and RMSE columns are equal-solute means in m$E_h$/e; NRMSE is the equal-solute mean of configuration-wise normalized errors.

| Waters | QM RMS | GLIDER | | MACE-POLAR-1-L | | GLIDER wins |
|---|---|---|---|---|---|---|
| | | RMSE | NRMSE | RMSE | NRMSE | /9 |
| 1 | 1.916 | 0.861 | 0.5479 | 1.806 | 1.2615 | 9 |
| 3 | 2.773 | 1.031 | 0.4108 | 2.349 | 1.0343 | 9 |
| 6 | 2.907 | 1.355 | 0.5180 | 2.616 | 0.9852 | 9 |
| 12 | 3.385 | 1.927 | 0.5976 | 2.596 | 0.8002 | 8 |

Table S13: **Per-solute shell-size behaviour.** Mean response-ESP NRMSE over the three fixed parent snapshots. G denotes GLIDER and M denotes MACE-POLAR-1-L.

| | 1 water | | 3 waters | | 6 waters | | 12 waters | |
|---|---|---|---|---|---|---|---|---|
| Solute | G | M | G | M | G | M | G | M |
| Acetylene | 0.8909 | 2.2649 | 0.5951 | 1.7506 | 0.6973 | 1.3476 | 0.8144 | 1.0438 |
| 2-butanone | 0.4490 | 1.0402 | 0.2941 | 0.7486 | 0.5055 | 0.8923 | 0.4643 | 0.6554 |
| Urea | 0.3038 | 0.4794 | 0.3432 | 0.6037 | 0.3512 | 0.7158 | 0.5141 | 0.7015 |

## 11.2 Nested 1/3/6/12-water diagnostic

Acetylene, 2-butanone and urea were selected before scoring to span weakly polar, carbonyl and strongly hydrogen-bonding regimes. For each solute, the first three pre-existing equally spaced parent frames were used. Waters in a parent were ordered by minimum-image distance from oxygen to the nearest solute atom, with water index as a deterministic tie-break. Prefixes of this single ordering generated exactly nested 1-, 3-, 6- and 12-water clusters. No snapshot or water was selected using model performance.

The deterministic two-shell probe surface was regenerated for the entire cluster at each size and shared by QM and both models for that configuration. Thus the endpoint remains the near-cluster response ESP, while the number and spatial extent of probe points grow naturally with the physical system. Every reference used restricted DF-RKS $\omega$B97X-D3BJ/def2-TZVPD, grid level 4, SCF tolerance $10^{-10}$ and the same full-complex-minus-two-common-basis-ghost subtraction. All 36 configurations converged. GLIDER and MACE-POLAR-1-L were run without retraining. Table S12 summarizes each cluster size, while Table S13 preserves the three solute-level trajectories.

GLIDER had lower error for every paired configuration through six waters and eight of nine at 12 waters. Its aggregate NRMSE was lowest at the original three-water size in this series and increased by 0.187 at 12 waters. Eight of nine parent configurations worsened from three to 12 waters. Absolute RMSE rose from 1.03 to 1.93 m$E_h$/e as the response RMS increased from 2.77 to 3.39 m$E_h$/e. These results establish a useful but degrading finite-size extension, not shell convergence, bulk-liquid transfer or thermodynamic accuracy.

## 11.3 Controlled removal of the global dipole branch

To isolate the global branch from the other changes in the original development controls, we retrained the final model with one intervention: the global dipole readout and its moment-reconciliation step were removed, while the local correction network and averaged frozen prior stayed fixed. Both variants used the original 48 configurations, five seeds, 64 epochs, the same shared-parameter initialization, optimizer and loss weights. Panel response labels scored the completed models only. This is a post hoc comparison, not new prospective evidence.

The matched full model reproduced the published checkpoint on a common newly extracted feature cache. Removal raised ESP NRMSE in all three panels (Table S14). Across 56 solutes, the mean increase was 0.0114, with a solute-bootstrap 95% interval of [0.0008, 0.0203]. Fourteen solutes improved without the branch. Its compact-domain average benefit is therefore modest and not uniform. Fresh feature

extraction can cause small rounding differences from archived panel tables, whereas the matched variants use one common cache.

Table S14: Matched global-branch ablation on the three water panels. Both variants share the original training schedule and frozen prior.

| Panel | Matched variant | ESP NRMSE | ESP RMSE (m$E_h/e$) | Dipole RMSE (D) |
|---|---|---|---|---|
| I | Full model | 0.3218 | 1.1833 | 0.1894 |
| I | Without global branch | 0.3288 | 1.2121 | 0.2170 |
| II | Full model | 0.3422 | 1.1908 | 0.1125 |
| II | Without global branch | 0.3631 | 1.2594 | 0.1256 |
| III | Full model | 0.4266 | 1.1562 | 0.2267 |
| III | Without global branch | 0.4293 | 1.1666 | 0.2283 |

### 11.4 Fragment-separation diagnostic

The response induced by interaction should vanish as the inducing fragments separate. We therefore moved those fragments, unlike the external-probe sweep below, which moves a readout around an unchanged responding complex. For each of the 14 original training solutes, the compact configuration supplied two test systems: solute plus only its first water, omitting the other waters, and solute plus the intact original three- or four-water environment. Either the single water or intact environment was translated away from a stationary solute without internal deformation. Frozen features, averaged prior and published head were recomputed at minimum interfragment distances 3, 4, 5, 6, 8, 10, 15, 20, 50 and 100 Å. The response potential was sampled on the same solute-centred surface throughout, for 280 configurations in total. The 14-solute mean curves use frozen predictions only.

Main-text Fig. 8c,d plots the 14-solute mean predicted RMS. Between 20 and 100 Å, GLIDER increases from 0.564 to 1.540 m$E_h$/e for the single-water system and from 0.925 to 2.843 m$E_h$/e for intact environments. The averaged prior decreases from 0.195 to 0.113 and from 0.301 to 0.239 m$E_h$/e, respectively. The prior retains a smaller nonzero residual, whereas the learned correction grows. These panel-wide values are predicted amplitudes.

To test this behavior against the physical response, we subsequently calculated QM references for one original training solute, `dev_cyclic_carbamate`, at all ten distances in both systems (Supplementary Fig. S18). We kept the same translated geometries, 512 solute-centred probe points, frozen GLIDER predictions and averaged prior. The reference is the counterpoise-consistent complex-minus-fragments response at restricted $\omega$B97X-D3BJ/def2-TZVPD, grid level 4 and SCF tolerance $10^{-10}$; all 60 component calculations converged. At 100 Å, the QM response RMS is 0.000119 m$E_h$/e for the single water and 0.000092 m$E_h$/e for the intact four-water environment. GLIDER's corresponding predicted RMS values are 0.234 and 1.556 m$E_h$/e, against 0.126 and 0.288 m$E_h$/e for the averaged prior. The four-water GLIDER amplitude rises from 0.697 at 20 Å to 1.556 m$E_h$/e at 100 Å as its QM reference falls from 0.00654 to 0.000092 m$E_h$/e. The single-water GLIDER amplitude instead declines over that interval but remains above the QM response. Both counterpoise interaction energies are below 0.0001 kcal mol$^{-1}$ in magnitude at 100 Å. Independent CPU recalculations of 12 distinct conditions agree with the GPU references within 0.00112 m$E_h$/e pointwise and 0.00017 m$E_h$/e in per-case RMS. This direct check establishes a separation-limit error for this solute, while the 14-solute curves show how the frozen predictions behave more broadly.

The RMS global-moment reconciliation component increases from 0.419 to 1.470 m$E_h$/e for one water and from 0.670 to 2.708 m$E_h$/e for intact clusters between 20 and 100 Å. Removing reconciliation only at inference leaves 100-Å amplitudes of 0.422 and 0.619 m$E_h$/e. It drives most of the increase but does not fully explain or repair the nonzero response. The matched compact-domain ablation above measures the separate accuracy effect of removing this branch during training.

One intact-environment configuration, `dev_thio_amide`, has about 1.91 $E_h/e$ RMS at 5 Å in both the prior and GLIDER, raising the 14-solute plotted mean to about 138 m$E_h$/e. The raw MACE-L isolated-environment output is already anomalous. At 6 Å the spike disappears without a change in graph edges. Its numerical cause remains unresolved. The case is retained in the data and log-scale plot so that the upstream anomaly is distinguishable from the long-range growth of the learned correction.

# 12 Post hoc held-out-water downstream response

## 12.1 Frozen pilot and confirmatory protocol

The six-solute pilot is preserved at Git commit `e83ec611a03ba7cfb1578ec1dc112713af641ed4`. It used all four existing liquid-bridge parents for each solute and established the exact-density coupling and derivative machinery. Before any new confirmatory reference or model score was evaluated, the broader protocol, solute list, configurations, endpoints, orientation grid, dipole origins and statistical analysis were committed at `0d3ff3918ece6d6d28fe7d748b9b5ed1a8288ba7`. No pilot solute entered the confirmatory statistics.

A full workspace and Git-history audit found ten additional eligible neutral solutes with existing explicit-water states. These came from completed OpenMM equilibration states archived in Transformato history, one parent per solute. The companion contains their canonical SMILES, SDF structures, topology and state records, source revisions and hashes. The identifier `2OJ9-original` refers to the ligand in the archived `BMI.sdf`, not an entire protein complex. Aliases and a tautomeric duplicate were not counted as further solutes (Table S15).

**Table S15:** Confirmatory held-out-water solutes. Canonical SMILES identify the molecular structures. Source SDF, topology, state, commit and hash records accompany each solute in the release.

| Solute identifier | Canonical SMILES |
|---|---|
| 2OJ9-original | Cc1cc(-n2ccnc2)cc2nc(-c3c(NCc4ccccn4)cc[nH]c3=O)[nH]c12 |
| cdk2-1h1q | c1ccc(Nc2nc(OCC3CCCCC3)c3nc[nH]c3n2)cc1 |
| cdk2-28 | CNS(=O)(=O)c1ccc(Nc2nc(OCC3CCCCC3)c3nc[nH]c3n2)cc1 |
| ethane | CC |
| ethanol | CCO |
| jnk1-17124 | CCOc1nc(NC(=O)Cc2cc(OC)c(Br)cc2OC)cc(N)c1C#N |
| jnk1-18631 | CCOc1nc(NC(=O)Cc2ccccc2OC)cc(N)c1C#N |
| methane | C |
| tyk2-ejm_42 | CCC(=O)Nc1cc(NC(=O)c2c(Cl)cccc2Cl)ccn1 |
| tyk2-ejm_45 | O=C(CC1CC1)Nc1cc(NC(=O)c2c(Cl)cccc2Cl)ccn1 |

The ten retained identifiers are `2OJ9-original`, `cdk2-1h1q`, `cdk2-28`, ethane, ethanol, `jnk1-17124`, `jnk1-18631`, methane, `tyk2-ejm_42` and `tyk2-ejm_45`. Every eligible distinct identity was retained, with one pre-existing state per solute. The ten solutes come from five broader source families. Solute is the primary unit and equal-family weighting is a sensitivity analysis.

