E2FORMER-V3: ATOM-ROLE-AWARE EQUIVARIANT INTERATOMIC POTENTIALS

Anonymous authors
Paper under double-blind review

ABSTRACT

Machine-learned interatomic potentials enable high-level accuracy for large-scale biological simulation, yet remain costly. In aqueous systems, hydrogens and the water solvent dominate the atom count, while uniform equivariant architectures give every atom the same representation capacity and spend most of their computation on them. We introduce E2Former-V3, an atom-role-aware equivariant potential architecture that represents every atom explicitly and keeps forces conservative. A static atom role, assigned before the backbone runs, sets each atom's latent width and how often it is updated. Two configurations follow. E2Former-V3-Standard routes hydrogens through half as many full-width block updates, reaching a force MAE of 8.146 meV/Å, slightly better than E2Former-V2 (Huang et al., 2026), at 1.75–$1.84\times$ its throughput. E2Former-V3-Flash routes hydrogens and water oxygens through a narrow stream with fewer updates, trading 3.126 meV/Å higher force error for 4.00–$5.85\times$ the throughput. We run molecular dynamics over water, a salt solution and a solvated protein for 20 ps, across three ensembles and three seeds. The runs hold their target temperature, keep chemical structure intact, and reproduce the measured liquid structure to within a few hundredths of an Ångstrom. On the solvated protein, both production checkpoints complete the full 1 ns without water fusion, nonbonded clashes or solute collapse.

1 INTRODUCTION

Machine-learning interatomic potentials (MLIPs) now approach density-functional-theory accuracy on molecular and biomolecular systems, returning an energy whose gradient is a conservative force on every atom (Schütt et al., 2017; Batzner et al., 2022; Batatia et al., 2022). Yet most equivariant MLIPs make the same choice without comment: every atom gets the same latent width and the same number of expensive block updates. That uniformity was inherited from earlier architectures, and nothing in the chemistry demands it.

Explicit-solvent systems are where that default costs the most, and where longer trajectories and larger boxes are most wanted. Figure 1(a) shows the composition of one such system, a solvated nucleic acid. Of its 15,961 atoms, 10,334 are hydrogen and 15,057 belong to water, its hydrogens counted in both; the solute contributes 608 heavy atoms. Per-atom work is uniform, so the cost follows those counts: the equivariant updates, the stored activations and the autograd graph behind the conservative forces all scale with atom number.

Computational chemistry has long handled that cost by treating atoms unequally: implicit solvent, coarse graining and constrained dynamics all show that hydrogens and solvent need not be resolved at full fidelity (Marrink et al., 2007; Kmiecik et al., 2016; Tan et al., 2006; Katzberger & Riniker, 2023; Ryckaert et al., 1977; Andersen, 1983; Miyamoto & Kollman, 1992; Hess et al., 1997; Hopkins et al., 2015). Each saves the cost by changing the physical model: a continuum replaces the solvent, pseudoparticles merge atoms, or bonds are frozen. The all-atom description goes with the cost, so none of these methods returns a first-principles force on every atom. Existing equivariant potentials, efficiency methods and conditional-computation models (Section 2) reduce the cost as well, and none of them allocates capacity by atom role while keeping every atom and its first-principles forces.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

