Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

A reusable solvent-response framework for molecular hydration prediction

Gal Oren
Stanford University
Boris Fain
Stanford University
Michael Levitt
levittm@stanford.edu

Stanford University

Article

Keywords:

Posted Date: November 10th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11334582/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

A reusable solvent-response framework for molecular hydration prediction

Gal Oren*, Boris Fain and Michael Levitt*

Stanford University, Stanford, California, USA

*Correspondence: galoren@stanford.edu (G.O.) and levittm@stanford.edu (M.L.)

Code and data: https://github.com/Scientific-Computing-Lab/SolvAI

Abstract

Solvent response governs hydration free energy, yet high-fidelity calculations resolve it afresh for every new molecule. We ask whether information already encoded in disparate solvent datasets can instead be learned once and reused. SolvAI trains molecular surrogates on six calculated and empirical sources, predicts 15 complementary response descriptors from a SMILES string, and combines them with structural features in a separate predictor fitted to experimental hydration free energies. Inference requires no new solvent simulation. In a matched ExtraTrees comparison, the descriptors reduce mean absolute error from 0.303 to 0.202 kcal mol^{-1} on 85 ARROW solutes and from 1.532 to 1.153 kcal mol^{-1} on an endpoint-disjoint external cohort of 220 molecules. Shuffling descriptors among molecules removes the gain; it persists on the 97-molecule subset also absent from supervised response sources. The same descriptors often help other endpoint families, although the best complete predictor and the most useful source mixture change with the evaluation chemistry. Existing solvent calculations can therefore provide reusable supervision for molecular prediction, while source choice and transfer beyond small neutral solutes remain important tests.

Introduction

Hydration free energy is the reversible work of transferring a molecule from the gas phase into water. It connects molecular structure to partitioning, binding and reactivity, and is a stringent test of how a model balances electrostatics, polarization, dispersion, repulsion and solvent reorganization. Experimental values are available for many small neutral compounds, notably through FreeSolv, but predicting them with comparable precision remains difficult.[1;2]

Physical simulation resolves solvent response directly. Alchemical methods vary the solute–water interaction along a coupling path and combine the resulting ensemble averages; path-integral molecular dynamics (PIMD) can also represent nuclear delocalization.[3;4] The cost is that each new molecule calls for another sampling calculation. On 85 neutral solutes, the ARROW study made the accuracy at stake concrete: classical ARROW has a reconstructed mean absolute error (MAE) of 0.785 kcal mol^{-1}, falling to 0.205 kcal mol^{-1} with eight-bead PIMD (PIMD8).[5] Learned-potential extensions reproduce parts of that physical description, but still use trajectories at inference.[6;7]

Prediction from molecular structure offers a different economy: train on earlier examples, then answer a new query without repeating their calculations. Graph and three-dimensional predictors,

Δ-learning and multi-fidelity methods use this principle in different ways.[8–11] Physically inspired descriptors can bring electrostatics, accessible surface or conformational response into such models.[12;13] Yet approaches that calculate a continuum or molecular solvent response for each query retain a per-molecule physical step.[14–17] Can that solvent information itself become a reusable molecular representation?

Predicting a physical descriptor before fitting a downstream property is an established route: quantum-chemical descriptors have been learned for reactivity, and a two-stage hydration model has predicted solute descriptors instead of calculating them for every query.[18;19] The usefulness of such a descriptor depends on its relevance, how accurately it can be predicted and how the downstream model uses it.[20;21] Here the question is whether several complementary kinds of solvent information, learned from different sources, remain useful across hydration predictors and unfamiliar molecules.

SolvAI uses the existing datasets as six groups of teachers (Fig. 1). They supply 15 targets spanning calculated transfer free energies, empirical solute properties and summaries of how water changes conformer energies. Molecular surrogates learn to predict these quantities from structure and are frozen before a separate endpoint learns experimental hydration free energy. We call each predicted quantity a *response descriptor*; the name does not imply a derivative with respect to an applied field. Because the two learning stages are separate, we can change the hydration predictor while holding the descriptors fixed, or change their surrogates while holding the source targets fixed.

We first ask whether the predicted responses improve a matched hydration predictor and whether the gain survives controls that break molecule alignment or separate training from test chemistry. We then hold the response representation fixed and vary the endpoint, distinguishing the value of adding responses to a given learner from the accuracy of the complete learner. Finally, we vary the information used to build the representation and test its behavior on larger molecules. The later architecture comparisons are retrospective; their selection protocol is detailed in Methods.

SolvAI: learning solvent response before hydration free energy

Two learning tasks with different supervision

For a neutral molecule represented by a SMILES string x, the target is its experimental gas-to-water transfer free energy $y = \Delta G_{\mathrm{hyd}}$ near 298 K. We first train separate molecular models on external source datasets. Each model predicts one or more source quantities, such as a continuum-solvent free energy or a hydrogen-bonding parameter. These models are *surrogates*: after training, they approximate the source quantities without repeating the source calculation or measurement. Their weights are frozen before learning the experimental hydration predictor.

Writing the frozen surrogate predictions as $\hat{r}_j(x)$, our model is

$$\hat{\boldsymbol{r}}(x) = (\hat{r}_1(x), \ldots, \hat{r}_{15}(x)), \qquad \hat{y}(x) = h([\boldsymbol{s}(x), \hat{\boldsymbol{r}}(x)]), \tag{1}$$

where $\boldsymbol{s}(x)$ denotes structural features; the standard tabular input contains a 2,048-bit Morgan fingerprint and 217 RDKit two-dimensional descriptors. The endpoint h learns how the predicted source quantities relate to experiment; it does not sum them as if they were components of one physical free energy. The primary comparisons keep the same 15 frozen predictions while varying h and, in some models, the structural representation (Table 2). All 15 quantities are predicted at deployment, including those whose original source targets were experimental.

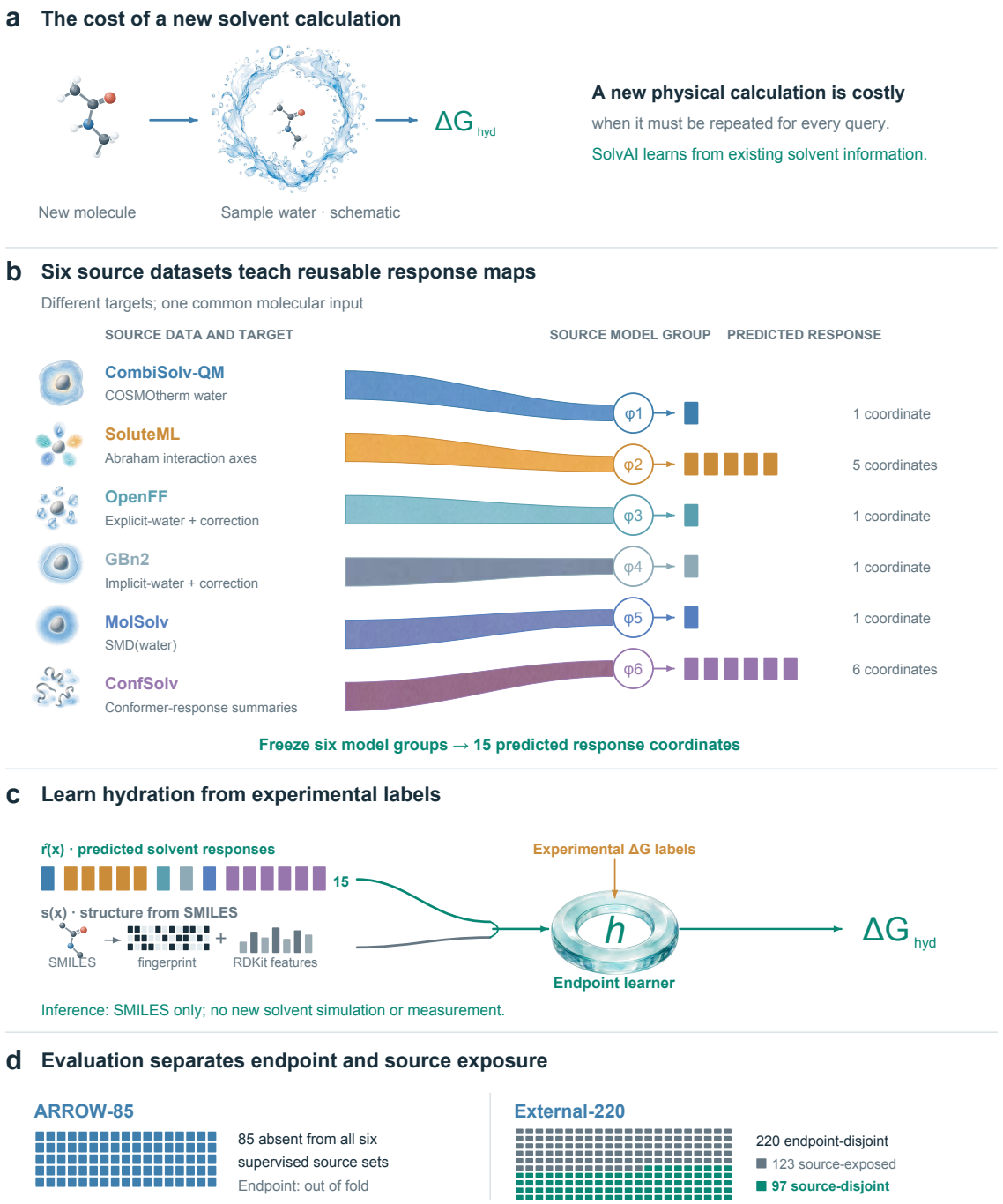


Figure 1: Existing solvent information becomes reusable molecular supervision. **a**, Explicit-solvent hydration calculations sample a water ensemble for each molecule to estimate ΔG_{hyd}. The molecular drawing and solvent halo are schematic, not simulation output. **b**, Six source datasets teach separate structure-to-response model groups, frozen before endpoint fitting. Their $1+5+1+1+1+6$ predicted coordinates total 15; a group can contain target-specific fits. OpenFF and GBn2 combine a physical-energy prediction with a learned experimental residual. Icons and colored ribbons distinguish sources; they do not depict measured response fields or magnitudes. **c**, A separate hydration endpoint learns from predicted responses and SMILES-derived structural features using experimental hydration labels. The molecule, fingerprint and RDKit-feature glyphs show input types, not actual feature values. Inference requires only SMILES, with no new solvent calculation. **d**, Each tile represents one evaluation molecule: 85 source-disjoint ARROW molecules scored out of fold; 220 endpoint-disjoint external molecules, including 123 source-exposed and 97 source-disjoint. Strict-97 is nested in External-220. Identity checks include standardized equivalents. The diagram is architecture-neutral; source learners and endpoints are compared in Supplementary Table 17 and Table 2.