Waters were ordered by minimum-image oxygen distance to the nearest solute atom, with zero-based source index as the tie-break. The base $B$ always contained the solute and W1–W3. External probes W4–W12 were excluded from the three base response calculations and from all model inputs. For a frozen external water $Wj$, the exact downstream observable was

$$E_{\mathrm{resp}}(B, Wj) = \sum_A Z_A \Delta V^B_{\mathrm{resp}}(\mathbf{R}_A) - \int \rho_{e,Wj}(\mathbf{r}) \Delta V^B_{\mathrm{resp}}(\mathbf{r})\, d\mathbf{r}. \tag{7}$$

The isolated water density and three counterpoise-consistent base densities used the manuscript electronic-structure level. Density-fitted AO Coulomb integrals evaluated the coupling directly. W$j$ remained frozen and did not polarize the base. The quantity is one response-electrostatic component rather than a full interaction or solvation energy.

The exact-dipole controls used the exact QM base-response dipole about three origins fixed in advance: base centre of mass, solute centre of mass and base centre of nuclear charge. All are released. A summary could select the best single origin for an endpoint across the entire panel, but never a different origin for each case. W4 torque was evaluated about oxygen, and the orientation test used all 24 proper octahedral rotations with oxygen fixed. Force and torque finite differences used 0.002 Å and 0.1° central steps, respectively, with smaller and larger steps predeclared for convergence checks.

### 12.2 Independent-chemistry breadth

The confirmatory W4 signal had RMS magnitude 0.123 kcal mol$^{-1}$, median absolute magnitude 0.0524 kcal mol$^{-1}$ and range -0.242 to +0.207 kcal mol$^{-1}$. Table S16 reports equal-solute statistics. The solute-COM dipole was the most accurate single dipole origin for energy. The base-COM origin was most accurate for torque. The zero predictor is shown because a finite-distance dipole approximation need not improve on zero.

Table S16: **Confirmatory W4 response-coupling energy.** Units are kcal mol$^{-1}$ except for correlations.

| Method | MAE | RMSE | Bias | Pearson $r$ | Spearman $\rho$ |
|---|---|---|---|---|---|
| GLIDER | 0.0487 | 0.0602 | 0.000075 | 0.890 | 0.842 |
| MACE-POLAR-1-L | 0.114 | 0.166 | -0.0441 | 0.499 | 0.552 |
| Exact QM dipole, solute COM | 0.149 | 0.207 | -0.00365 | 0.0636 | -0.0788 |
| Zero response | 0.0853 | 0.123 | 0.0359 | – | – |

GLIDER was the only tested method with energy MAE below zero response. The paired GLIDER-minus-MACE energy-MAE difference was -0.0653 kcal mol$^{-1}$ (95% solute-blocked interval -0.149 to -0.00416), and the difference from the best exact-dipole control was -0.100 (-0.185 to -0.0192). GLIDER improved 7/10 and 8/10 solutes, respectively. Equal-family MAE was 0.0473 for GLIDER, 0.0918 for MACE-POLAR-1-L and 0.144 for the exact dipole. These results independently resolve the pilot energy comparison with the exact dipole, whose blocked interval had included zero. Figure S8 shows the paired errors for all ten solutes.

Torque provides the clearest directional derivative result. QM torque RMS magnitude was 0.244 kcal mol$^{-1}$. GLIDER vector RMSE was 0.0686, compared with 0.156 for MACE-POLAR-1-L and 0.961 for the best fixed-origin exact dipole. Median angular errors were 9.38°, 21.2° and 75.0°. Zero torque was therefore more accurate than the exact-dipole approximation, while GLIDER improved substantially on both zero and MACE-POLAR-1-L. The paired GLIDER-minus-MACE vector-error difference was -0.0683 (95% interval -0.125 to -0.0223), and GLIDER improved 8/10 solutes against MACE and 10/10 against the exact dipole.

For the 24-orientation profiles, mean absolute-profile RMSE was 0.0496 for GLIDER, 0.127 for MACE-POLAR-1-L and 0.253 kcal mol$^{-1}$ for the best exact dipole. GLIDER improved 8/10 solutes against each comparator, and both paired blocked intervals excluded zero. Mean-centred profile RMSE was 0.0464, 0.0580 and 0.251, respectively. The improvement over the exact dipole remained clear, whereas the paired interval against MACE included zero and GLIDER improved only 4/10 solutes. Median profile correlations were 0.959 for GLIDER and 0.955 for MACE. The absolute profile result is strong. The incremental mean-centred shape advantage over MACE is mixed.

Figure S9 displays the paired torque and orientation errors for each solute, including the cases in which the mean-centred gain is uncertain.

Confirmatory force was favourable: vector RMSE was 0.0443 kcal mol$^{-1}$ Å$^{-1}$ for GLIDER, 0.216 for MACE-POLAR-1-L and 0.730 for the best exact dipole, with median angular errors of 19.5°, 64.9° and 68.5° among the seven directionally meaningful cases. The pilot force comparison was much weaker (0.317 versus 0.341 for GLIDER and MACE), so the complete evidence does not support force as a primary result.

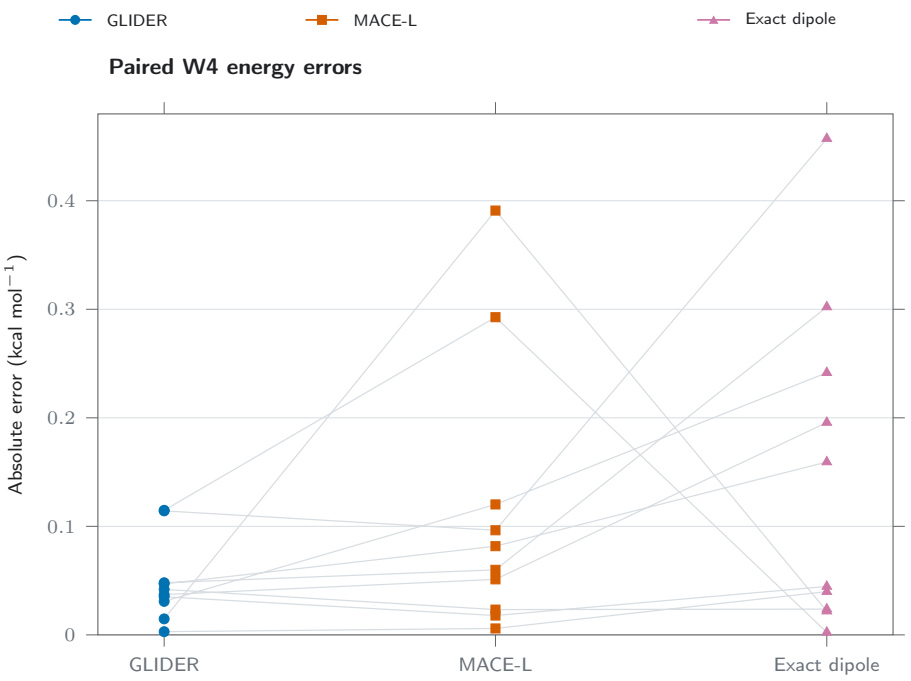



**Figure S8: Per-solute frozen-W4 response-coupling energy errors.** Lines join the three methods for the same solute. The exact response dipole uses the solute-COM origin, the best single predeclared energy origin for the ten-solute panel.

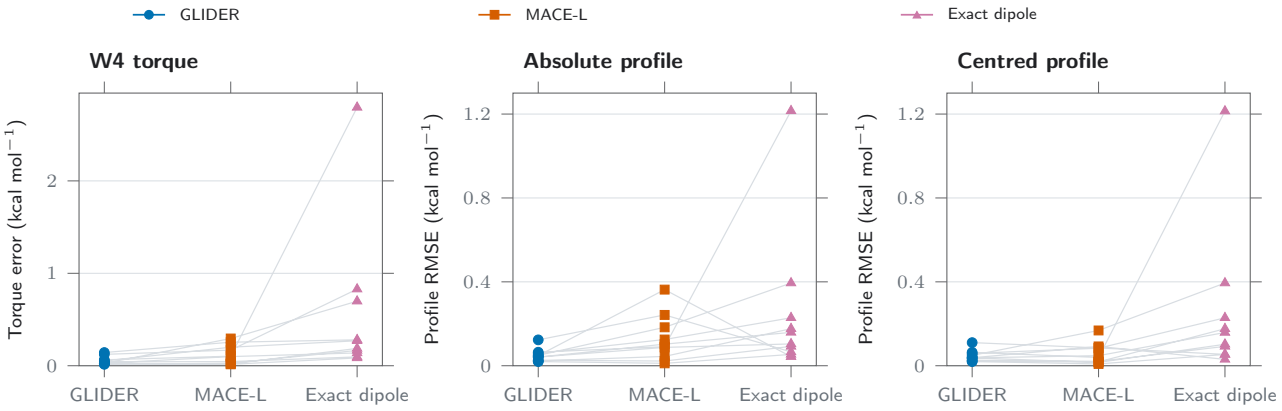



**Figure S9: Confirmatory W4 torque and orientation errors.** From left to right: torque-vector error, absolute orientation-profile RMSE and mean-centred profile RMSE. Lines join the same solute. The exact dipole uses the best single predeclared panel-wide origin for each endpoint.