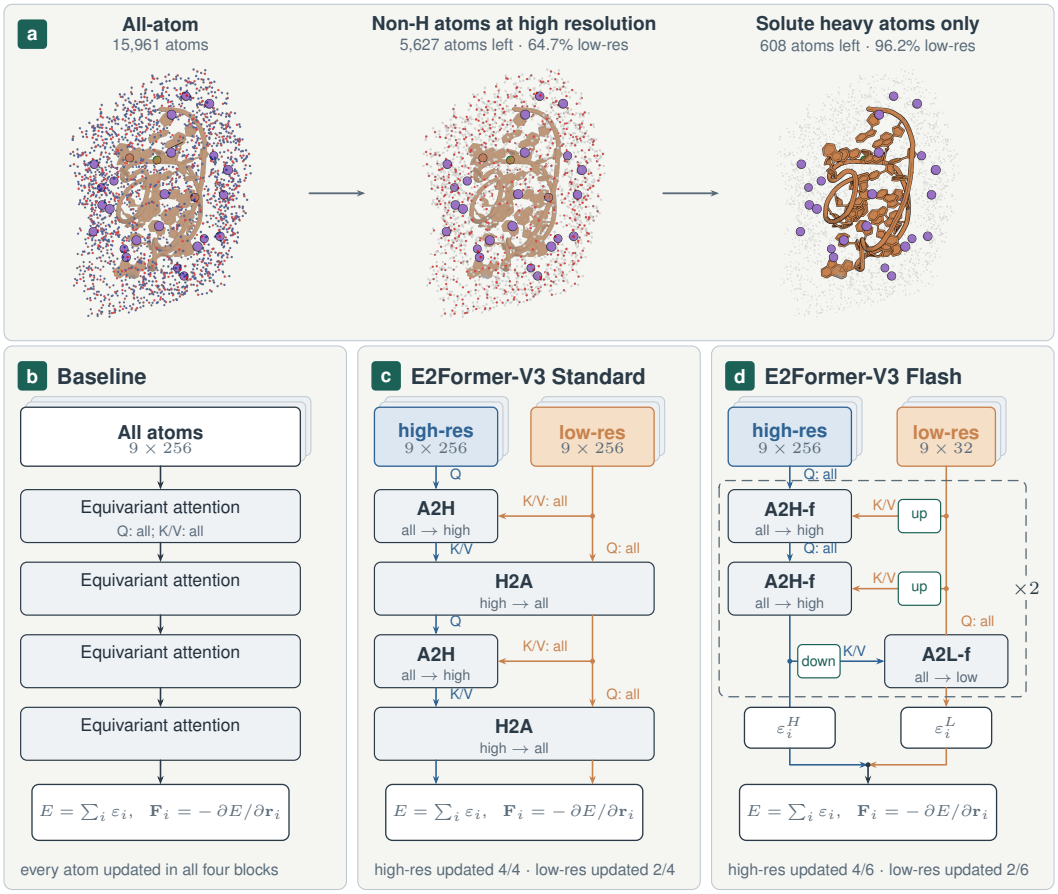


Figure 1: **E2Former-V3 under three capacity allocations.** **(a)** One solvated nucleic acid (15,961 atoms; water drawn within 8 Å of the RNA, counts are the full system's), at the three levels of atomic detail the allocations use: every atom; then hydrogen at low resolution, so 64.7% of atoms are low-resolution and 5,627 stay high-resolution; then the oxygens of water at low resolution as well, so 96.2% are low-resolution and only the 608 solute heavy atoms stay high-resolution. The panels below use the same three columns. **(b)** Baseline: one shared 9×256 state, every row updated in all four blocks $(4/4)$. The 9 is the $1+3+5$ angular components of $l = 0, 1, 2$ at $l_{\max} = 2$. **(c)** Standard: $[\mathrm{A2H}, \mathrm{H2A}] \times 2$ on one shared 256-channel state, so the role streams have equal width and differ only in update count, $4/4$ against $2/4$. A2H queries an all-atom bank from the high-resolution rows and updates those rows only; H2A is its transpose, so every atom queries a high-resolution-only bank and both lanes are written. **(d)** Flash: $[\mathrm{A2H\text{-}f}, \mathrm{A2H\text{-}f}, \mathrm{A2L\text{-}f}] \times 2$ on two persistent streams, 256 channels per degree at high resolution against 32 at low resolution (drawn narrow, since width is the second capacity axis); the dashed frame marks the three blocks of one repetition, run twice. A2H-f updates the high-resolution stream from a bank built from both, A2L-f the low-resolution stream from a bank built from both; up and down are the degree-wise projections carrying that exchange. Updates are therefore $4/6$ and $2/6$. In both routed allocations the role heads merge before forces are differentiated, so no atom is removed and no coordinate is integrated out.

This work. We introduce an atom-role-aware machine-learning interatomic potential architecture. As in the reduced-fidelity methods above, hydrogens and solvent carry less of the model, but nothing about the physics is dropped: every atom and every degree of freedom stay, the forces still come from one conservative energy, and only the capacity an atom consumes is smaller. Within that constraint we route atoms statically by atom role: the atoms that need full detail keep the high-resolution tensors and the full block schedule, and the rest are modelled at low resolution. We propose two configurations, drawn in Figure 1(c) and (d).

Standard sends hydrogen to low resolution, in the solvent and on the solute alike, and holds the width at 256 channels per degree for every atom. Low-resolution atoms take two of the four block updates, high-resolution atoms all four, so the saving comes from participation alone. **Flash** sends hydrogen and the oxygens of water to low resolution, and narrows that stream as well, to 32 channels per degree beside the persistent 256-channel one. Its schedule runs six blocks, of which low-resolution atoms take two and high-resolution atoms four. In both configurations the two role streams exchange information in both directions, and their energy heads merge before forces are differentiated. In an A2H block only the high-resolution rows query, so most of the key/value side never has to be built at all; skipping it cuts latency by 23.1–27.3% across the five measured systems. The implementation runs on custom Triton kernels (Tillet et al., 2019), and Section 3 gives the architecture and both optimizations in full.