What the response descriptors represent

The sources offer different views of solvation rather than successive rungs on one fidelity scale (Table 1; Fig. 1b). Continuum-solvent calculations estimate transfer energetics; Abraham parameters describe empirical interaction tendencies; corrected explicit- and implicit-solvent estimates combine a physical prediction with an experimental residual. ConfSolv adds information about how water redistributes conformer energies and solvation responses. Together they give the endpoint several kinds of evidence about a molecule, rather than asking one source to stand for the whole solvent problem. The source methods and their physical definitions are documented in the original literature and Supplementary Methods S3 and Table 1.[9;22–27]

Table 1: The 15 response descriptors and the information they provide. Every listed quantity is predicted from molecular structure at inference. CombiSolv-QM supplies COSMOtherm targets, SoluteML supplies Abraham parameters, and MolSolv supplies SMD(water) targets. The last column describes the released implementation, not an architectural requirement of the framework. Complete definitions, units and source counts are in Supplementary Table 1.

Source	Number	Predicted quantity	Released surrogate
CombiSolv-QM	1	COSMOtherm water transfer free energy	D-MPNN
SoluteML	5	Abraham E, S, A, B, L interaction parameters	ExtraTrees
OpenFF	1	Explicit-solvent estimate plus experimental residual	ExtraTrees
GBn2	1	Implicit-solvent estimate plus experimental residual	ExtraTrees
MolSolv	1	SMD continuum-water transfer free energy	D-MPNN
ConfSolv	6	Conformer corrections and water-response distributions	LightGBM

The source archives contain many calculations per molecule. MolSolv's 1,729,545 records are SMD(water) calculations; ConfSolv's 5,392,567 are water-conformer records. After quality, target and identity filtering, the respective surrogate tables retain 350,359 and 17,829 molecular connectivities. Archive records and training rows are different units of counting (Supplementary Table 1 and Fig. 2).[28]

Hydration free energy is an equilibrium property of a molecule, not a separate experimental label for each constrained geometry. ConfSolv's six coordinates summarize conformational response without reconstructing an ensemble at inference. Supplementary Fig. 7 illustrates this distinction, which also matters when comparing the response features with Uni-Mol's generated three-dimensional conformer.[29]

Results

Separating source supervision from hydration evaluation

Two experimental settings ask different questions. ARROW-85 provides experimental hydration values and a published high-fidelity simulation result for the same small molecules; its folds test the learned predictor out of sample. A separate cohort of 220 neutral solutes, drawn from original Henry-law measurements, tests prediction outside the endpoint training labels. All comparisons of SolvAI in this paper use experiment as the reference; the ARROW simulation is an accuracy context, not a model input.

The endpoint training pool contains 1,280 public hydration labels plus available ARROW training folds. The response-source tables exclude every exact or standardized equivalent of the 85 ARROW test solutes. All 220 external molecules are absent from endpoint labels, and a nested set of 97 is absent from the supervised response sources as well. We distinguish

these two forms of exposure throughout because a source calculation for a molecule is not its experimental hydration label. Exclusions and affected surrogate refits are documented in Methods, Supplementary Table 2 and Supplementary Data 4.

A matched test of reusable response information

We begin with a matched question: do the predicted responses add information when the hydration learner stays the same? An ExtraTrees ensemble of randomized decision trees[30] is trained with or without the 15 response columns, using identical labels, folds, structural features and training recipe. Each ARROW solute is predicted out of fold by a fit that did not receive its experimental label. Structural features alone yield an MAE of 0.303 kcal mol^{-1}; the first eight response descriptors reduce it to 0.212, SMD(water) to 0.207, and the six conformer summaries to 0.202 kcal mol^{-1} (Fig. 2b). That final value is numerically close to the published ARROW/PIMD8 error of 0.205 on these solutes. PIMD8 is a separate physical simulation, while this endpoint learns from experimental hydration labels; the matched test here is the comparison with and without response descriptors.

The gain is distributed across molecules rather than driven by a single large correction: absolute error falls for 53 of 85 solutes, and the paired MAE change is −0.101 kcal mol^{-1} (95% bootstrap interval, −0.215 to −0.020; Fig. 2c). The full set also improves more than the tested computed-energy, SMD-only or ConfSolv-only blocks in isolation (Fig. 2d). The empirical and residual-corrected block contributes on its own, suggesting that the sources supply partly complementary information. Molecule-level predictions and residuals are in Supplementary Fig. 1 and Data 2.

Shuffling tests whether the benefit depends on assigning the responses to the correct molecule (Fig. 2a). Across five independent permutations of the 15-column block, the fixed-partition mean is 0.306 kcal mol^{-1}, close to the 0.303 matched control and far from the aligned-response fit. The paired difference between aligned and mean-shuffled descriptors is −0.104 kcal mol^{-1} (95% interval, −0.214 to −0.025). Alignment between a molecule and its predicted response is therefore necessary for the observed gain.

Changing the ARROW partition changes the absolute estimate but preserves the matched contrast. Across five complete five-fold partitions, the response-augmented endpoint has an average MAE of 0.207 ± 0.004 kcal mol^{-1}, versus 0.313 for structure alone, and improves in every partition (Fig. 2e; Supplementary Table 4). Reducing the preset weight of available ARROW training labels from 3 to 1 also preserves the contrast (Supplementary Table 9).

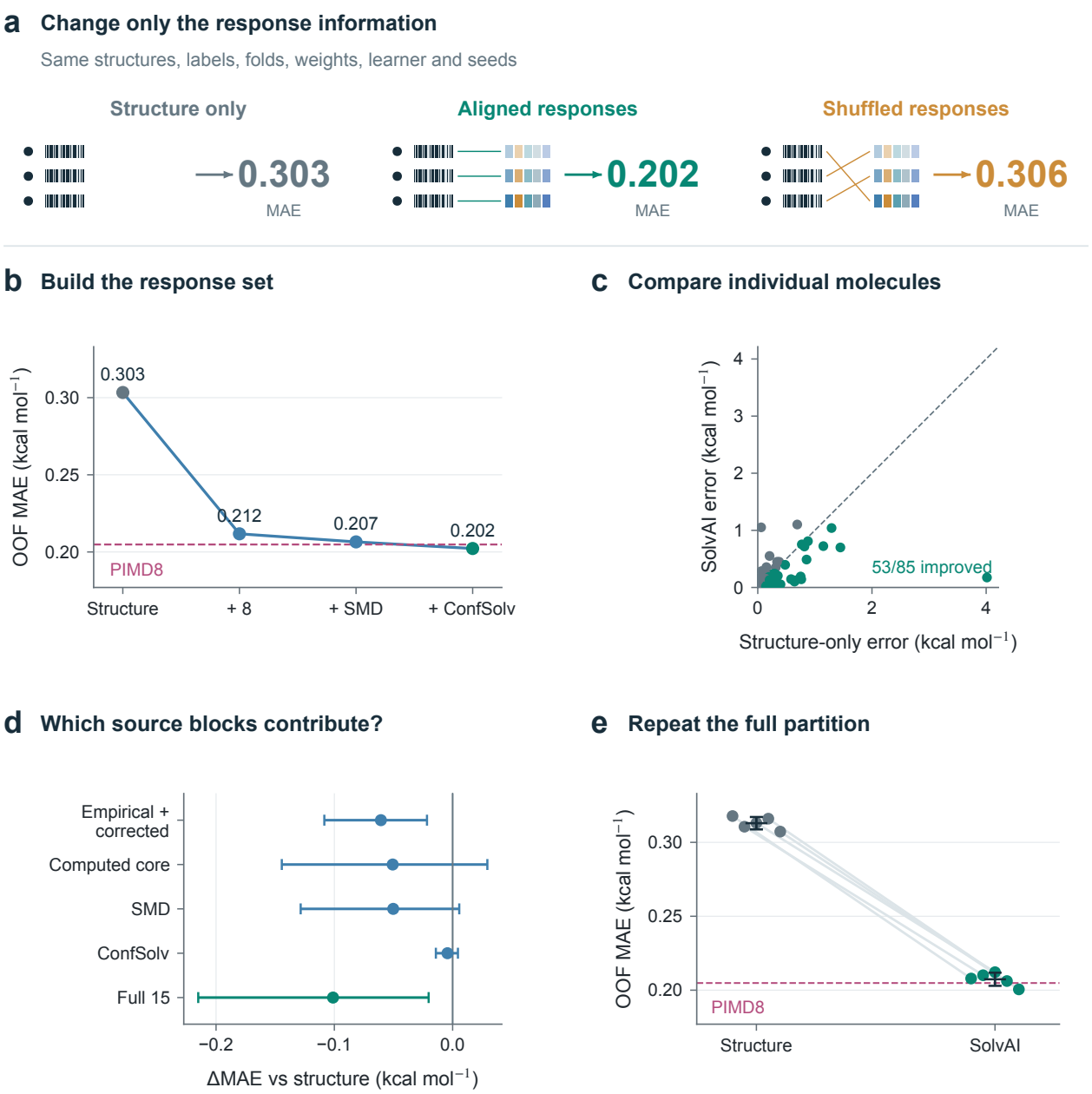


Figure 2: Molecule-aligned response descriptors improve a matched hydration endpoint. **a**, The matched control removes the responses; the shuffled control reassigns the complete response block among molecules. Labels, structural descriptors, architecture, weights, folds and seeds are fixed. The assignment sketches are illustrative; shuffled MAE is the mean of five permutations. **b**, Fixed-partition progression adds eight COSMOtherm, Abraham and corrected-solvent descriptors, then SMD(water), then six ConfSolv summaries. The dashed magenta line marks ARROW/PIMD8 on the same 85 solutes, as an accuracy reference rather than an input or training source. **c**, Paired absolute errors in kcal mol^{-1} for every molecule. Points below the diagonal mark lower error with responses (teal); gray points do not improve. **d**, MAE change for predeclared response blocks relative to structure only; bars are 95% molecule-bootstrap intervals from 100,000 resamples. **e**, Paired results for five complete split repeats. Thin lines join methods within the same partition; black marks show mean ± sample s.d.

Does the response help beyond one endpoint?

The matched test raises a second question: does the information in the 15 frozen descriptors remain useful when the hydration learner changes? We tested nine fixed endpoint configurations, from trees and neural networks to molecular graphs and pretrained encodings (Table 2). For each one, a paired fit removes only the response columns while retaining its selected structural inputs and training recipe. Thus the comparison asks both how accurate the complete learners are and what the responses contribute within each learner. TabM uses a shared-parameter neural ensemble, and TabPFN predicts from a pretrained transformer conditioned on labelled training rows.[31–33] The study is retrospective; configurations were selected using training-side validation, with the full protocol in Methods and Supplementary Methods S18.