### 12.3 Exact global multipole hierarchy

A final post hoc control tested whether the gap to the exact response dipole was recovered by the next global moments. The protocol was committed before comparative scoring at `b77dd55e4c0fdff09845629ea3b1dbfe473a84a5`. No new base or probe QM was generated. The stored response matrix represents a positive electron-number difference, so the physical electronic response charge density used for the moments is $\rho^q_{\mathrm{resp}}(\mathbf{r}) = -\Delta\rho_{\mathrm{resp}}(\mathbf{r})$. The nuclear terms cancel in the response difference. For $\mathbf{x} = \mathbf{r} - \mathbf{O}$, raw Cartesian moments were

$$M^{(n)}_{i_1\ldots i_n} = \int \rho^q_{\mathrm{resp}}(\mathbf{r}) x_{i_1} \cdots x_{i_n}\, d\mathbf{r}. \tag{8}$$

The neutral response has $M^{(0)} = 0$. We used the traceless tensors

$$Q_{ij} = 3M^{(2)}_{ij} - \delta_{ij} M^{(2)}_{aa}, \tag{9}$$

$$O_{ijk} = 5M^{(3)}_{ijk} - \delta_{ij}T_k - \delta_{ik}T_j - \delta_{jk}T_i, \qquad T_i = M^{(3)}_{iaa}, \tag{10}$$

with repeated Cartesian indices summed. In this convention, the truncated source potential at $\mathbf{R}$ relative to the same origin is

$$\Phi^{(3)}(\mathbf{R}) = \frac{p_i R_i}{R^3} + \frac{Q_{ij}R_iR_j}{2R^5} + \frac{O_{ijk}R_iR_jR_k}{2R^7}. \tag{11}$$

Equivalently, coupling these moments to the exact frozen-water potential $\Phi_W$ gives

$$E^{(3)} = p_i \partial_i \Phi_W(\mathbf{O}) + \frac{1}{6} Q_{ij} \partial_i \partial_j \Phi_W(\mathbf{O}) + \frac{1}{30} O_{ijk} \partial_i \partial_j \partial_k \Phi_W(\mathbf{O}). \tag{12}$$

All moments were evaluated directly from analytic AO multipole integrals. Water-potential derivatives used a tensor-product seven-point stencil at 0.01 bohr, checked at 0.0075 and 0.015 bohr. The three predefined origins were unchanged. Each endpoint and hierarchy used at most one panel-wide origin, with no case-specific choice.

Table S17 shows that the quadrupole and octupole did not close the dipole gap. Rather than recovering the 0.100 kcal mol$^{-1}$ energy-MAE gap between the exact dipole and GLIDER, adding the quadrupole raised MAE by 0.103 and adding the quadrupole and octupole raised it by 0.359 relative to the dipole. Torque and orientation errors also increased. Solute-blocked 95% intervals for GLIDER minus every exact hierarchy excluded zero for energy MAE, torque vector RMSE and both absolute and mean-centred orientation-profile RMSE. GLIDER improved 8/10 solutes over each hierarchy for energy, 10/10 for torque, and 8/10, 10/10 and 9/10 over dipole, dipole–quadrupole and dipole–quadrupole–octupole, respectively, for the absolute orientation profile. Equal-family and leave-one-solute-out summaries preserved the ordering. Figure S14 shows the error hierarchy at W4 and across W4–W12.

Table S17: **Exact global response-multipole control at W4.** Energy is MAE. Torque is vector RMSE. Orientation values are mean per-solute profile RMSE. All units are kcal mol$^{-1}$.

| Method | Energy | Torque | Absolute profile | Centred profile |
|---|---|---|---|---|
| GLIDER | 0.0487 | 0.0686 | 0.0496 | 0.0464 |
| MACE-POLAR-1-L | 0.114 | 0.156 | 0.127 | 0.0580 |
| Zero response | 0.0853 | 0.244 | 0.131 | 0.130 |
| Exact dipole | 0.149 | 0.961 | 0.253 | 0.251 |
| Exact dipole + quadrupole | 0.252 | 2.14 | 0.394 | 0.393 |
| Exact dipole + quadrupole + octupole | 0.508 | 5.80 | 0.812 | 0.810 |

The worsening hierarchy is physically interpretable only with its convergence geometry. A finite global multipole series is an exterior expansion, but these extended base complexes place most held-out waters inside the practical source extent. At W4, the oxygen lay outside the radius containing 99% of the absolute response density for only 1/10 base-COM, 1/10 nuclear-centre and 0/10 solute-COM expansions.

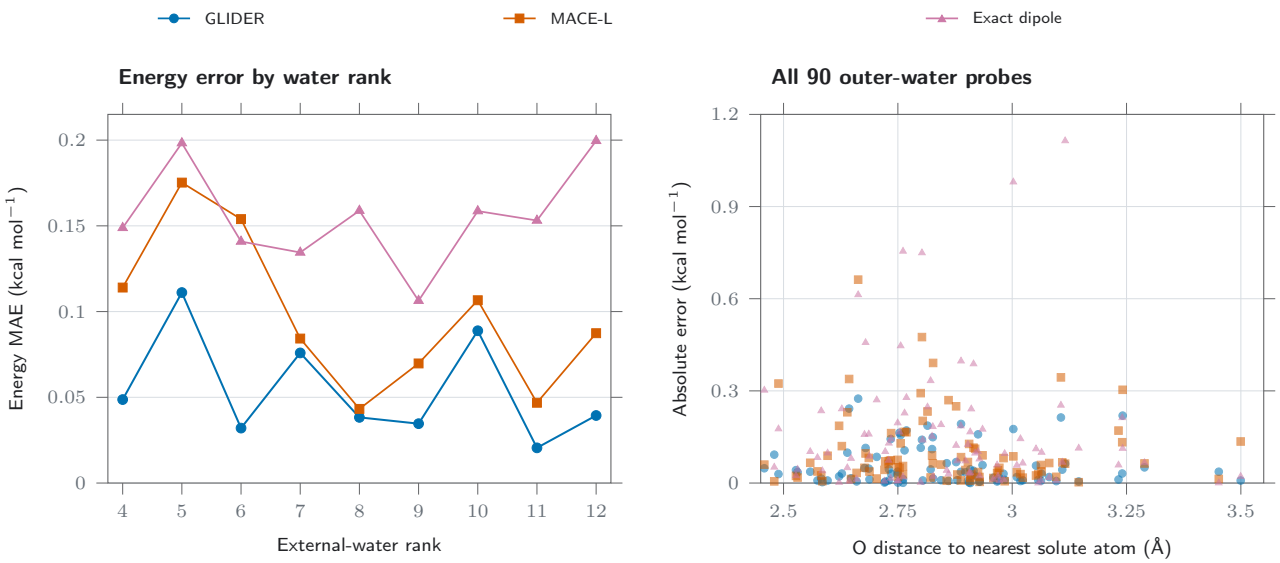



**Figure S10: Range of the frozen outer-water response coupling.** Left, equal-solute energy MAE by external-water rank. Right, all 90 per-probe absolute errors against oxygen distance to the nearest solute atom. The base remains solute plus W1–W3 for every point.

The median oxygen-to-$r_{99}$ ratio was 0.65–0.67. Across all 90 W4–W12 probes, only 15/90 lay outside $r_{99}$ for the base-COM or nuclear-centre origins and none did for the solute COM. Increasing order did not approach the exact coupling systematically with rank. Low-order single-origin truncations are therefore not expected to behave like a converged exterior expansion here. The result shows that these tested truncations do not recover the sampled near-field response. It does not show that multipoles in general fail, or rule out a converged high-order, distributed or alternative compact representation.

The implementation passed the frozen numerical identities. The physical response charge was neutral within $6.67 \times 10^{-14}$ e, and the density-derived dipole agreed with the stored QM response dipole within $4.76 \times 10^{-8}$ D. Rank-1 energies reproduced all existing W4–W12, torque and orientation values within $4.44 \times 10^{-16}$ kcal mol$^{-1}$. Origin translations agreed through rank three within $8.27 \times 10^{-11}$ in the corresponding atomic units. Rotation-covariance relative errors were below $9.52 \times 10^{-4}$. Step changes altered quadrupole couplings by at most $2.97 \times 10^{-8}$ and octupole couplings by at most $4.08 \times 10^{-5}$ kcal mol$^{-1}$ on the predefined smallest and largest validation cases.

## 12.4 Spatial range and cumulative outer coupling

The same base response was coupled independently to W4–W12. GLIDER had lower equal-solute energy MAE than both MACE-POLAR-1-L and the exact dipole at every rank (Fig. S10). Against MACE, rank-specific blocked intervals excluded zero for W4, W6, W9, W11 and W12 and included zero for the other four ranks. Point estimates favoured GLIDER at all nine ranks, although only five rank-specific intervals excluded zero. In the well-sampled 2.5–3.0 Å bin, MAE was 0.0543 for GLIDER, 0.108 for MACE and 0.155 kcal mol$^{-1}$ for the exact dipole. Values in the 3.0–3.5 Å bin were 0.0512, 0.0850 and 0.176, but only four solutes contributed. The observed support therefore contains no crossover to dipole sufficiency and cannot locate one at longer range.

Response torque was also evaluated for every W4–W12 probe (Fig. S11). GLIDER had lower vector RMSE than MACE-POLAR-1-L at seven of nine ranks and lower error than the exact-dipole field at every rank. The GLIDER–MACE blocked interval excluded zero at W4, W8 and W9. W7 favoured MACE slightly, and W12 was nearly tied. The range-dependent torque result is therefore supportive but

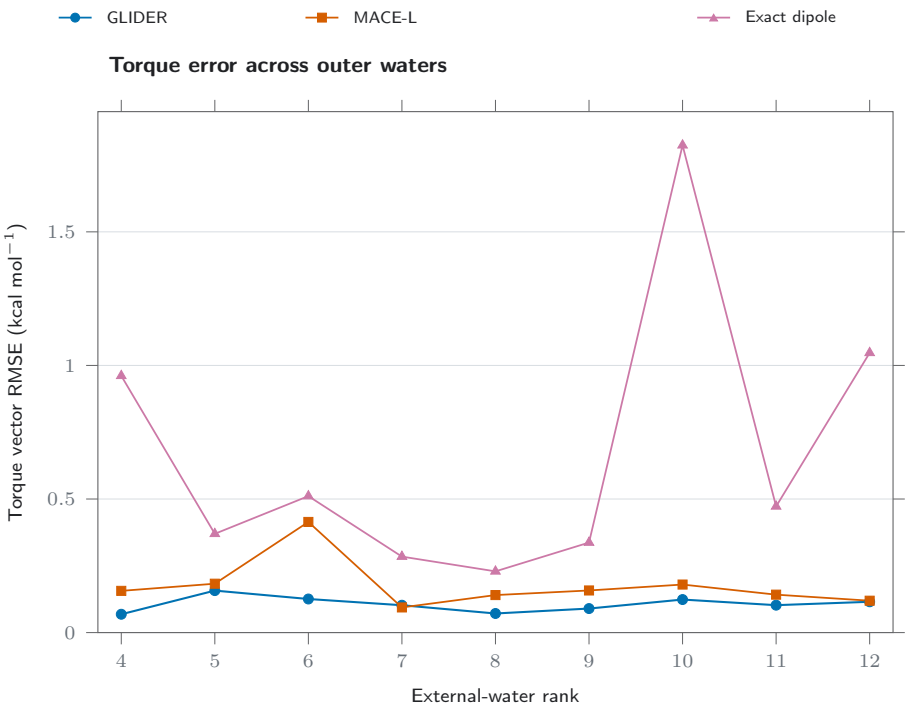



**Figure S11: Response-torque error by external-water rank.** Equal-solute vector RMSE for the same fixed base and independent W4–W12 probes. The exact dipole uses the base-COM origin selected once for the torque endpoint.

less uniform than the W4 torque and energy results.