Results. The two routed configurations give two non-dominated accuracy–speed operating points against E2Former-V2, the uniform-capacity predecessor, and they trade along different axes. **Standard** favors force accuracy: a force MAE of 8.146 meV/Å on the OMol25 validation split (a molecular dataset built for density-functional reference data), against E2Former-V2's published 8.38, at 1.75–$1.84\times$ the throughput on a set of solvated benchmark systems and less peak memory on each (Table 6). **Flash** favors speed, memory and fixed-device reach at a higher but still chemically accurate force error: 11.272 meV/Å, inside the 1 kcal mol^{-1} Å^{-1} chemical-accuracy threshold (Chmiela et al., 2017), for 4.00–$5.85\times$ the throughput. Its reach is absolute as well as relative: on one A100-40GB, where the full-capacity baselines stop at 12000 and 18000 atoms, Flash evaluates a 74001-atom water box.

We ran molecular dynamics on several systems to check that the model stays physically reliable. Both configurations were run over water, a salt solution and a solvated protein for 20 ps across all three ensembles and three seeds. All 54 trajectories completed. The thermostatted legs held the target temperature, and none of the trajectories produced a non-finite value, atomic overlap or water fusion. The two production checkpoints then completed 1 ns on the solvated protein in NVT without water fusion, nonbonded clash or solute collapse (Table 20). Oxygen–oxygen and ion–oxygen radial distributions sit within a few hundredths of an Ångstrom of their measured positions, and the full curves are in the appendix (Figure 14). Within these measured horizons, neither configuration showed the listed dynamical failures. Both retain the cross-role interactions and the energy-conserving force construction that the all-atom description depends on. E2Former-V3 therefore keeps the first-principles, all-atom description and extends the accuracy–efficiency Pareto frontier in solvated biomolecular systems, holding chemical accuracy and stable dynamics (Figure 2).

2 Related Work

Equivariant Machine-Learning Interatomic Potentials. Invariant message passing (Schütt et al., 2017; Gasteiger et al., 2020; 2021), tensor operations on irreducible representations (Thomas et al., 2018; Geiger & Smidt, 2022), data-efficient, higher body-order, strictly local and attention-based potentials (Batzner et al., 2022; Batatia et al., 2022; Musaelian et al., 2023; Batatia et al., 2025; Fuchs et al., 2020; Liao & Smidt, 2022; Liao et al., 2023), and smoothness-, conservation- and universality-oriented models (Fu et al., 2025; Bigi et al., 2024; Wood et al., 2025; Levine et al., 2025) all fix latent width, angular degree and block-update count globally, identically for a solvent hydrogen and a solute non-H atom. This global choice motivates the atom-role-aware allocation studied here.

Efficiency for Equivariant Potentials at Uniform Capacity. Algebraic restructuring cheapens the tensor product (Passaro & Zitnick, 2023; Luo et al., 2024; Li et al., 2025); fused attention recomputes edge intermediates rather than materialising them (Huang et al., 2026; Dao et al., 2022) via tile-level GPU languages and graph compilers (Tillet et al., 2019; Ansel et al., 2024); one line drops the symmetry constraints (Qu & Krishnapriyan, 2024); and graph-level work reduces *which edges exist* (Kahle et al., 2024; Hu et al., 2025). Every one of these changes a uniform operation's constant factor, or the edge set, and none changes how wide or how often a node is updated. All of them therefore stay bounded by the activation and autograd-graph footprint our setting shrinks. Ours changes how wide and how often a node is updated, which is orthogonal to all of the above:

both configurations keep the uniform baseline's attention implementation, so the gaps we report are the allocation rather than the kernel.

Reduced-Fidelity Treatments of Hydrogen and Solvent. These differ mainly in what they delete: coarse-grained fields merge atoms into pseudo-particles (Marrink et al., 2007; Kmiecik et al., 2016); constrained dynamics and mass repartitioning freeze or slow the fastest motions (Ryckaert et al., 1977; Andersen, 1983; Miyamoto & Kollman, 1992; Hess et al., 1997; Hopkins et al., 2015); implicit solvation substitutes a continuum that disagrees with explicit water exactly at the water-bridged contacts machine-learned solvents are built to recover (Tan et al., 2006; Katzberger & Riniker, 2023); and multi-resolution partitioning keeps two descriptions plus a boundary (Warshel & Levitt, 1976; Senn & Thiel, 2009; Böselt et al., 2021). They share the premise that resolution should follow atom role, which our static router adopts, and they share the mechanism of deleting degrees of freedom or splitting the Hamiltonian, which it does not. That premise is why our chemistry-versus-random-routing statement is a conjecture rather than a guess. What none of them supplies is one end-to-end differentiable all-atom energy whose gradient is a conservative force on every original atom, because the savings come from deleting degrees of freedom or splitting the Hamiltonian.