Table 2: Endpoint mechanisms and their measured accuracy. All entries include the 15 response descriptors. MAE is in kcal mol^{-1}. Boldface marks the lowest displayed MAE in each column among these fixed full-response configurations; it does not imply a statistically resolved rank. The first three columns use the ARROW and external evaluation identities; Strict-97 is nested within External-220. Neural and TabPFN predictions average five seeds, tree controls three. The final column is a separate exploratory size-held-out refit on 1,160 training and 205 held-out molecules, using three seeds for every method. A dash means not evaluated, not zero error. Supplementary Table 13 includes matched no-response controls and common-three-seed results.

Endpoint	Mechanism and observed pattern	ARROW	External	Strict	Size
ExtraTrees	Leaf averages; bounded output	0.202	1.153	1.536	1.772
Residual MLP	Residual links; external response gain	0.340	1.428	2.016	1.813
Dual branch	Separate inputs; improved size transfer	0.288	1.234	1.616	1.560
Ridge + neural residual	Linear trend + correction; low Strict error	0.336	1.172	**1.429**	1.391
Message passing	Size-aware graph; external response gain	0.288	1.168	1.505	1.729
MoLFormer + neural head	Frozen sequence encoding; low external error	0.273	**1.111**	1.450	1.609
TabM	Shared neural ensemble; low size error	0.291	1.132	1.454	1.389
TabPFN-3.5	Label-conditioned transformer; low size error	**0.199**	1.134	1.493	**1.318**
MoLFormer + ExtraTrees	Frozen encoding + trees; input contrast	0.234	1.174	1.564	–

The lowest full-model point estimate changes with the cohort: TabPFN reaches 0.199 kcal mol^{-1} on ARROW-85, a MoLFormer neural head 1.111 on External-220, and ridge plus a neural residual 1.429 on Strict-97 (Table 2). These are descriptive rankings of complete predictors. They cannot tell us whether the same predictor needed the response descriptors.

The paired fits answer that question (Fig. 3). On External-220 and its nested Strict-97 subset, seven of nine configurations have below-zero descriptive intervals for the response-minus-no-response MAE difference. The neural MoLFormer endpoint, for example, changes from 1.332 to 1.111 externally; ExtraTrees, message passing and several other learners also benefit. On ARROW-85, the intervals are below zero for ExtraTrees and the two MoLFormer endpoints. The contrast has exceptions that make the distinction useful: TabPFN has the lowest ARROW point estimate without a resolved response increment, while the lowest full-model

Strict-97 configuration, ridge plus a neural residual, has lower error when its responses are removed. The representation can therefore help across architectures without defining a single best complete model. Complete paired results, mechanism diagrams and selection variants are in Supplementary Tables 13–15, Figs. 8–10 and Data 6.

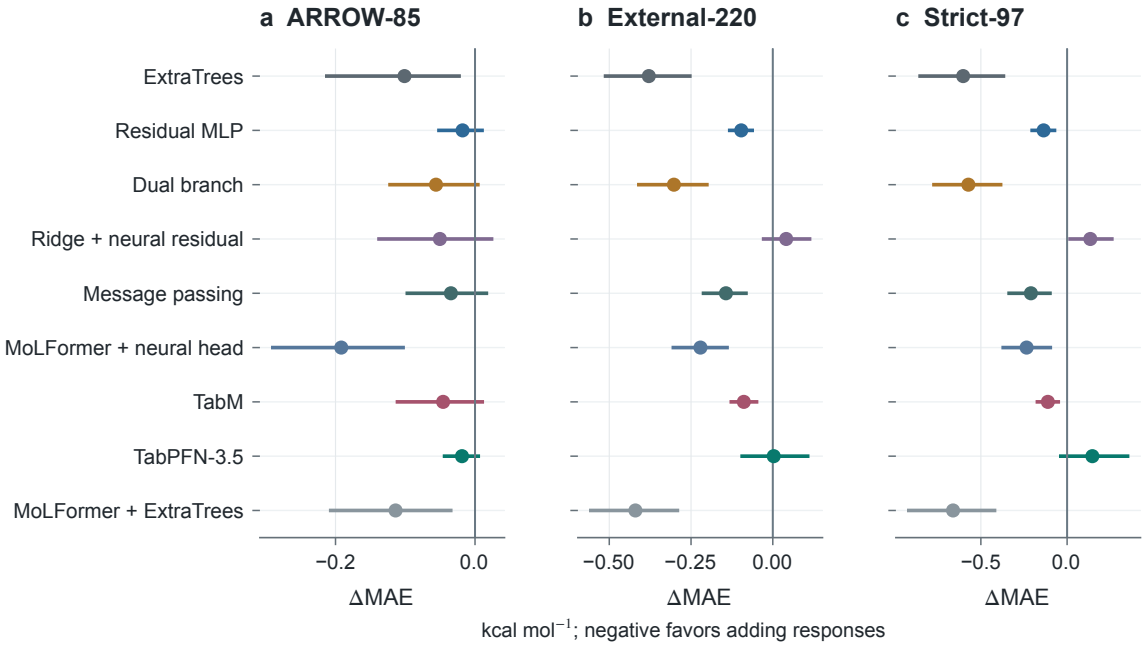


Figure 3: The response contribution varies among endpoint families. a–c, Paired MAE changes from adding the same 15 response descriptors to each selected endpoint recipe on ARROW-85, External-220 and Strict-97. Negative values favor response augmentation. Bars are descriptive 95% percentile intervals from 100,000 paired molecule resamples, conditional on the fitted models and without multiplicity correction. They are not uncertainty intervals for individual molecular predictions. Strict-97 is nested within External-220. Horizontal scales differ between cohorts. The no-response controls inherit the full model's recipe. Cross-architecture accuracy and seed counts are given in Table 2; family-specific configuration choices are retained in Supplementary Data 6.

The responses might help only because the structural input to ExtraTrees is limited. A focused comparison holds that endpoint fixed while adding the 15 descriptors to frozen MoLFormer sequence encodings or Uni-Mol three-dimensional encodings (Fig. 4a). Direct graph predictors are shown as separate references; they are not the paired ablation for either frozen encoding.

The added responses lower ARROW-85 MAE from 0.348 to 0.234 kcal mol^{-1} with MoLFormer features and from 0.334 to 0.220 with Uni-Mol features. Both paired advantages also persist on External-220 and Strict-97, with intervals below zero (Fig. 4b,c). Thus the predicted solvent quantities provide useful information even beside pretrained molecular representations. The direct graph comparisons and their training protocols are reported in Methods and Supplementary Tables 11–12; the ten-descriptor source control shown in Fig. 4 is examined with the full source set below.

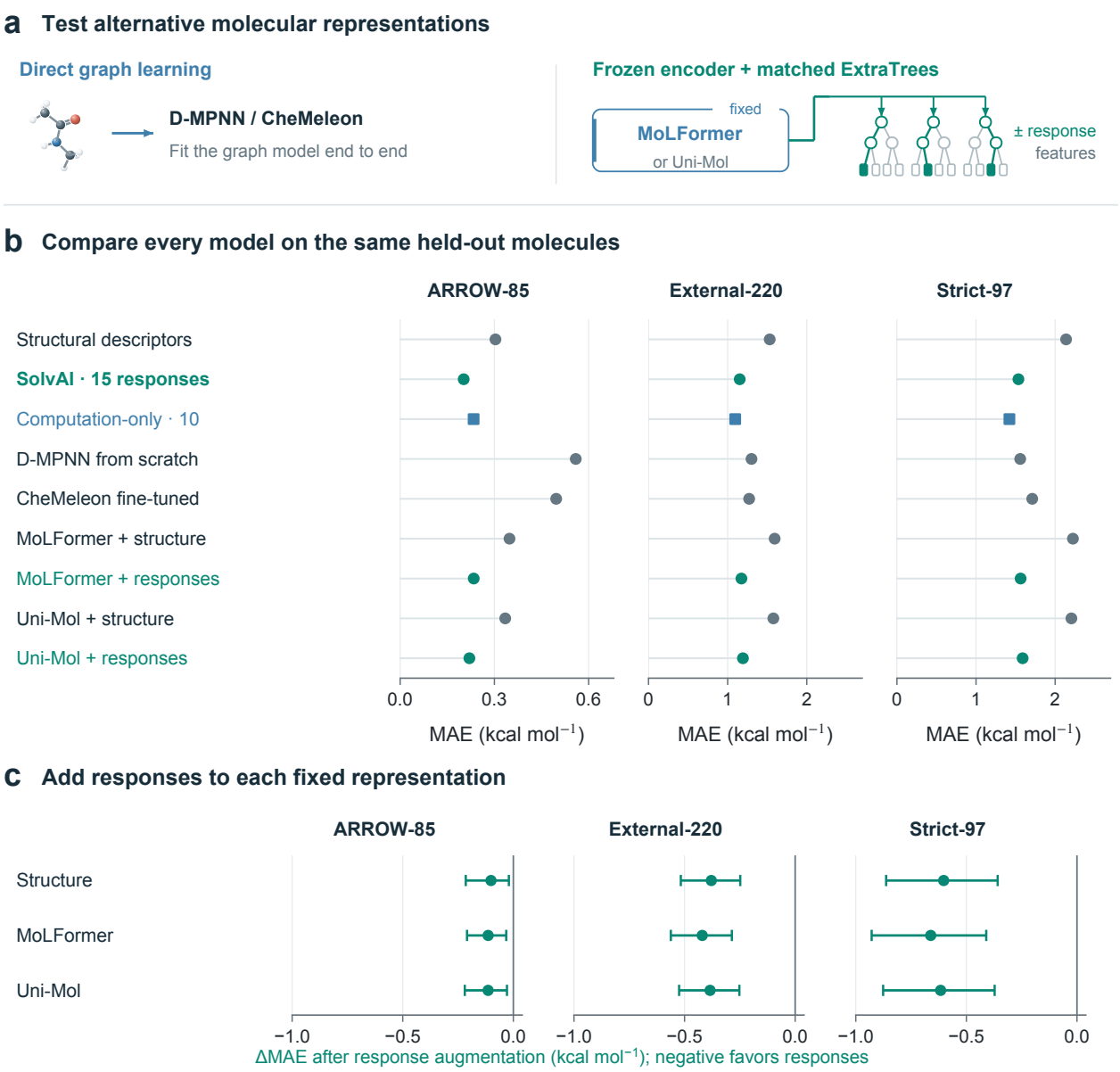


Figure 4: Response descriptors complement pretrained molecular representations. a, Direct graph models fit the entire network. Frozen MoLFormer and Uni-Mol representations instead augment the same structural features and ExtraTrees endpoint; paired fits add or remove only the 15 responses. **b**, All nine model conditions on ARROW-85, External-220 and Strict-97. Values are MAE in kcal mol^{-1}; note the different ARROW and external horizontal scales. ARROW predictions are out of fold. External predictions use all 1,365 endpoint-training labels; Strict-97 is a subset of External-220, not a separate fit. The ten-response control removes the Abraham parameters and experimental residual corrections (Fig. 6a). **c**, Paired changes after response augmentation, relative to each unaugmented representation. Negative values favor adding responses; bars are 95% intervals from 100,000 paired molecule resamples. Full numerical results and intervals are retained in Supplementary Tables 11–12.