For the cumulative quantity $\sum_{j=4}^{k} E_{\mathrm{resp}}(B, Wj)$, GLIDER also had the lowest MAE at every $k$ (Fig. S12). At $k = 12$, MAE was 0.138 for GLIDER, 0.290 for MACE-POLAR-1-L and 0.520 kcal mol$^{-1}$ for the exact dipole. The paired GLIDER-minus-comparator intervals were [-0.290,-0.0205] and [-0.641,-0.147], respectively. This sum remains a cumulative frozen Coulomb component, not a shell interaction or solvation energy.

## 12.5 Response fraction of energy and directionality

The W4 response was a modest fraction of the full frozen electrostatic energy but a larger fraction of its directional structure (Fig. S13). The median response/full magnitude was 1.76% for the energy offset, 11.7% for torque and 4.84% for orientation modulation. Median response and full values were 0.0524 and 2.06 kcal mol$^{-1}$ for energy, 0.175 and 1.14 for torque, and 0.439 and 4.45 for orientation modulation. Prespecified small-denominator rules left one energy ratio undefined and retained all torque and orientation ratios. The response fraction was larger for the directional observables than for the energy offset, most notably for torque.

## 12.6 Numerical checks and interpretation boundary

All ten bases, 90 external probes and 900 orientation/torque variants are accounted for. All SCFs converged. The maximum absolute base-response electron number was $6.67 \times 10^{-14}$ e, and the density-to-dipole reconstruction discrepancy was $4.76 \times 10^{-8}$ D. The maximum coupling-linearity error was $2.69 \times 10^{-15}$ $E_h$. Direct GLIDER site and field evaluation agreed to $1.21 \times 10^{-17}$ $E_h$/e. Force and torque step-size changes were at most $1.31 \times 10^{-5}$ kcal mol$^{-1}$ Å$^{-1}$ and $1.92 \times 10^{-4}$ kcal mol$^{-1}$, respectively. A common translation changed the scalar coupling by at most $6.91 \times 10^{-10}$ kcal mol$^{-1}$. The rotation check

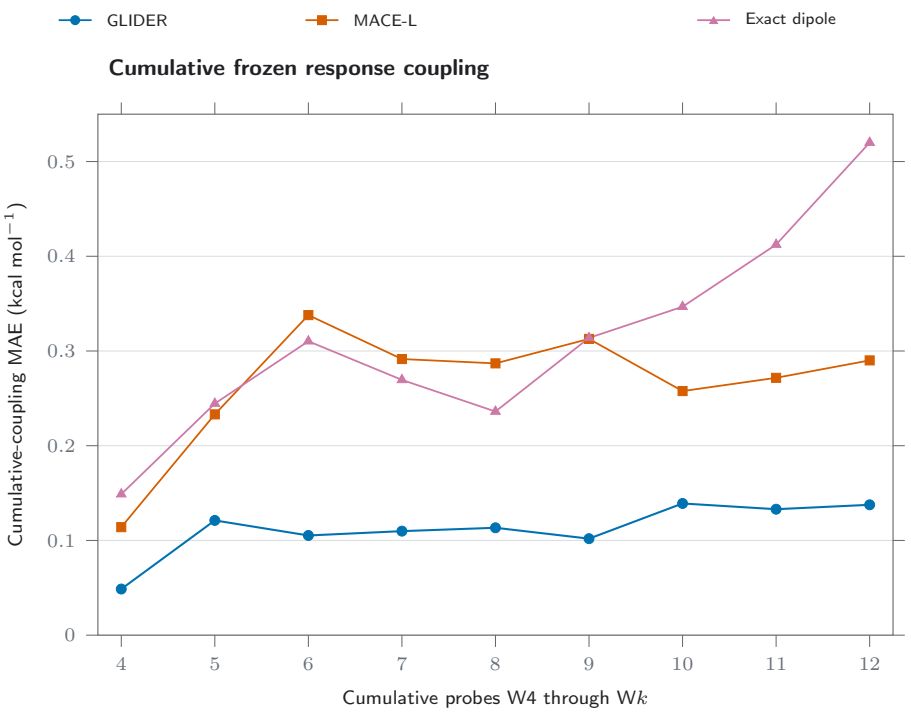



**Figure S12: Cumulative frozen outer-environment response coupling.** Equal-solute MAE for sums from W4 through W$k$.

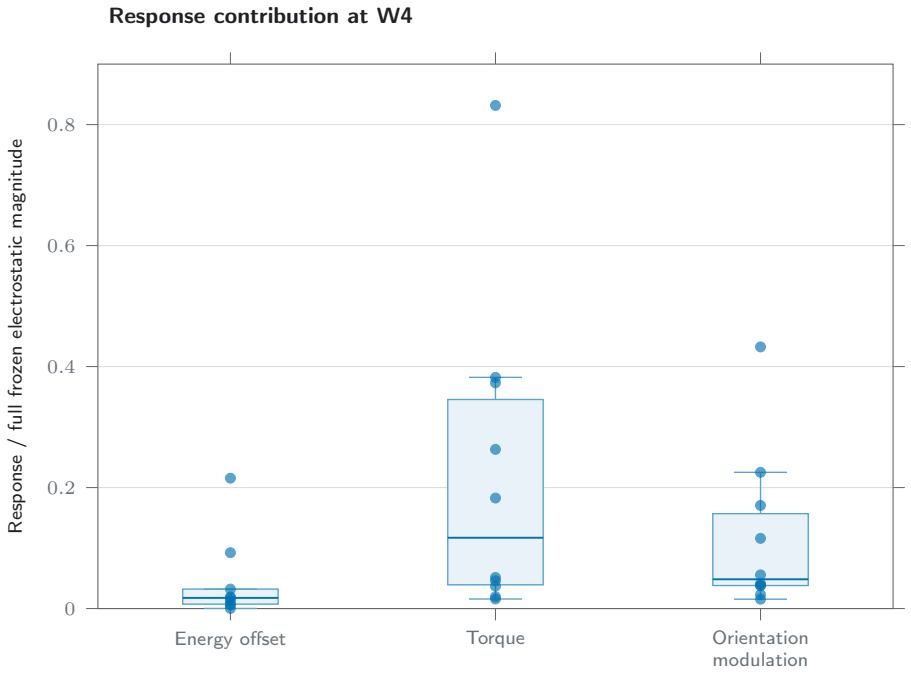



**Figure S13: Response as a fraction of full frozen electrostatics at W4.** Points are the ten solutes except where the frozen denominator rule leaves a ratio undefined. Boxes show their distributions.

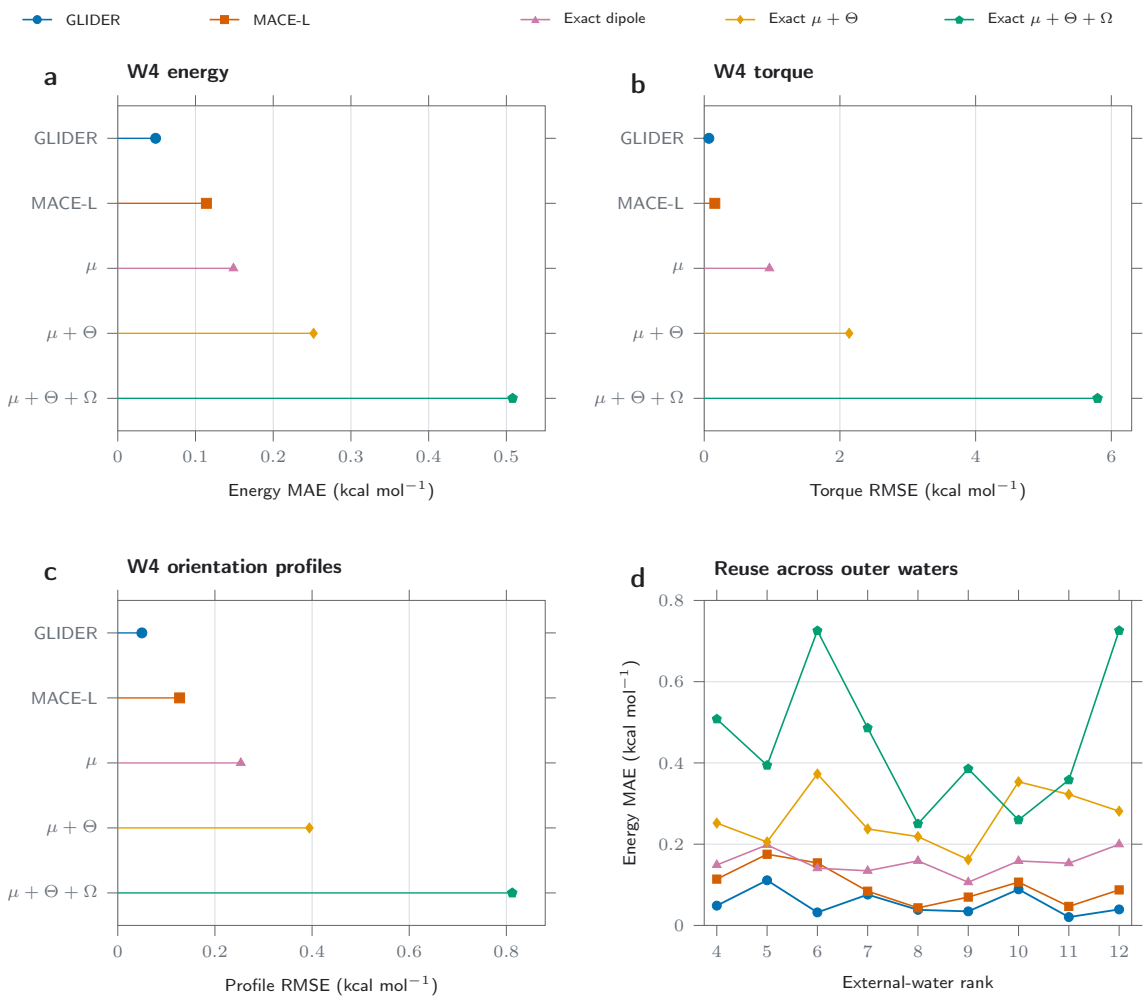



**Figure S14: Exact global response-multipole hierarchy. a–c**, W4 energy, torque and absolute orientation-profile errors. Exact moments are shown cumulatively through dipole ($\mu$), quadrupole ($\Theta$) and octupole ($\Omega$). Each order uses the most accurate single predefined origin for that endpoint over the panel. **d**, Energy MAE for W4–W12, using the origin selected once from W4 energy.

differed by $3.49 \times 10^{-5}$ kcal mol$^{-1}$. Rigid-water bond and angle errors were below $2.22 \times 10^{-15}$ Å and $2.27 \times 10^{-13}$ degrees.