Transfer beyond nearby labels

Random molecular folds can place close chemical neighbors on opposite sides of a train–test split. We therefore repeated the ExtraTrees comparison after excluding endpoint-training molecules in the test family's scaffold, functional family or molecular cluster, and after removing neighbors above fixed Morgan-similarity thresholds. Each rule applies to the entire endpoint-supervised pool; the response surrogates remain fixed.

Adding responses lowers MAE under every predeclared separation rule (Fig. 5a,b; Supplementary Table 5). At a similarity exclusion threshold of 0.70, the matched error falls from 0.343 to 0.211 kcal mol^{-1}. Scaffold and functional-family splits are harder in absolute terms, yet their paired intervals also favor the response features. The gain therefore persists without nearby endpoint labels; it does not make these difficult splits equally accurate. Full regime values and molecule-level family errors are in Supplementary Tables 5–6 and Fig. 5.

Removing all ARROW experimental labels from endpoint training asks whether the representation needs fold-local adaptation to those 85 solutes. Trained only on the 1,280 public labels, the structure-only endpoint has an MAE of 0.385 kcal mol^{-1}, versus 0.257 with responses (paired interval for the change, −0.246 to −0.037; Fig. 5c). This is an ARROW-label removal test; the separate external cohort tests a more distinct set of molecules.

The external cohort supplies the stronger transfer test. It was frozen from original Henry-law measurements for a separate solvation-model study before that study's predictions were examined.[34] On its 220 endpoint-disjoint neutral solutes, responses lower matched MAE from 1.532 to 1.153 kcal mol^{-1} (paired change −0.379; 95% interval, −0.517 to −0.248; Fig. 5d).

Among the nested 97 solutes absent from all six supervised response sources, the corresponding contrast is 2.138 to 1.536 (paired change −0.603; interval, −0.863 to −0.359). The strict subset's median absolute error moves in the opposite direction, from 1.102 to 1.244; the mean improvement is therefore not a gain for every molecule or every error summary (Supplementary Table 10). Absolute errors are also much larger here than on ARROW-85.

The largest absolute gain occurs where external molecules are least like the endpoint training pool (Fig. 5e). Among 35 solutes with nearest training-set Morgan similarity below 0.30, MAE is 4.027 kcal mol^{-1} from structural features and 2.501 with responses. This is an encouraging reduction in a difficult region, but the remaining error is still substantial. The full similarity-bin counts and estimates are shown in the figure.

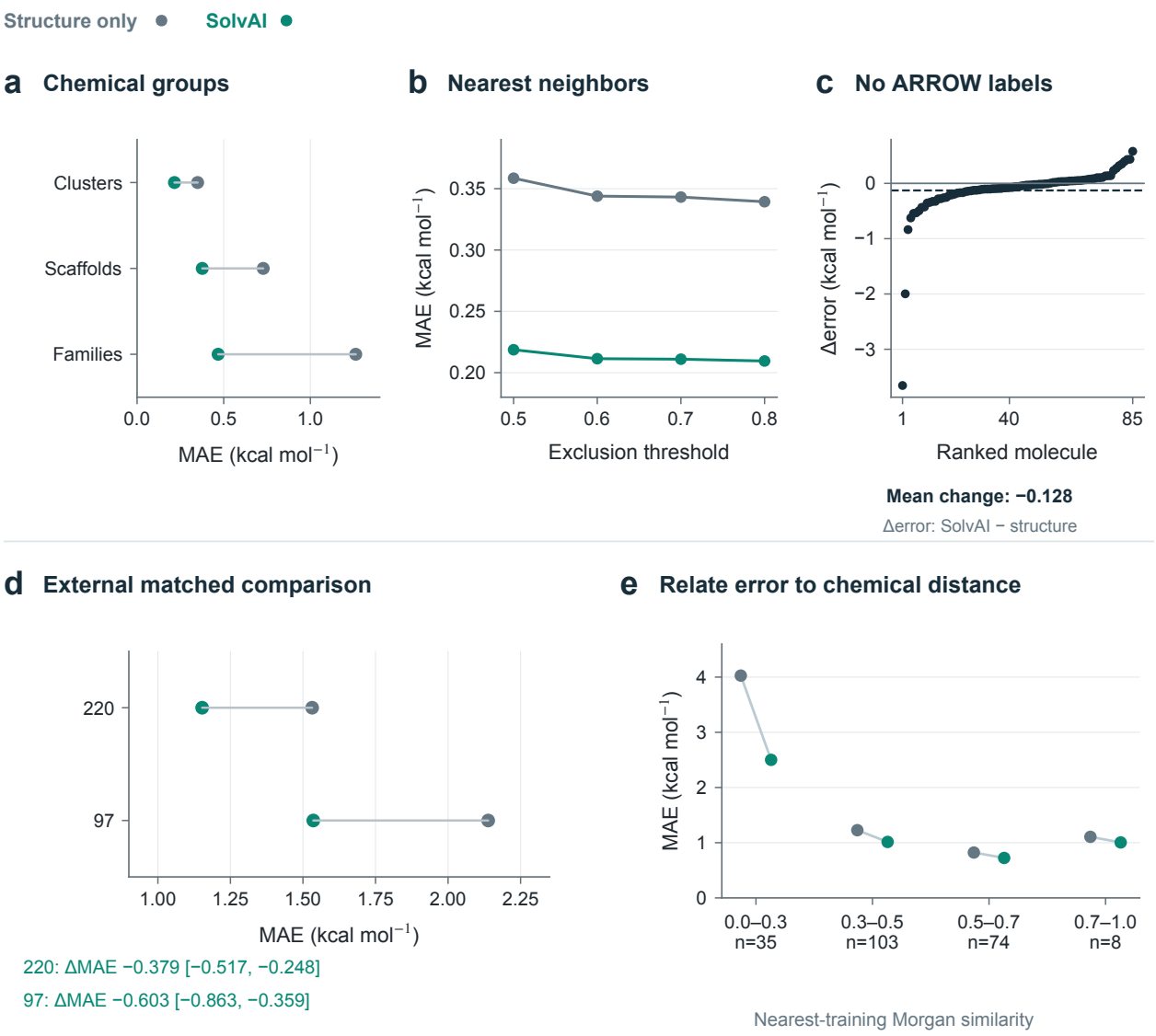


Figure 5: Response descriptors reduce error under chemical separation. **a**, Matched MAEs after removing every endpoint training label in the held-out molecular cluster, scaffold or functional family. **b**, Similarity exclusion removes every training row at or above the indicated Morgan similarity threshold to a test molecule. **c**, Molecule-level change in absolute error when no ARROW experimental label is used for training; negative values favor adding responses. **d**, Matched performance on the prospectively frozen external cohort and the subset also absent from every response-source table. Bracketed values below the plot are 95% intervals from 100,000 paired molecule resamples for the MAE difference, not intervals for the plotted MAEs. **e**, External errors in fixed nearest-training Morgan similarity bins, with all sample counts. The largest improvement occurs in the least similar bin, where absolute errors also remain largest. These bins describe the cohort, not calibrated applicability thresholds. All response surrogates use the standardized-equivalent exclusions described in Methods.

Source composition changes transfer behavior

The response set mixes computed quantities with empirical parameters and experimental residual corrections. To distinguish their roles, an additional control (Fig. 6) retains ten computed-source descriptors: COSMOtherm, uncorrected OpenFF and GBn2, SMD(water), and the six ConfSolv summaries. The predictor, experimental training labels and weights remain unchanged. On ARROW-85 this control reaches 0.234 kcal mol^{-1}; its improvement over structural features alone has a paired interval spanning zero (−0.191 to 0.022 kcal mol^{-1}). The full 15-descriptor model

is better than this control by 0.032 kcal mol^{-1} (95% interval for full minus computation-only, −0.063 to −0.006).

The ordering reverses outside the reference chemistry. The ten-descriptor model reaches 1.095 kcal mol^{-1} on the external 220 molecules, compared with 1.153 for the full model. The paired full-minus-computation-only difference is 0.057 kcal mol^{-1} (95% interval, 0.023 to 0.093). On the strict 97-molecule subset, the corresponding errors are 1.422 and 1.536 kcal mol^{-1}, with a paired interval of 0.058 to 0.172 for the difference. Computed-source descriptors therefore reduce external error relative to structural features alone, while the full source composition is more effective on ARROW. This comparison identifies a dependence on source composition and evaluation chemistry, rather than a single source set that is uniformly preferable. Both fixed configurations are reported; external errors were not used to retune either model.