These tests establish a downstream consequence of the fixed base response for a frozen external quantum water. They do not include mutual base–probe polarization, exchange, dispersion, nuclear relaxation, total molecular forces, a conservative Hamiltonian, dynamics or thermodynamics.

# 13 Numerical and physical-consistency checks

The reported checkpoint SHA-256 is `a288afa285128e13cb7a05ba9459f1dce07e79b8638c6e272c2509ae68d7a288`. Prospective predictions conserve response charge below $10^{-8}$ e. The site-reconstructed molecular response agrees with the stored global dipole below $3 \times 10^{-7}$ D. On a target-free rotation/translation subset, maximum deviations were $6.21 \times 10^{-5}$ e for charge, $2.40 \times 10^{-4}$ e bohr for atomic dipoles and 0.00581 D for the molecular dipole. Charge neutrality and site/global moment consistency are analytic. Small covariance deviations originate in finite-precision frozen encoder features.

The release-facing implementation reproduced all 144 stored Panel-I/II predictions with maximum differences $7.51 \times 10^{-10}$ Hartree/e in ESP and $1.53 \times 10^{-7}$ D in dipole. An independent repeat of all 80 Panel-III predictions was bitwise identical at stored precision. The statistics script rebuilds aggregate, molecule, regime, chemistry, bootstrap and leave-one-out tables from stored per-case arrays.

For the prospective non-water test, an independent verifier accounted for all 72 configurations, checked every reference hash and every shared probe array, and reconstructed GLIDER and MACE-POLAR-1-L NRMSE to machine precision. All 72 reference triplets converged, six used the predeclared same-Hamiltonian continuation, and none was excluded or replaced.

Table S18 records the principal artifact hashes used in the release checks.

**Table S18:** Key reported-artifact hashes.

| Artifact | SHA-256 |
|---|---|
| GLIDER checkpoint | `a288afa285128e13cb7a05ba9459f1dce07e79b8638c6e272c2509ae68d7a288` |
| Panel III manifest | `873e0b05843f0ebf00bfa8b1a5a5c4a09c4b4946b685d132407a01d8bb2118b7` |
| Liquid bridge geometry | `59b75023bb3ab12be5e616b6e923cd07e8675fbebb38a6c72cc9a2d3536a2be2` |
| Reported statistics | `3fca48d99cc833b7470adb3a2a7874fe76ec576bea3d94f395faebfc698fc95f` |

# 14 External-probe map of the fixed response

The W4–W12 experiment measures coupling at waters drawn from existing liquid structures. To ask what happens between and beyond those positions, we sampled a wider range of external-water locations around the same ten confirmatory bases. Each base geometry, QM response density, GLIDER prediction and polar-baseline prediction was held fixed. Only the water used to *read out* the field moved. This post hoc test maps one-way electrostatic coupling of an already induced response. It does not test whether the response itself vanishes when the inducing waters are separated from the solute. That is the distinct fragment-separation diagnostic of Section 11.4.

## 14.1 Probe construction and response coupling

Twelve Fibonacci directions from each base centre of mass defined radial placements. Along each direction, the probe oxygen was placed at a prescribed minimum clearance from the union of base-atom van der Waals surfaces. The nine clearances were 0.8, 1.2, 1.8, 2.5, 3.5, 5, 7.5, 10 and 15 Å. This coordinate is different from the oxygen-to-nearest-solute-atom distance used for W4–W12. Twelve evenly indexed members of the 24-orientation grid provided rigid water orientations. A placement was rejected before

scoring when any probe–base atom distance was below 1.2 Å, using the same rule for every method. The dense grid retained 12,903 of 12,960 planned placements, with all ten bases represented.

The dense probe used fixed three-site charges of −0.834 e on oxygen and +0.417 e on each hydrogen. Its response coupling was $\sum_a q_a \Delta V^B_{\mathrm{resp}}(\mathbf{r}_a)$. Reference potentials and fields were evaluated from the stored base QM response density rather than interpolated from the training probes. GLIDER used its released point-charge and point-dipole decoder, and the polar baseline used its native 1.5-Å Gaussian decoder. Exact dipole, dipole plus quadrupole and dipole plus quadrupole and octupole fields were reconstructed from that same QM response density about one fixed base-centre-of-mass origin. Zero response was also scored. Central differences of the potential at 0.001 Å gave the electric field for fixed-charge force and torque calculations. Neither probe model induced a new base response.

A smaller complementary grid coupled the response to the full frozen isolated-QM-water charge distribution. It used three directions, clearances of 0.8, 1.8, 3.5, 7.5 and 15 Å, and four orientations per position. The common steric rule retained 596 of 600 planned orientations. Coupling used the density-fitted Coulomb integrals of Section 12.1. Torques used the existing ±0.1° isolated-water variants and were complete at 147 of 150 positions, with identical coverage for every method. All three original exact-dipole origins were retained. Higher-order quantum-water multipole couplings were not part of this sweep. A separate preflight duplicated one included placement and was not counted twice.

### 14.2 Aggregation and uncertainty

At each position, the accepted orientations supplied coupling MAE and RMSE. Normalized coupling RMSE divided the position-level error by its QM-profile RMS. Centred-profile RMSE removed each method's orientation mean before comparison. Position-level errors were averaged within a base and then equally across the ten solutes. The resulting values are not pooled errors over thousands of independent waters. Dense-grid torque error is the orientation RMS of the vector error at each position. The quantum-water torque statistic is the vector error at the reference orientation, so the two torque summaries are not directly interchangeable.

Paired 95% percentile intervals used 20,000 solute-blocked bootstrap draws with seed 20260925. All positions and orientations of a base stayed in one block. With ten independent solutes and multiple exploratory endpoints, these intervals are descriptive and have no multiplicity adjustment. Table S19 gives the sweep-wide means, while Fig. S15 resolves distance and probe type.

Table S19: **Frozen external-probe sweep across ten solutes.** Sweep entries average position-level errors within each solute, then average solutes equally. The 15-Å column isolates the largest tested clearance. MAE, centred orientation-profile error and torque error are in kcal mol$^{-1}$. NRMSE is dimensionless. These are post hoc summaries, separate from the original W4 confirmatory endpoints.

| Method | Sweep MAE | 15 Å MAE | NRMSE | Centred profile | Torque |
|---|---|---|---|---|---|
| *Fixed-charge water probe* | | | | | |
| GLIDER | 0.02459 | 0.000531 | 0.3904 | 0.02746 | 0.05038 |
| Polar baseline (MACE-L) | 0.03150 | 0.000543 | 0.4011 | 0.03474 | 0.06133 |
| Zero response | 0.06650 | 0.002298 | 1.0000 | 0.07302 | 0.1151 |
| Exact D, base COM | 0.2278 | 0.001062 | 2.834 | 0.3528 | 2.277 |
| Exact D+Q, base COM | 2.909 | 0.000359 | 43.49 | 5.707 | 106.2 |
| Exact D+Q+O, base COM | 37.34 | 0.000174 | 739.4 | 97.13 | 2369 |
| *Frozen quantum-water probe* | | | | | |
| GLIDER | 0.05595 | 0.000431 | 0.4902 | 0.03308 | 0.07054 |
| Polar baseline (MACE-L) | 0.06137 | 0.000476 | 0.5868 | 0.04344 | 0.09022 |
| Zero response | 0.09475 | 0.001796 | 1.0000 | 0.08053 | 0.1496 |
| Exact D, base COM | 0.5044 | 0.000907 | 2.629 | 0.2400 | 0.7129 |
| Exact D, solute COM | 0.2012 | 0.000892 | 1.676 | 0.1521 | 0.3330 |
| Exact D, nuclear centre | 0.4196 | 0.000906 | 2.298 | 0.2551 | 0.9457 |

### 14.3 What changes with probe distance

Across the dense fixed-charge grid, GLIDER's equal-solute coupling MAE was 0.0246 kcal mol$^{-1}$, compared with 0.0315 for the polar baseline and 0.0665 for zero response (Table S19 and Fig. S15a). The paired GLIDER-minus-baseline difference was $-0.00691$ kcal mol$^{-1}$, with a 95% interval of $[-0.01467, -0.00069]$ and seven of ten solute wins. Normalized coupling errors were closer, 0.390 versus 0.401, and their paired interval included zero. Centred-profile and torque means also favoured GLIDER, each with seven solute wins. Fixed-charge force errors were 0.0459 versus 0.0512 kcal mol$^{-1}$ Å$^{-1}$, with five solute wins and an interval that included zero.

The full quantum-water subset had lower mean coupling MAE for GLIDER than the polar baseline, 0.0560 versus 0.0614 kcal mol$^{-1}$, compared with 0.0947 for zero response (Fig. S15c). Its paired absolute-error interval, $[-0.01400, 0.00448]$ kcal mol$^{-1}$, included zero. The sweep-wide centred-profile and torque comparisons were also unresolved. Quantum-water normalized coupling RMSE was 0.490 versus 0.587, with a paired interval of $[-0.1766, -0.0206]$. The absolute and normalized metrics weight weak-response positions differently. At the shortest 0.8-Å clearance, the quantum-water absolute-error point estimate instead favoured the polar baseline, 0.204 versus 0.227 kcal mol$^{-1}$ for GLIDER. This exploratory grid does not replace the original confirmatory W4 test.

Exact global moments show why the contact-range result should not be extrapolated unchanged to the far field (Fig. S15b). Adding quadrupole and octupole terms worsened dense-grid errors at short clearance but recovered accuracy farther out. At 15 Å, coupling MAE was 0.000531 kcal mol$^{-1}$ for GLIDER, 0.000543 for the polar baseline, 0.001062 for the exact dipole, 0.000359 after adding the exact quadrupole and 0.000174 after adding the exact octupole. The exact dipole alone remained less accurate than GLIDER across the sampled dense grid, while higher exact moments surpassed it at the farthest point. At 0.8 Å, 94% of probe atoms lay inside the sphere enclosing all base-atom van der Waals surfaces about the base centre of mass. None did at 10 or 15 Å. This geometric envelope differs from the response-density $r_{99}$ diagnostic of Section 12.3.