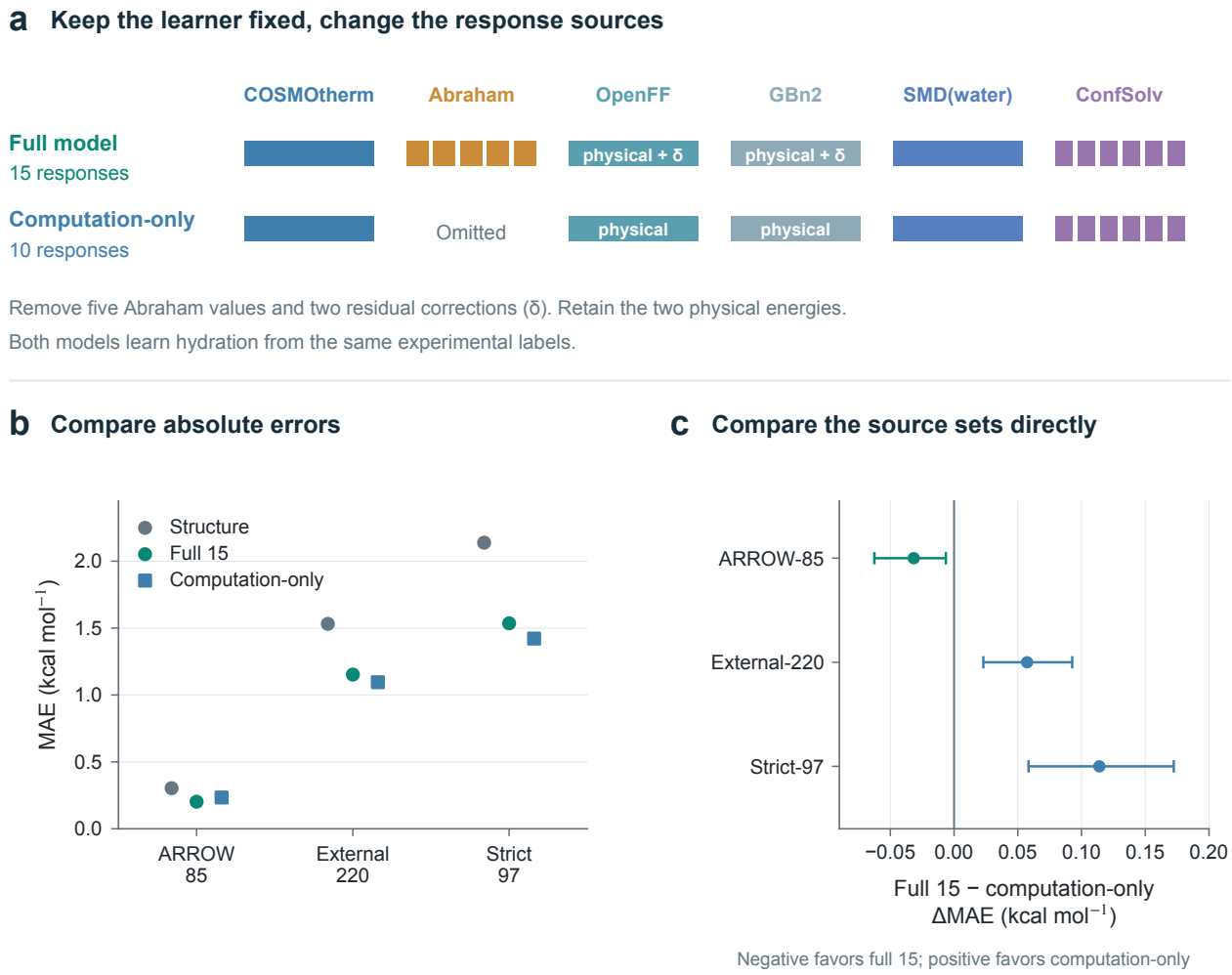


Figure 6: The preferred source composition changes with evaluation chemistry. a, The computation-only control removes five Abraham descriptors and both experimental residual corrections, δ, while retaining the underlying OpenFF and GBn2 physical-energy predictions. The resulting computation-only set contains ten descriptors. Both models retain the same structural features, experimental endpoint labels, weights and learner. Each tile denotes one descriptor, not its magnitude. **b**, Absolute MAEs for structural features, all 15 responses and the ten-response control on all three reported evaluation sets. **c**, Paired full-minus-computation-only MAE differences with 95% intervals from 100,000 molecule resamples. Negative values favor the full model on ARROW-85; positive values favor computation-only responses externally. The strict 97 are nested within the external 220. The external extension is retrospective and compares two fixed configurations without retuning on external errors. The reversal identifies a dependence on source composition and chemistry, not one uniformly optimal response set.

Changing source surrogates changes the transfer trade-off

The source mixture is only one design choice. Another is how accurately each source quantity can be inferred from structure, and whether improving that prediction helps hydration. In a retrospective challenge we held the 15 target definitions fixed and compared alternate learners for the four source families that had used trees. Source-only validation retained the released recipes for Abraham and ConfSolv but selected ridge-plus-neural-residual surrogates for OpenFF and GBn2 (Supplementary Method S22; Table 17). Their individual source targets were predicted more accurately, whereas the corrected OpenFF and GBn2 descriptors passed downstream showed smaller, unresolved changes. Better targetwise source scores did not automatically produce better hydration features.

Replacing only those two surrogate families made the trade-off visible in the hydration endpoint. With a protocol-matched ExtraTrees refit, ARROW-85 MAE moves from 0.203 to 0.238 kcal mol^{-1}, while External-220 improves from 1.192 to 1.075 and its source-disjoint Strict-97 subset from 1.595 to 1.352 (paired Strict-97 interval for the change, −0.358 to −0.137; Supplementary Table 18). A second endpoint has no clearly resolved gain, and the ExtraTrees size-held-out error rises. These results show that source-learner choice can favor one transfer setting without improving the others. The challenge does not replace the released model; source exposure and the full selection protocol are documented in Methods and Supplementary Method S22.

When response prediction is insufficient

Some plausible sources do not survive the two-stage route. In exploratory development screens, compact ConfSolv summaries helped more than the tested latent representations from the same source; additional force-field, pair-interaction and implicit-solvent features did not improve their matched endpoints (Supplementary Fig. 3). A useful source target must be relevant to hydration, predictable from structure and complementary to the information already supplied by other descriptors.

We also predicted coupling-path energy derivatives from short two-bead path-integral simulations (PIMD2). Although the surrogates captured broad trends, their component MAEs ranged from 1.27 to 5.20 kcal mol^{-1}. Adding the predicted response increased its separate exploratory endpoint error from 0.196 to 0.201 kcal mol^{-1}; no PIMD-derived coordinate enters the final model. The fuller hierarchy and direct-integration tests are in Supplementary Fig. 4 and Data 1. These outcomes test the learned descriptors in this setting, rather than rank the physical source methods themselves.

Testing transfer to larger molecules

We separate accuracy on labelled larger molecules from the shape of predictions for still larger, unlabelled series. For the first test, we held out the largest 15% within each endpoint-label source and refitted eight fixed configurations on 1,160 remaining labels. Of the 205 held-out molecules, 188 exceed the largest training molecule's 17 heavy atoms. Their labels had been available in earlier random-fold development, so this is an exploratory test of endpoint size transfer with the source surrogates held fixed, not an independent test of the whole pipeline beyond all prior source exposure.

On this held-out split, TabPFN lowers MAE from 1.772 for ExtraTrees to 1.318 kcal mol^{-1} (paired difference −0.453; 95% interval, −0.648 to −0.281; Fig. 7a). The responses also help TabPFN itself: its matched error without them is 1.959 (response-minus-control interval, −0.847 to −0.451). Several other neural endpoints improve on ExtraTrees here as well. They are useful alternatives for this split, although their separately selected input recipes mean the

cross-endpoint difference is not solely an architecture effect (Supplementary Table 15 and Data 6).

Figure 7 also follows capped glycine and alanine series without experimental peptide reference values. Those panels test how predictions change with chain length, not whether they are physically accurate.

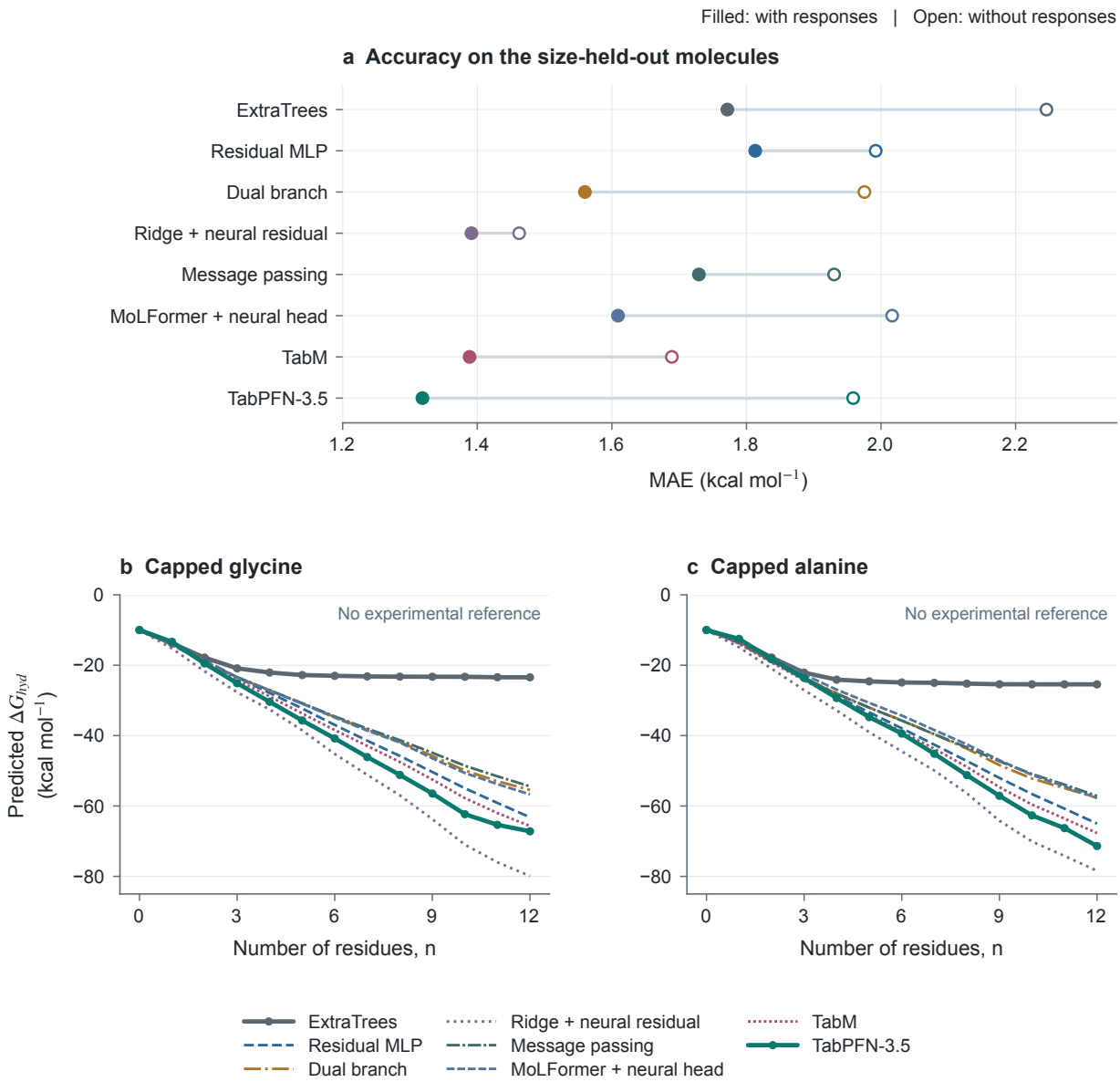


Figure 7: Labelled size transfer and unlabelled series answer different questions. a, MAE on the 205-molecule size holdout after refitting on the same 1,160 training labels. Filled and open marks denote models with and without responses. All predictions average three seeds. This diagnostic reuses development labels under a different split. **b,c**, Predictions along capped glycine and alanine series, using the full-data models with responses. Neural and TabPFN curves average five seeds; ExtraTrees averages three. These curves have no peptide reference values and therefore do not rank physical accuracy. The abscissa is the number of residues between the acetyl and N-methylamide caps; $n = 0$ is N-methylacetamide. Colors identify the same families in both panels.