Absolute GLIDER–baseline differences narrowed at long range as the coupling signal declined (Fig. S15a,c). The clearance-specific normalized paired intervals in Fig. S15d do not establish a uniform

near-field advantage. This is a map of frozen electrostatic consequences at sampled positions, not a claim about self-consistent polarization, a complete interaction energy or asymptotic accuracy.

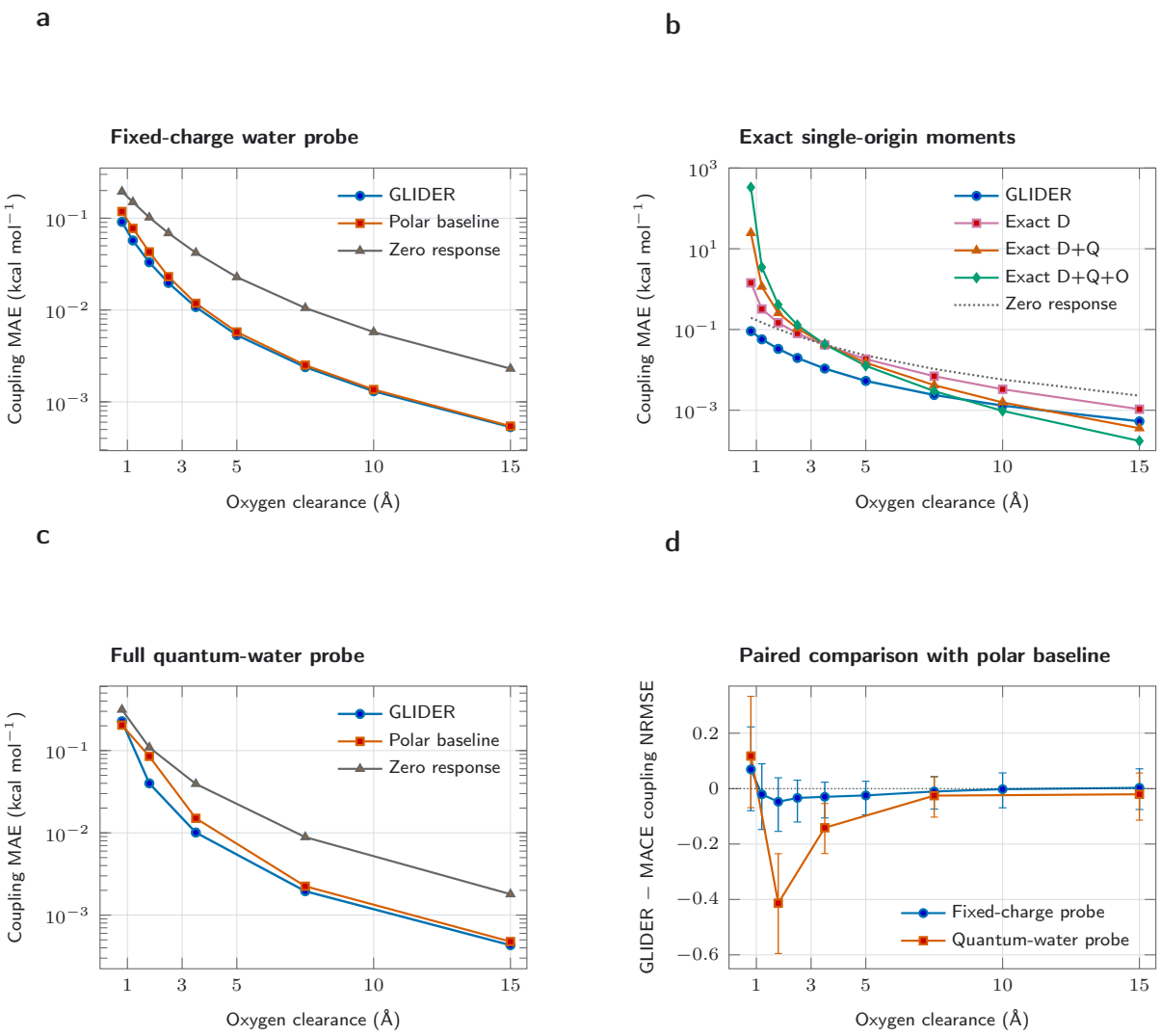



**Figure S15: Frozen response coupling across external-water positions.** The same ten confirmatory bases are used throughout. **a**, Coupling MAE for the dense fixed-charge probe. **b**, Coupling MAE for GLIDER and exact moments through dipole (D), quadrupole (Q) and octupole (O), all at the fixed base-centre-of-mass origin. The dotted line marks zero response. **c**, Coupling MAE for the smaller full quantum-water probe. **d**, Paired GLIDER-minus-polar-baseline normalized coupling errors, with 95% solute-blocked bootstrap intervals. The polar baseline is MACE-L. The horizontal coordinate is oxygen clearance from the *base* van der Waals surface, rather than the nearest-solute-atom distance in the W4–W12 analysis. Every retained placement contributes. Both probes are frozen and exclude mutual base–probe polarization.

## 14.4 Source data

The public companion's `experiments/distance_sweep/` directory contains the ten base identities, accepted probe positions and orientations, stored reference potentials and fields, density-based couplings, model predictions, per-position and per-solute errors, aggregate tables and the reproduction script. Its raw-input hash list links the sweep files to their base references. A CPU spot check reproduces the stored couplings and field values to numerical precision. The analysis reuses the released checkpoint and

archived QM densities without new training or base-response calculation.

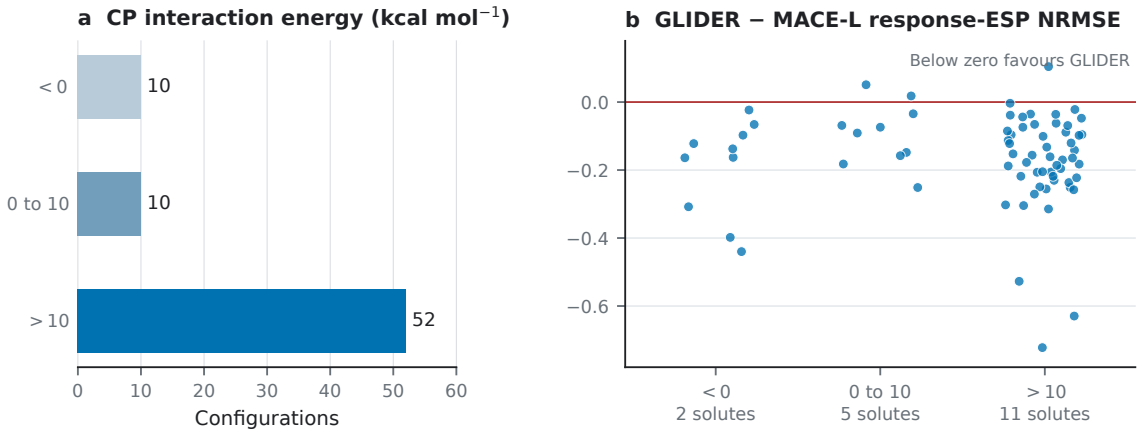



**Figure S16: Post hoc contact audit of the frozen 72-case non-water panel.** **a**, Counterpoise interaction energy from the stored full-complex and ghost-fragment QM energies, grouped into negative, 0 to 10 and above 10 kcal mol$^{-1}$ bands. **b**, Paired GLIDER-minus-MACE-L response-ESP NRMSE for every case in those energy bands. The original score retains all points. Negative paired error favours GLIDER; this plot diagnoses geometry rather than selecting a new benchmark subset. All underlying energies, geometries and errors are released.

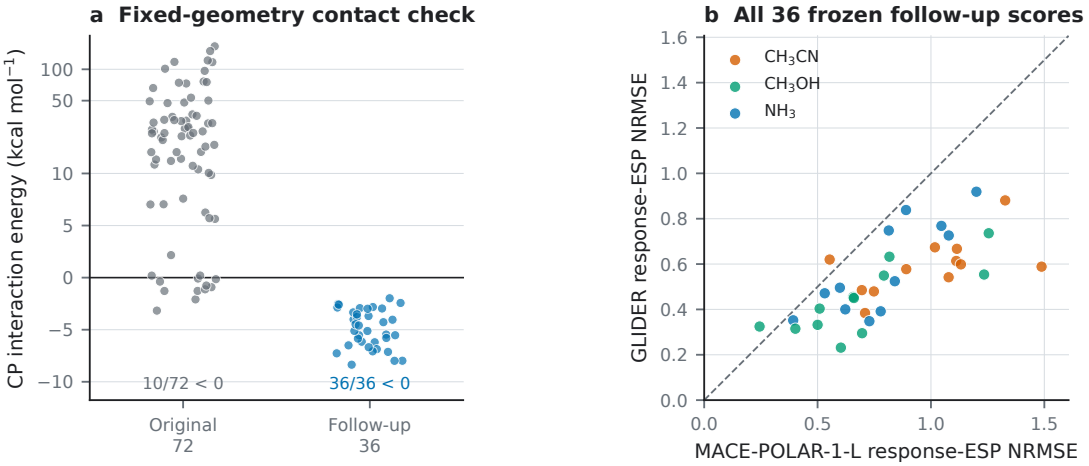



**Figure S17: Complete post-audit non-water contact follow-up.** **a**, Counterpoise interaction energies for all 72 original unrelaxed replacement geometries and all 36 separately constructed contacts. The follow-up contacts were selected by a geometry-only MMFF rule before new QM references or response scores were obtained; every follow-up QM interaction energy is negative. **b**, Paired response-ESP NRMSE for each of the 36 follow-up configurations on identical probes. Colours identify the three neighbour species, and the diagonal denotes equal error. All cases, including the two with lower MACE-L error, are shown. The published GLIDER checkpoint and unfitted MACE-L were fixed before the new QM calculations. The 72-case panel remains an unrelaxed geometry challenge, while the 36-case panel tests bounded transfer at constructed attractive contacts.