In those unlabelled series, ExtraTrees predictions approach a plateau, while several neural endpoints and TabPFN keep changing as residues are added (Fig. 7b,c). That shape is less constrained by the fitted range, but it cannot be scored as more accurate without reference values. An alkane series makes the point from another direction: TabPFN is non-monotonic between C10

and C14, and most tested alkanes already occur in endpoint training. The individual predictions, source-label audit and sensitivity to two alkane labels are in Supplementary Methods S20 and Data 6.

Predictive uncertainty adds a final check on using this endpoint for new molecules. TabPFN's nominal 80% and 95% predictive intervals cover only 58.2% and 80.0% of External-220, respectively, with similar undercoverage on Strict-97 (Supplementary Fig. 11 and Table 16). These native intervals have not been calibrated on the external labels; their widths should not yet be read as reliable per-molecule error guarantees.

Discussion

The value of SolvAI lies in changing when solvent information is used. Instead of resolving a new solvent environment for every query, it learns molecular coordinates from earlier calculations and measurements, then lets an experimental hydration endpoint decide how those coordinates matter. Predicted physical descriptors and multi-fidelity transfer have precedents;[8;9;18;19] the test here is whether several heterogeneous solvent sources can be combined and reused across new molecules and learners. The matched ExtraTrees contrast shows a gain beyond structural features, and shuffling the response block among molecules removes it. Chemical-separation and external-cohort tests show that part of this information remains useful beyond nearby endpoint labels.

Two kinds of portability matter here. A response representation is reusable if different hydration learners can extract value from it; a complete predictor transfers only if it remains accurate in new chemistry. The first property is supported by the paired gains across several endpoint families, including fits with pretrained molecular encodings. It does not follow from ranking complete models by error: the lowest-error learner changes across cohorts, and its own response ablation can be neutral or unfavorable. Keeping these questions apart shows what the frozen surrogates contribute without attributing every model improvement to them.

The source-composition reversal shows why adding more teachers is not a monotonic route to transfer. Empirical parameters and residual corrections help the full mixture on ARROW-85, while the fixed computed-source subset has lower error on both external cohorts (Fig. 6). Likewise, a more accurate surrogate for its own source target need not yield a better hydration feature. These comparisons do not identify a single mechanism for the reversal; the source distributions, prediction errors and endpoint use of each coordinate can all matter. They do show that source selection must be evaluated on the intended deployment chemistry, with independent labels held out from that choice.

The same distinction between promising behavior and verified accuracy governs larger molecules. The labelled size split tests a different endpoint regime, but earlier development had already seen its labels. Peptide and alkane curves show how predictions behave as structures grow; without reference free energies, they cannot validate that behavior. The failed PIMD2 descriptors and undercovered predictive intervals further show that a physically motivated target or a numerical uncertainty estimate is not automatically useful at deployment. The present evidence is for neutral molecules near room temperature. It does not establish performance for ions, other solvents, conformer-resolved energies or forces.

Independent larger-molecule and new-solvent references are the next decisive tests. They would allow source learners to be judged by downstream utility as well as recovery of their own targets, and would show how far solvent information learned once can replace a calculation repeated for every new molecule.

Methods

Reference set and targets

The primary evaluation uses the 85 neutral organic solutes reported in the ARROW hydration analysis at 298 K.[5] One water condition was retained per connectivity. The released benchmark table contains canonical structure, experimental hydration free energy, experimental uncertainty where available, classical ARROW, four-bead PIMD (PIMD4) and PIMD8 values, and all fold assignments. The exact reconstructed PIMD8 MAE is 0.205 kcal mol^{-1}; the source article reports approximately 0.20 kcal mol^{-1}.

Endpoint labels and features

The public training pool contains 1,280 neutral, benchmark-disjoint experimental hydration labels: 1,147 connectivity-resolved records from a FreeSolv/CombiSolv-Exp merge and 133 SoluteML dGsolvDB3 records supported by at least two source measurements.[2;9;23] The ExtraTrees input has a 2,048-bit radius-2 Morgan fingerprint and 217 RDKit two-dimensional descriptors. Non-finite values are handled by a median imputer fitted inside each endpoint fit.

Response surrogates

Fifteen response descriptors are generated by six surrogate families (Supplementary Table 1). CheMeleon-initialized D-MPNNs predict two source families: CombiSolv-QM/COSMOtherm and MolSolv SMD(water).[35;36] ExtraTrees models predict five Abraham axes and the physical and residual components of corrected OpenFF and GBn2 responses. LightGBM models predict six ConfSolv conformer and response-distribution summaries. The latter are derived from the H_2O subset of ConfSolv. The final coordinate list, target units, source sizes, transformations and inference calculations are given in Supplementary Table 1. PIMD-derived targets are absent from this list.

Retrospective source-surrogate architecture challenge

We separately compared five fixed learner candidates for each of four previously tree-based source families: the released tree recipe, an MLP using binary Morgan bits, an MLP using count features, count-aware ridge and count-aware ridge plus a bounded neural residual. The two D-MPNN source families, all source target definitions and the endpoint labels remained unchanged. Candidate choice used three source-development folds and a fixed 0.005 target-standardized-MAE tie margin favoring the simpler learner. A 20% source test, reserved before fitting, was opened only after choice was locked. The source-development choice of Abraham's tree was retained despite the lower test point estimate of an unselected hybrid. Full candidates, preprocessing, targetwise and summed-source diagnostics appear in Supplementary Method S22 and Table 17.

Endpoint comparisons used three source conditions: the released response matrix; a protocol-matched refit of the four tree families; and that refit with only the preselected OpenFF and GBn2 hybrids substituted. Every exact endpoint molecule appearing in a source table received source out-of-fold predictions; none of the ARROW-85 or Strict-97 identities appeared in any supervised source table. ExtraTrees and ridge-plus-neural-residual endpoints were refitted with their own locked recipes, identical labels, folds and seeds across source conditions. External-220 includes source-exposed molecules and is secondary to Strict-97 for this source-stage challenge. Its paired intervals use 10,000 molecule resamples and are descriptive, not adjusted for model

exploration. The 205-molecule size refit and unlabelled alkane/peptide series are diagnostics, not independent validation of extensivity (Supplementary Method S22 and Table 18).

Matched ExtraTrees endpoint

For the ExtraTrees endpoint, an ensemble of three `ExtraTreesRegressor` pipelines[30] is used, each with 360 trees, maximum feature fraction 0.7, minimum leaf size 2, no bootstrap and random seeds 11, 29 and 47. Predictions are averaged. Public endpoint rows have sample weight 1 and available ARROW outer-training rows weight 3. The matched structure-only model is identical except that the 15 response columns are removed.

Identity standardization

RDKit 2026.03.5 was used to generate canonical isomeric SMILES, full InChIKeys and connectivity blocks.[37] The confirmatory audit additionally applied MolStandardize cleanup, largest-fragment selection, uncharging and canonical tautomer enumeration. Exact and standardized benchmark equivalents were removed from every supervised external source. The audit found no such match in the endpoint pool. It identified 2 CombiSolv-QM, 32 MolSolv and 22 ConfSolv source records beyond the original exact-connectivity exclusion. The affected surrogates were refitted with all remaining molecules held in their original train, validation or test split. Full pair lists, identifiers and split-verification hashes are released.

Confirmatory evaluation

Confirmatory analyses followed a prespecified protocol archived in Git. Documented implementation amendments preceded the corrected runs (Supplementary Methods S9). The primary partition is the existing five-fold assignment on the 85 molecules. Five complete shuffled K-fold partitions use seeds 314159, 271828, 161803, 141421 and 173205. In each outer fold, all 1,280 public labels and the available ARROW outer-training labels are fitted; the test target is absent. Feature blocks, ExtraTrees hyperparameters, sample weights and model seeds are identical across all comparisons.

A separately prespecified one-factor sensitivity repeated the fixed folds after changing only the available ARROW outer-training weight from 3 to 1. It retained all source rows, surrogates, features, hyperparameters and seeds. Its paired interval used the same 100,000-resample convention.

For the negative control, the 15 response columns are permuted as one block, thereby preserving their joint distribution while destroying molecule alignment. Public, outer-training and outer-test response rows are permuted separately within each fold. Five permutation seeds (88001–88005) are evaluated for the fixed partition and each of the five repeats.

Chemical separation

Separation is applied to the complete endpoint-supervised pool. Functional families are assigned by a deterministic SMARTS precedence; five outer folds are formed by group. Bemis–Murcko scaffolds are canonicalized, with acyclic structures grouped by primary family. Taylor–Butina clusters use radius-2, 2,048-bit Morgan fingerprints at Tanimoto similarity 0.70. For each outer test fold, every public or ARROW training row sharing a held-out group is removed. A separate nearest-neighbor analysis removes every training row with similarity at or above 0.50, 0.60, 0.70 or 0.80 to any test molecule. No threshold was selected by performance. Exact SMARTS, assignments, training counts and maximum retained similarities are released in Supplementary Data 3 and the accompanying `results/confirmatory/` tables.

Zero-ARROW transfer

The frozen endpoint configuration was fitted once to the 1,280 public experimental labels without any ARROW experimental value. Both matched structure-only and full response-descriptor models then predicted all 85 ARROW molecules. This analysis measures transfer without reference-set endpoint adaptation; it is not described as an independent external test because model development used the reference chemistry.

External validation

The candidate cohort was frozen for a separate solvation-model validation study, before that study's model predictions, using original-measurement records from Sander's Henry-law compilation v5.0.0.[34] The intrinsic Henry solubility constant H_s^{cp} is the aqueous concentration divided by the gas partial pressure. Values at 298.15 K were converted using $K_c = H_s^{cp} RT$ and $\Delta G_{\mathrm{hyd}} = -RT \ln K_c$, corresponding to 1 M ideal gas and 1 M ideal-dilute aqueous standard states. Here K_c is the dimensionless aqueous-to-gas concentration ratio, R the gas constant and T the temperature. Qualification preceded SolvAI evaluation and required one connected neutral solute, supported elements, compatible target metadata and no exact, fragment-parent, uncharged-parent or canonical-tautomer equivalent in the 1,280 public or 85 ARROW endpoint labels. One previously documented registry conflict was excluded, leaving 220 molecules. A strict 97-molecule subset additionally excludes any identity-equivalent occurrence in all six supervised response-source tables.

For the matched ExtraTrees comparison, deployment-style fitting used the same 1,280 public plus all 85 ARROW labels, weights, preprocessing, architecture and seeds; only the 15 response columns differ. External-cohort labels were not used for fitting, surrogate or feature selection, thresholding or model choice in that comparison. The subsequent architecture study reuses this already examined cohort and is therefore retrospective, even though its final models do not fit these labels. The primary paired analysis and bootstrap convention match the confirmatory analysis. Complete eligibility decisions, source-exposure matches and molecule-level predictions are released as Supplementary Data 5. Because the compilation is external and molecular identities are disjoint, we call this an external molecule-disjoint cohort; we do not claim that all historical measurement provenance is independent.

Additional molecular-model comparisons

The graph and pretrained-representation comparisons followed a fixed protocol specified before their fits. These are retrospective extensions on the existing evaluation sets, not a second prospective validation. The same public labels, ARROW folds and sample weights are used for every model. Two direct predictors map molecular graphs to experimental hydration free energy: a D-MPNN trained from random initialization and a CheMeleon-initialized D-MPNN with its encoder fine-tuned.[35;36] For each outer training set and seed, 15% of each label source is held out internally to select the epoch count by validation MAE, with an 80-epoch cap and patience 12. The model is then initialized afresh from the same random or pretrained starting weights and trained on the complete outer training set for that count. Only this refitted model predicts the outer test set. Three seeds (11, 29, 47) are averaged. Supplementary Fig. 6 illustrates the selection and refit protocol; Supplementary Methods S14 gives the complete configurations.

Two further comparisons use frozen pretrained representations: MoLFormer encodes a SMILES sequence, and Uni-Mol encodes an RDKit-generated conformer.[29;38] Each is concatenated with the 2,265 structural features and fitted with the same ExtraTrees ensemble as the matched control. We repeat the fit after adding the 15 response descriptors. This paired design tests whether the responses add value to pretrained features without changing the downstream learner.

The pretrained encoders receive no hydration labels during these fits; their original pretraining membership has not been audited for molecular disjointness. All model families and both members of each paired comparison are reported, irrespective of their ranking.

These matched evaluations compare alternative molecular-learning approaches without conflating different data splits or training resources.[39] We do not compare the reported errors directly with unfiltered FreeSolv leaderboard values: 80 ARROW identities occur in FreeSolv, and the supervised response sources require their own exclusions. A conventional FreeSolv benchmark would require fold-specific refitting of every affected response source. The external 97-molecule subset instead tests all six response sources with explicitly audited exclusions.

The computation-only source control was specified separately before its ARROW fits. After that comparison, we extended the same ten-descriptor configuration to the external cohort without retuning. Its OpenFF and GBn2 components were recovered from the packaged surrogates and checked against the frozen corrected-response sums before evaluation (Supplementary Methods S17). This sequence distinguishes the retrospective source analysis from the original prospective external evaluation.

Broader endpoint search and numerical reproducibility

The extended study retains the five original ARROW outer folds, the external 220 identities, the strict subset and the 15 frozen response predictions. For each outer training pool, random inner splits hold out 15% of each source. The selection score is the equally weighted mean of public-source and ARROW-source validation MAE, not an outer-test score. Neural fits use source weights 1 and 3, respectively. Candidate screening is followed by repeated inner validation, longer-training challenges up to 1,200 epochs and a separate width challenge. Each family has a locked configuration and training schedule for each outer partition. Final fits restart from their specified initialization, restore the inner-validation rows to training, and average seeds 11, 29, 47, 71 and 101. Matched no-response fits use the same recipe and schedule. The ExtraTrees and frozen-encoder/tree controls use seeds 11, 29 and 47; common-three-seed comparisons are also reported. Supplementary Methods S18 specifies the search and its limits.

MoLFormer representations are generated with the encoder's deterministic evaluation option in addition to evaluation mode. An audit found that evaluation mode alone left stochastic attention projections active in the earlier cache. We regenerated all study and stress representations from the same checkpoint and repeated both the neural-head and fixed-tree comparisons. Only these corrected MoLFormer results enter the manuscript's comparisons. Reload, repeated-query and reordered-batch checks verify prediction replay; the packaged ExtraTrees endpoint does not use MoLFormer and is unaffected. The discrepancy and corrected hashes are recorded in Supplementary Methods S19 and Data 6.

TabPFN-3.5 uses the pinned full and Fast checkpoints with `tabpfn` 9.1.0, run locally under the provider's non-commercial license. We do not fine-tune the foundation model. Eight candidates combine two checkpoints with four feature recipes; training-side challenges compare eight versus 32 ensemble passes and single versus repeated ARROW context rows. Context repetition represents source emphasis but is not mathematically equivalent to weighted neural loss. Splits precede repetition. The pretrained model's native preprocessing, including feature selection, is retained; this is not an entirely tree-free preprocessing pipeline. Fitted contexts contain training labels and are not redistributed.

Size, series and predictive-uncertainty diagnostics

The size diagnostic holds out the largest 15% within each source, with stable identity ordering for ties. Every model uses the same 1,160/205 split and three seeds; no held-out label is used

in that refit. Configurations come from the locked external-fit recipe. Because the held-out labels were accessible in earlier random-fold development, this is an exploratory conditional test of size transfer. The source models and their source-level out-of-fold predictions are not rebuilt around this size split; exclusion from the endpoint refit does not imply exclusion from all source supervision (Supplementary Methods S5 and S20). The unlabelled grid comprises 30 linear alkanes and 13 members each of capped glycine and alanine series, including the common zero-residue molecule. Series identities, source overlap, fingerprint counts and all predictions are released.

For uncertainty evaluation, we transform each TabPFN distribution to the common raw-target scale and form an equally weighted mixture across five seeds. Central 80% and 95% intervals are quantiles of this mixture, not averages of seedwise quantiles. Empirical coverage is the fraction of external labels inside each interval. We also report widths and verify the mixture mean against stored point predictions. The external cohort is used only to evaluate these intervals, not to calibrate them. Supplementary Methods S20–S21 details these diagnostics.

Statistics

MAE in kcal mol^{-1} is primary. Root mean square error (RMSE), median absolute error and the fraction of molecules improved are secondary. All comparisons are paired by molecule. Ninety-five percent percentile intervals use 100,000 molecule bootstrap resamples with seed 20260828. For repeated partitions, each molecule's absolute error is first averaged across repeats before resampling. A predeclared comparison is positive when the complete paired interval for candidate-minus-control MAE is below zero, negative when above zero and neutral otherwise. We report all prespecified comparisons without multiplicity adjustment and do not treat the panel as a single confirmatory hypothesis test. No post-result feature selection is used within the confirmatory comparisons. The later architecture search uses training-side validation and a locked selection policy; its test intervals are descriptive, conditional on fitted models and not adjusted for the many architectures compared.

Released ExtraTrees deployment artifact

After evaluation was frozen, the response surrogates were fitted with standardized benchmark equivalents removed and the endpoint was refitted on the 1,280 public plus 85 ARROW labels. This all-data artifact is used for deployment and the endpoint-disjoint external evaluation; ARROW accuracy uses only the appropriate fold-specific fits. Inference parses the supplied SMILES with RDKit and converts it to canonical isomeric SMILES before computing descriptors and evaluating the six frozen response-surrogate artifacts; three endpoint regressors are averaged. This step is not a claim of invariance to tautomer, protonation or salt form. The runtime does not read an experimental value, family label, scaffold, fold identifier, trajectory or PIMD result. Software versions, hashes and a SMILES-only smoke test are reported in Supplementary Methods S12–S13 and Tables 7–8. On the release host, warm CPU inference took a median of 15.29 s for one molecule and 15.82 s for a batch of 32 (0.494 s per molecule). These measurements describe this artifact and hardware; they are not a same-hardware comparison with physical simulation or the other endpoint models.

Acknowledgments

This work was supported in part by US National Institutes of Health award R35GM122543 (M.L.).

Author contributions

G.O. and M.L. conceived and designed the study. G.O. carried out the computational work. G.O., B.F. and M.L. interpreted the results and revised the manuscript.

Competing interests

The authors declare no competing interests.

Use of artificial intelligence

Generative AI tools assisted with this research and manuscript preparation, and the authors take full responsibility for the content.

Data availability

Benchmark records, held-out predictions, split assignments, audit tables and derived source manifests for the released model and endpoint-architecture study are provided in the SolvAI repository (Supplementary Data 1–6). Aggregate results of the retrospective source-surrogate challenge are included in Supplementary Tables 17–18 and the repository's source-architecture evidence JSON. Complete fit-level and molecule-level records for that challenge are retained for the revised data release. Large third-party datasets are not redistributed where their size or license makes that inappropriate; URLs or DOIs, hashes, licenses and preprocessing instructions are supplied in `repro/DATA_PROVENANCE.md`.

Code availability

Source code for the released model, its packaged inference artifacts, confirmatory and endpoint-architecture analyses, figure generation, manuscript source and a locked environment are available at https://github.com/Scientific-Computing-Lab/SolvAI. A frozen-artifact path reproduces predictions, metrics, figures and PDFs without rerunning physical calculations for the released analyses. The endpoint-comparison directory additionally provides predictions, selection specifications and analysis code. TabPFN weights require the provider's license; fitted contexts that contain non-redistributable training labels are not published. The packaged public inference command uses the ExtraTrees endpoint, and its runtime measurements do not describe the alternative endpoints. The source-surrogate challenge code and its full fit records are not part of the current public release and will accompany the revised data deposit.

References

[1] 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. https://doi.org/10.1007/s10822-014-9747-x.

[2] Guilherme Duarte Ramos Matos, Daisy Y. Kyu, Hannes H. Loeffler, John D. Chodera, Michael R. Shirts, and David L. Mobley. Approaches for calculating solvation free energies and enthalpies demonstrated with an update of the FreeSolv database. *Journal of Chemical & Engineering Data*, 62:1559–1569, 2017. https://doi.org/10.1021/acs.jced.7b00104.

[3] Michael R. Shirts and John D. Chodera. Statistically optimal analysis of samples from multiple equilibrium states. *The Journal of Chemical Physics*, 129:124105, 2008. https://doi.org/10.1063/1.2978177.

[4] Thomas E. Markland and Michele Ceriotti. Nuclear quantum effects enter the mainstream. *Nature Reviews Chemistry*, 2:0109, 2018. https://doi.org/10.1038/s41570-017-0109.