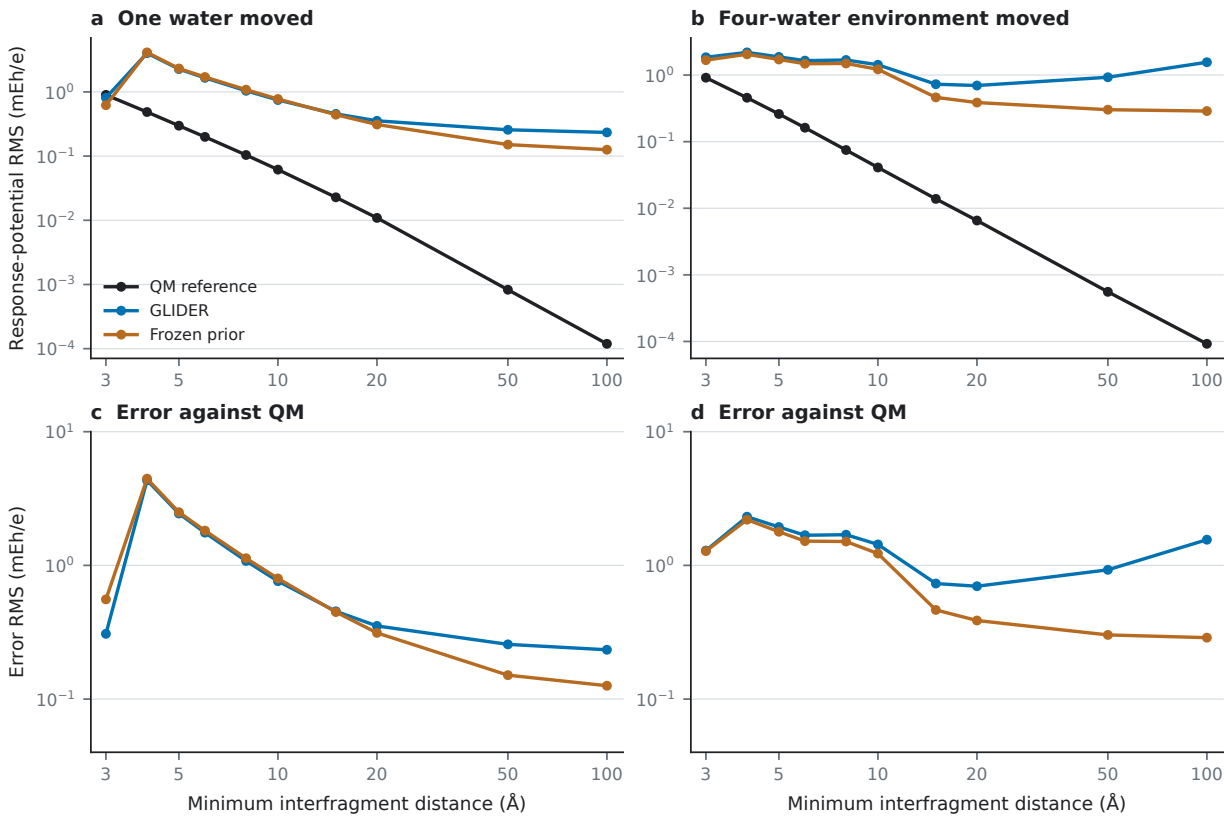



**Figure S18: Direct QM check of the fragment-separation limit for one training solute.** For `dev_cyclic_carbamate`, the first water alone (**a,c**) or the intact four-water environment (**b,d**) is translated away from the fixed solute. The horizontal coordinate is the minimum interfragment distance. **a,b**, Response-potential RMS for the new QM reference, frozen GLIDER and averaged polar prior. **c,d**, RMS prediction error against QM. All values are over the same 512 solute-centred probes in m$E_h$/e. The reference uses the full complex and two ghost-basis fragments at each of ten distances. The 20 QM references and predictions are released in the companion repository; this post hoc one-solute check is distinct from the 14-solute, prediction-only means in main-text Fig. 8c,d.

## 15 Independent post hoc comparison with ARROW

The classic analytical ARROW force field was supplied by Boris Fain and used without its neural pair-energy correction, parameter refitting or access to GLIDER response labels during ARROW parameterization [Pereyaslavets et al., 2022]. ARROW assigns discrete atom types, and this parameterization could type 11 of the 62 distinct solutes represented in the three prospective panels, liquid-derived clusters and nested shell-size set. We retained all 68 configurations of those 11 solutes that were successfully typed, including a solute's repeated appearance in both liquid and shell-size sets. No model was compared on a configuration absent from the common subset. The original GLIDER predictions preceded their QM labels; the ARROW evaluation was performed later. Panel I subsequently entered a separate development pool, and all ARROW comparisons here are explicitly post hoc.

At each frozen geometry, separate self-consistent ARROW cloud solutions were obtained for the complex, solute and water environment. The atom-centred response cloud dipoles are the complex cloud dipoles minus their fragment counterparts. ARROW has no response-charge channel in this parameterization. We evaluated those response dipoles with ARROW's native smeared-cloud electrostatic kernel at the exact reference probe points. For an operator control, the same dipoles were also evaluated with the point-dipole operator used for GLIDER. Configuration NRMSE follows Eq. 4 of this Supplementary

Information. We average configurations within each solute–set group, then give equal weight to those groups. Repeated appearances of urea and 2-butanone give 13 groups but only 11 distinct solutes. For the paired ARROW-minus-GLIDER interval, 100,000 bootstrap resamples draw the 11 solutes as independent units and retain both groups when a solute is redrawn.

Table S20: **Response-ESP NRMSE on identical ARROW-covered configurations.** Values are equal means of solute–set group means. The native smeared-cloud ARROW result is the primary ARROW score; the point column uses the same ARROW response dipoles with GLIDER's point operator. The final two rows are alternative summaries of the rows above, not additional cases. The 68-case group contains 11 distinct solutes; urea and 2-butanone contribute both liquid and shell-size groups. Lower is better.

| Set | Cases | Groups | GLIDER | ARROW | ARROW point | MACE-L |
|---|---|---|---|---|---|---|
| Panel I | 4 | 1 | 0.218 | 0.190 | 0.195 | 0.541 |
| Panel II | 16 | 4 | 0.288 | 0.404 | 0.412 | 0.527 |
| Panel III | 16 | 4 | 0.211 | 0.241 | 0.247 | 0.519 |
| Liquid-derived | 8 | 2 | 0.333 | 0.386 | 0.394 | 0.577 |
| Shell size | 24 | 2 | 0.403 | 0.388 | 0.396 | 0.730 |
| All covered | 68 | 13 | 0.283 | 0.332 | 0.339 | 0.564 |
| Without shell size | 44 | 11 | 0.262 | 0.322 | 0.329 | 0.534 |

The paired ARROW-minus-GLIDER difference over all covered groups was 0.0487, with a 95% solute-cluster bootstrap percentile interval $[-0.0174, 0.1292]$. Without the shell-size groups it was 0.0603 $[-0.0084, 0.1435]$. Excluding both the shell-size groups and later-reused Panel I left 40 configurations and ten solutes, with means 0.266 for GLIDER, 0.335 for ARROW and 0.534 for MACE-L; the paired interval was $[-0.0052, 0.1588]$. Thus the observed average ordering does not depend on Panel I or the shell-size series, but none of these small common-subset intervals resolves a GLIDER–ARROW difference. On the 44-case subset where the additional direct-ESP comparators were also archived, MACE-M, MACE-S and AIMNet2 scored 0.617, 0.676 and 0.786, respectively.

The nested size diagnostic provides a different comparison. The two ARROW- typable solutes supplied three parent snapshots each, with nested 1, 3, 6 and 12-water environments. A probe belongs to the solute surface when its nearest atom in the full cluster is a solute atom. For each configuration we average the signed predicted-minus-QM potential over those probes, then average the six configurations at each water count. This is a signed error, not a solute-only response or a total interaction energy. For 1, 3, 6 and 12 waters, respectively, mean response-ESP NRMSE over the same six configurations was 0.376, 0.319, 0.428 and 0.489 for GLIDER, compared with 0.384, 0.370, 0.415 and 0.413 for ARROW. These case-level scores are released with the shell-size predictions.

Table S21: **Mean signed response-potential error on solute-nearest probes for the two ARROW-typable shell-size solutes.** Values are in m$E_h$/e and average six configurations at each water count. A value near zero can reflect cancellation; the corresponding per-size NRMSE values are given in the preceding text.

| Waters | GLIDER | ARROW | MACE-L |
|---|---|---|---|
| 1 | +0.13 | −0.28 | +0.44 |
| 3 | +0.05 | −0.20 | +1.09 |
| 6 | −0.62 | −0.30 | +1.60 |
| 12 | −1.78 | +0.20 | +1.83 |

ARROW was also evaluated on the one typable training solute in the direct QM fragment-separation check, `dev_cyclic_carbamate`. In the single-water trajectory, its supplied response-potential NRMSE was 0.169 at 3 Å and 0.121 at 10 Å, versus 0.342 and 12.378 for GLIDER on the same QM reference and probes. The ARROW response is identically zero at and beyond 15 Å in these open-cluster runs because

their real-space electrostatic cutoff is 15 Å and periodic long-range electrostatics was not used. The cutoff therefore cannot establish ARROW's physical long-range limit. The ARROW separation metrics were supplied as case-level summaries; unlike the 68 compact-configuration predictions, the underlying ARROW separation probe arrays were not present in the transferred archive. The QM and GLIDER columns were checked against the released direct-QM summary, but the ARROW separation scores could not be independently rescored from that archive.

The public companion's `experiments/arrow_comparison/` directory gives all 136 compact-case ARROW field arrays, coverage records, the exact common subset, signed-bias source table, separation summary and the audit script. The compact-case scorer verified every archive prediction hash and exact probe-point alignment, then reproduced all ARROW, GLIDER and MACE-L NRMSEs from their arrays. No ARROW force-field parameter database or private raw simulation files are included. The stored QM water-cluster counterpoise interaction energies used the stated D3(BJ) correction; energy agreement is not used here as evidence that ARROW's stronger field has a particular exchange-compensation mechanism.

## 16 Reproduction

The public companion at https://github.com/Scientific-Computing-Lab/GLIDER_AI holds the scientific data and code, while the manuscript sources are maintained separately. From the companion repository root with Python 3.11 or newer, the lightweight release check and raw-array rescoring are:

```
python -m pip install -e .
python scripts/reproduce/verify_release.py
python scripts/verify_companion.py
python scripts/reproduce/score_experiment.py --experiment all
python scripts/reproduce/recompute_statistics.py --output build/statistics
python scripts/reproduce/score_arrow_release.py
```

These commands verify required hashes, the exact 24/24 training split, checkpoint and charge/moment constraints and recompute the stored experiment scores without rerunning SCF. The last command verifies the independent ARROW common-subset arrays and probe alignment and recomputes its paired intervals and signed shell-size errors; it does not rerun ARROW's self-consistent cloud model. The release's `docs/reproduce.md` also explains a density-to-coupling recalculation, figure-data regeneration and frozen-control commands. The post hoc non-water contact audit uses the stored counterpoise component energies and can be regenerated with

```
uv run --no-project --with 'ase==3.29.0' \
  python scripts/reproduce/audit_nonwater_contacts.py
python scripts/figures/plot_nonwater_contact_audit.py
```

The complete post-audit follow-up, including geometries, frozen predictions and all QM reference arrays, is in `experiments/nonwater_contact/`. Its scores and paired intervals can be regenerated without new SCF calculations with

```
python scripts/diagnostics/score_nonwater_contact_panel.py \
  --base experiments/nonwater_contact
```

The frozen-control commands require official MACE dependencies. Fresh feature extraction can vary slightly with runtime even though archived arrays rescore exactly. The translated-fragment separation test has frozen predictions for all 280 configurations. The later 20-reference check for `dev_cyclic_carbamate`, with geometries, probe arrays, component energies and a case-level results table, is in `experiments/dissociation_qm/`.

## References


S. F. Boys and F. Bernardi. The calculation of small molecular interactions by the differences of separate total energies. some procedures with reduced errors. *Molecular Physics*, 19:553–566, 1970. doi: 10.1080/00268977000101561.

Bogumil Jeziorski, Robert Moszynski, and Krzysztof Szalewicz. Perturbation theory approach to intermolecular potential energy surfaces of van der waals complexes. *Chemical Reviews*, 94:1887–1930, 1994. doi: 10.1021/cr00031a008.

Rustam Z. Khaliullin, Erika A. Cobar, Rohini C. Lochan, Alexis T. Bell, and Martin Head-Gordon. Unravelling the origin of intermolecular interactions using absolutely localized molecular orbitals. *The Journal of Physical Chemistry A*, 111:8753–8765, 2007. doi: 10.1021/jp073685z.

Jeng-Da Chai and Martin Head-Gordon. Long-range corrected hybrid density functionals with damped atom–atom dispersion corrections. *Physical Chemistry Chemical Physics*, 10:6615–6620, 2008. doi: 10.1039/B810189B.

Stefan Grimme, Jens Antony, Stephan Ehrlich, and Helge Krieg. A consistent and accurate ab initio parametrization of density functional dispersion correction for the 94 elements h–pu. *The Journal of Chemical Physics*, 132:154104, 2010. doi: 10.1063/1.3382344.

Florian Weigend and Reinhart Ahlrichs. Balanced basis sets of split valence, triple zeta valence and quadruple zeta valence quality for h to rn: Design and assessment of accuracy. *Physical Chemistry Chemical Physics*, 7:3297–3305, 2005. doi: 10.1039/B508541A.

Ilyes Batatia, William J. Baldwin, Domantas Kuryla, Joseph Hart, Elliott Kasoar, Alin M. Elena, Harry Moore, Mikolaj J. Gawkowski, Benjamin X. Shi, Venkat Kapil, Panagiotis Kourtis, Ioan-Bogdan Magdau, and Gabor Csanyi. Mace-polar-1: A polarisable electrostatic foundation model for molecular chemistry. *arXiv preprint arXiv:2602.19411*, 2026. URL https://arxiv.org/abs/2602.19411.

David L. Mobley and J. Peter Guthrie. FreeSolv: a database of experimental and calculated hydration free energies, with input files. *Journal of Computer-Aided Molecular Design*, 28:711–720, 2014. doi: 10.1007/s10822-014-9747-x.

Andrea Grisafi, Alberto Fabrizio, Benjamin Meyer, David M. Wilkins, Clemence Corminboeuf, and Michele Ceriotti. Transferable machine-learning model of the electron density. *ACS Central Science*, 5 (1):57–64, 2019. doi: 10.1021/acscentsci.8b00551.

Peter Bjørn Jørgensen and Arghya Bhowmik. Equivariant graph neural networks for fast electron density estimation of molecules, liquids, and solids. *npj Computational Materials*, 8(1):183, 2022. doi: 10.1038/s41524-022-00863-y.

Alberto Fabrizio, Andrea Grisafi, Benjamin Meyer, Michele Ceriotti, and Clemence Corminboeuf. Electron density learning of non-covalent systems. *Chemical Science*, 10:9424–9432, 2019. doi: 10.1039/C9SC02696G.

Michael G. Darley, Chris M. Handley, and Paul L. A. Popelier. Beyond point charges: Dynamic polarization from neural net predicted multipole moments. *Journal of Chemical Theory and Computation*, 4:1435–1448, 2008. doi: 10.1021/ct800166r.

Linfeng Zhang, Mohan Chen, Xifan Wu, Han Wang, Weinan E, and Roberto Car. Deep neural network for the dielectric response of insulators. *Physical Review B*, 102:041121, 2020. doi: 10.1103/PhysRevB.102.041121.

Michael Gastegger, Kristof T. Schütt, and Klaus-Robert Müller. Machine learning of solvent effects on molecular spectra and reactions. *Chemical Science*, 12:11473–11483, 2021. doi: 10.1039/D1SC02742E.

Kirill Zinovjev. Electrostatic embedding of machine learning potentials. *Journal of Chemical Theory and Computation*, 19:1888–1897, 2023. doi: 10.1021/acs.jctc.2c00914.

Ang Gao and Richard C. Remsing. Self-consistent determination of long-range electrostatics in neural network potentials. *Nature Communications*, 13:1572, 2022. doi: 10.1038/s41467-022-29243-2.

Kit Joll, Philipp Schienbein, Kevin M. Rosso, and Jochen Blumberger. Machine learning the electric field response of condensed phase systems using perturbed neural network potentials. *Nature Communications*, 15:8192, 2024. doi: 10.1038/s41467-024-52491-3.

Alan M. Lewis, Paolo Lazzaroni, and Mariana Rossi. Predicting the electronic density response of condensed-phase systems to electric field perturbations. *The Journal of Chemical Physics*, 159:014103, 2023. doi: 10.1063/5.0154710.

Yunqi Shao, Linnéa Andersson, Lisanne Knijff, and Chao Zhang. Finite-field coupling via learning the charge response kernel. *Electronic Structure*, 4:014012, 2022. doi: 10.1088/2516-1075/ac59ca.

Chaoqiang Feng, Yaolong Zhang, and Bin Jiang. Efficient sampling for machine learning electron density and its response in real space. *Journal of Chemical Theory and Computation*, 21:691–702, 2025. doi: 10.1021/acs.jctc.4c01355.

Kaycee Low, Michelle L. Coote, and Ekaterina I. Izgorodina. Inclusion of more physics leads to less data: Learning the interaction energy as a function of electron deformation density with limited training data. *Journal of Chemical Theory and Computation*, 18:1607–1618, 2022. doi: 10.1021/acs.jctc.1c01264.

Kaycee Low, Michelle L. Coote, and Ekaterina I. Izgorodina. Accurate prediction of three-body intermolecular interactions via electron deformation density-based machine learning. *Journal of Chemical Theory and Computation*, 19:1466–1475, 2023. doi: 10.1021/acs.jctc.2c00984.

Almaz Khabibrakhmanov, Matteo Gori, Carolin Müller, and Alexandre Tkatchenko. Noncovalent interactions in density functional theory: All the charge density we do not see. *Journal of the American Chemical Society*, 147:40763–40775, 2025. doi: 10.1021/jacs.5c13706.

Peichen Zhong, Dongjin Kim, Daniel S. King, and Bingqing Cheng. Machine learning interatomic potential can infer electrical response. *npj Computational Materials*, 11:384, 2025. doi: 10.1038/s41524-025-01911-z.

Xiaoyu Wang and Bingqing Cheng. Distilling latent electrostatics from foundation machine learning interatomic potentials. *arXiv preprint arXiv:2606.15001*, 2026. URL https://arxiv.org/abs/2606.15001.

Yoonjae Park, Dongjin Kim, Daniel S. King, Nam H. Đào, Roya Savoj, Sebastien Hamel, Xiaoyu Wang, and Bingqing Cheng. Polarizable atomic multipoles for learning long-range electrostatics. *arXiv preprint arXiv:2605.05746*, 2026. URL https://arxiv.org/abs/2605.05746.

Ge Song and Weitao Yang. Nepoip/mm: Toward accurate biomolecular simulation with a machine learning/molecular mechanics model incorporating polarization effects. *Journal of Chemical Theory and Computation*, 21:5588–5598, 2025. doi: 10.1021/acs.jctc.5c00372.

Dylan M. Anstine, Roman Zubatyuk, and Olexandr Isayev. Aimnet2: a neural network potential to meet your neutral, charged, organic, and elemental-organic needs. *Chemical Science*, 16:10228–10244, 2025. doi: 10.1039/D4SC08572H.

Nils Gonnheimer, Karsten Reuter, Venkat Kapil, and Johannes T. Margraf. Mace-mdp: A general dipole and polarizability model for organic molecules and materials. *ChemRxiv*, 2026. doi: 10.26434/chemrxiv.15000716/v2. URL https://doi.org/10.26434/chemrxiv.15000716/v2.

Christoph Bannwarth, Sebastian Ehlert, and Stefan Grimme. Gfn2-xtb—an accurate and broadly parametrized self-consistent tight-binding quantum chemical method with multipole electrostatics and density-dependent dispersion contributions. *Journal of Chemical Theory and Computation*, 15: 1652–1671, 2019. doi: 10.1021/acs.jctc.8b01176.

Leonid Pereyaslavets, Ganesh Kamath, Oleg Butin, Alexey Illarionov, Michael Olevanov, Igor Kurnikov, Serzhan Sakipov, Igor Leontyev, Ekaterina Voronina, Tyler Gannon, Grzegorz Nawrocki, Mikhail Darkhovskiy, Ilya Ivahnenko, Alexander Kostikov, Jessica Scaranto, Maria G. Kurnikova, Suvo Banik, Henry Chan, Michael G. Sternberg, Subramanian K. R. S. Sankaranarayanan, Brad Crawford, Jeffrey Potoff, Michael Levitt, Roger D. Kornberg, and Boris Fain. Accurate determination of solvation free energies of neutral organic compounds from first principles. *Nature Communications*, 13:414, 2022. doi: 10.1038/s41467-022-28041-0.