[5] Leonid Pereyaslavets, Ganesh Kamath, Oleg Butin, et al. Accurate determination of solvation free energies of neutral organic compounds from first principles. *Nature Communications*, 13:414, 2022. https://doi.org/10.1038/s41467-022-28041-0.

[6] Alexey Illarionov, Serzhan Sakipov, Leonid Pereyaslavets, et al. Combining force fields and neural networks for an accurate representation of chemically diverse molecular interactions. *Journal of the American Chemical Society*, 145:23620–23629, 2023. https://doi.org/10.1021/jacs.3c07628.

[7] Igor V. Kurnikov, Leonid Pereyaslavets, Ganesh Kamath, et al. Neural network corrections to intermolecular interaction terms of a molecular force field capture nuclear quantum effects in calculations of liquid thermodynamic properties. *Journal of Chemical Theory and Computation*, 20:1347–1357, 2024. https://doi.org/10.1021/acs.jctc.3c00921.

[8] Raghunathan Ramakrishnan, Pavlo O. Dral, Matthias Rupp, and O. Anatole von Lilienfeld. Big data meets quantum chemistry approximations: The delta-machine learning approach. *Journal of Chemical Theory and Computation*, 11:2087–2096, 2015. https://doi.org/10.1021/acs.jctc.5b00099.

[9] Florence H. Vermeire and William H. Green. Transfer learning for solvation free energies: From quantum chemistry to experiments. *Chemical Engineering Journal*, 418:129307, 2021. https://doi.org/10.1016/j.cej.2021.129307.

[10] Clemens Rauer and Tristan Bereau. Hydration free energies from kernel-based machine learning: Compound-database bias. *The Journal of Chemical Physics*, 153:014101, 2020. https://doi.org/10.1063/5.0012230.

[11] Logan Ward, Naveen Dandu, Ben Blaiszik, et al. Graph-based approaches for predicting solvation energy in multiple solvents: Open datasets and machine learning models. *The Journal of Physical Chemistry A*, 125:5990–5998, 2021. https://doi.org/10.1021/acs.jpca.1c01960.

[12] Zhan-Yun Zhang, Ding Peng, Lihong Liu, Lin Shen, and Wei-Hai Fang. Machine learning prediction of hydration free energy with physically inspired descriptors. *The Journal of Physical Chemistry Letters*, 14:1877–1884, 2023. https://doi.org/10.1021/acs.jpclett.2c03858.

[13] Jan Weinreich, Nicholas J. Browning, and O. Anatole von Lilienfeld. Machine learning of free energies in chemical compound space using ensemble representations: Reaching experimental uncertainty for solvation. *The Journal of Chemical Physics*, 154:134113, 2021. https://doi.org/10.1063/5.0041548.

[14] Jenke Scheen, Wilson Wu, Antonia S. J. S. Mey, Paolo Tosco, Mark Mackey, and Julien Michel. Hybrid alchemical free energy/machine-learning methodology for the computation of hydration free energies. *Journal of Chemical Information and Modeling*, 60:5331–5339, 2020. https://doi.org/10.1021/acs.jcim.0c00600.

[15] Amin Alibakhshi and Bernd Hartke. Improved prediction of solvation free energies by machine-learning polarizable continuum solvation model. *Nature Communications*, 12:3584, 2021. https://doi.org/10.1038/s41467-021-23724-6.

[16] Vigneshwari Subramanian, Ekaterina L. Ratkova, David Scott Palmer, Ola Engkvist, Maxim V. Fedorov, and Antonio Llinàs. Multisolvent models for solvation free energy predictions using 3D-RISM hydration thermodynamic descriptors. *Journal of Chemical Information and Modeling*, 60:2977–2988, 2020. https://doi.org/10.1021/acs.jcim.0c00065.

[17] Amin Alibakhshi and Bernd Hartke. Implicitly perturbed Hamiltonian as a class of versatile and general-purpose molecular representations for machine learning. *Nature Communications*, 13:1245, 2022. https://doi.org/10.1038/s41467-022-28912-6.

[18] Thijs Stuyver and Connor W. Coley. Quantum chemistry-augmented neural networks for reactivity prediction: Performance, generalizability, and explainability. *The Journal of Chemical Physics*, 156:084104, 2022. https://doi.org/10.1063/5.0079574.

[19] Luyang Jia, Zhan-Yun Zhang, Bin-Bin Xie, Lin Shen, and Wei-Hai Fang. A two-stage machine learning framework with physically inspired descriptors for efficient prediction of hydration free energies. *Chemistry–Methods*, 6:e202500150, 2026. https://doi.org/10.1002/cmtd.202500150.

[20] Shih-Cheng Li, Haoyang Wu, Angiras Menon, Kevin A. Spiekermann, Yi-Pei Li, and William H. Green. When do quantum mechanical descriptors help graph neural networks to predict chemical properties? *Journal of the American Chemical Society*, 146:23103–23120, 2024. https://doi.org/10.1021/jacs.4c04670.

[21] Guanming Chen and Thijs Stuyver. Harnessing surrogate models for data-efficient predictive chemistry: descriptors vs. learned hidden representations. *Digital Discovery*, 4:3227–3237, 2025. https://doi.org/10.1039/D5DD00256G.

[22] Andreas Klamt. Conductor-like screening model for real solvents: a new approach to the quantitative calculation of solvation phenomena. *The Journal of Physical Chemistry*, 99:2224–2235, 1995. https://doi.org/10.1021/j100007a062.

[23] Yunsie Chung, Florence H. Vermeire, Haoyang Wu, Pierre J. Walker, Michael H. Abraham, and William H. Green. Group contribution and machine learning approaches to predict Abraham solute parameters, solvation free energy, and solvation enthalpy. *Journal of Chemical Information and Modeling*, 62:433–446, 2022. https://doi.org/10.1021/acs.jcim.1c01103.

[24] Simon Boothroyd, Pavan Kumar Behara, Owen C. Madin, et al. Development and benchmarking of Open Force Field 2.0.0: The Sage small molecule force field. *Journal of Chemical Theory and Computation*, 19:3251–3275, 2023. https://doi.org/10.1021/acs.jctc.3c00039.

[25] Hai Nguyen, Daniel R. Roe, and Carlos Simmerling. Improved generalized Born solvent model parameters for protein simulations. *Journal of Chemical Theory and Computation*, 9:2020–2034, 2013. https://doi.org/10.1021/ct3010485.

[26] Aleksandr V. Marenich, Christopher J. Cramer, and Donald G. Truhlar. Universal solvation model based on solute electron density and on a continuum model of the solvent defined by the bulk dielectric constant and atomic surface tensions. *The Journal of Physical Chemistry B*, 113: 6378–6396, 2009. https://doi.org/10.1021/jp810292n.

[27] Lagnajit Pattanaik, Angiras Menon, Volker Settels, et al. ConfSolv: Prediction of solute conformer-free energies across a range of solvents. *The Journal of Physical Chemistry B*, 127:10151–10170, 2023. https://doi.org/10.1021/acs.jpcb.3c05904.

[28] Xiaolin Pan. The solvation energy dataset for machine learning model—MolSolv, 2022. https://doi.org/10.5281/zenodo.7262826.

[29] Gengmo Zhou, Zhifeng Gao, Qiankun Ding, et al. Uni-Mol: A universal 3D molecular representation learning framework. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=6K2RM6wVqKu.

[30] Pierre Geurts, Damien Ernst, and Louis Wehenkel. Extremely randomized trees. *Machine Learning*, 63:3–42, 2006. https://doi.org/10.1007/s10994-006-6226-1.

[31] Yury Gorishniy, Akim Kotelnikov, and Artem Babenko. TabM: Advancing tabular deep learning with parameter-efficient ensembling. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2410.24210. Preprint: https://doi.org/10.48550/arXiv.2410.24210.

[32] Noah Hollmann, Samuel Müller, Lennart Purucker, Arjun Krishnakumar, Max Körfer, Shi Bin Hoo, Robin Tibor Schirrmeister, and Frank Hutter. Accurate predictions on small data with a tabular foundation model. *Nature*, 637:319–326, 2025. https://doi.org/10.1038/s41586-024-08328-6.

[33] Prior Labs. TabPFN-3.5 model card and checkpoints, 2026. URL https://huggingface.co/Prior-Labs/tabpfn_3_5/tree/06bf2ba35c80a92a3b9abb436b99cf49e7a0365e. Revision 06bf2ba35c80a92a3b9abb436b99cf49e7a0365e; accessed 6 October 2026.

[34] Rolf Sander. Compilation of Henry's law constants (version 5.0.0) for water as solvent. *Atmospheric Chemistry and Physics*, 23:10901–12440, 2023. https://doi.org/10.5194/acp-23-10901-2023.

[35] Kevin Yang, Kyle Swanson, Wengong Jin, et al. Analyzing learned molecular representations for property prediction. *Journal of Chemical Information and Modeling*, 59:3370–3388, 2019. https://doi.org/10.1021/acs.jcim.9b00237.

[36] Jackson W. Burns, Akshat Shirish Zalte, Charlles R. A. Abreu, et al. Deep learning foundation models from classical molecular descriptors. arXiv:2506.15792, 2025. https://doi.org/10.48550/arXiv.2506.15792.

[37] RDKit contributors. RDKit: Open-source cheminformatics, 2026. URL https://www.rdkit.org/.

[38] Jerret Ross, Brian Belgodere, Vijil Chenthamarakshan, Inkit Padhi, Youssef Mroueh, and Payel Das. Large-scale chemical language representations capture molecular structure and properties. *Nature Machine Intelligence*, 4:1256–1264, 2022. https://doi.org/10.1038/s42256-022-00580-7.

[39] Zhenqin Wu, Bharath Ramsundar, Evan N. Feinberg, et al. MoleculeNet: a benchmark for molecular machine learning. *Chemical Science*, 9:513–530, 2018. https://doi.org/10.1039/C7SC02664A.

[40] Jiahui Yu, Chengwei Zhang, Yingying Cheng, et al. SolvBERT for solvation free energy and solubility prediction: a demonstration of an NLP model for predicting the properties of molecular complexes. *Digital Discovery*, 2:409–421, 2023. https://doi.org/10.1039/D2DD00107A.

[41] Jennifer A. Clark. OpenFE ASFE – solvation_set – freesolv – 2026-08-06-openff-2.3.0-solvation_set_freesolv. Zenodo, 2026. https://doi.org/10.5281/zenodo.21810272.

[42] Paul Katzberger and Sereina Riniker. A general graph neural network based implicit solvation model for organic molecules in water. *Chemical Science*, 15:10794–10802, 2024. https://doi.org/10.1039/D4SC02432J.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SolvAISupplementaryInformation.pdf
- SolvAISupplementaryData6EndpointArchitectures.zip