Conditional Computation and Non-Uniform Per-Token Compute. Non-uniform per-input compute is well validated — sparsely gated experts (Shazeer et al., 2017; Fedus et al., 2021; Zhou et al., 2022), learned or capped depth (Graves, 2016; Elbayad et al., 2019; Raposo et al., 2024), token pruning and merging (Rao et al., 2021; Bolya et al., 2022). All of them use attention with distinct query and key/value sets (Vaswani et al., 2017; Jaegle et al., 2021), which is the structure our cross-role blocks take. These mechanisms are conceptual precursors rather than direct MLIP competitors: molecular use additionally requires rotationally equivariant exchange across the resolution boundary, a conservative scalar with a dense per-node gradient, and retention of every atom; pruning and merging instead discard or fuse tokens. Three schemes inside interatomic potentials already vary compute per atom. Mixed precision gives each atom a precise or a fast potential, chosen from a local-structure descriptor (Immel et al., 2025). Element-indexed routing sends each atom to expert sub-networks by atomic number (Liu et al., 2026). Multifidelity domain decomposition splits the system in space and runs the higher-capacity model on the reactive region (de Miranda Nascimento et al., 2026). Mixed precision and domain decomposition distribute two separate potentials, so both need a switching rule or an agreement loss where the two meet; our role streams live inside one network and exchange through learned projections, so there is no seam and the forces are the gradient of one conservative energy. Element-indexed routing has been demonstrated in invariant architectures, where a nonlinear expert can act on a feature without breaking symmetry; the same experts inside an equivariant block would have to preserve each angular degree. A router fixed by atom role learns no gate, so ours runs inside the equivariant stack without that constraint. That router also separates species, and our Flash router separates two environments of the same species: solute oxygens hold high resolution while the oxygens of intact water drop to low. None of the mechanisms above decides capacity from an atom role.

3 METHOD

3.1 THE UNIFORM MLIP

An input is a set of atoms: atomic numbers Z_i, coordinates $\mathbf{r}_i$, charge, spin multiplicity, a validity mask, and optionally a periodic cell. A radial cutoff on the coordinates turns that into a neighbour graph. The model returns one total energy and one force on every atom. Force accuracy is always a *per-component* MAE, the mean absolute deviation over the three Cartesian components.

Between input and output the chain is three steps. *Embedding* initialises a per-atom intermediate representation from neighbour distances, neighbour directions and element identity. *Attention* then refines that representation by attending over the neighbours the graph supplies. *The head* reads energy and forces out of the refined representation.

The representation is a set of $\mathrm{SO}(3)$ irreducible representations under the $\mathrm{SO}(2)$ reduction of Passaro & Zitnick (2023), one block per angular degree up to a maximum degree $l_{\max} = 2$. Each atom therefore carries an intermediate representation of shape $[(l_{\max}+1)^2, C] = [9, C]$, where C is the

number of channels. Two quantities set an atom's capacity: C, its *width*, and the number of attention updates it receives, its *update count*. These are the two axes this paper allocates.

The readout fixes what an allocation may and may not do. Every atom contributes one scalar atomic energy and the model returns their sum, $E = \sum_i \varepsilon_i$, with $\mathbf{F}_i = -\partial E/\partial \mathbf{r}_i$, so every force is read out of that one scalar by differentiation.

A uniform MLIP gives every atom the same capacity — the same width and the same number of attention updates, whatever the atom is. In a solvated biological system, hydrogens and the oxygens of intact water are most of the atoms, so a uniform allocation spends most of its compute on them. The question studied here is whether assigning those atoms less capacity produces a more efficient MLIP without sacrificing accuracy.

3.2 The Atom-Role-Aware MLIP

We manually specify two chemical-prior routers. A wrapper routes each atom before the backbone runs. The first router $\mathrm{is_lr}^{\mathrm{Standard}}(i) = m_{\mathrm{H}}(i)$ sends every hydrogen to the low-resolution role, in solvent and solute alike; the second router $\mathrm{is_lr}^{\mathrm{Flash}}(i) = m_{\mathrm{H}}(i) \vee m_{\mathrm{W}}(i)$ sends those hydrogens and the oxygens of intact water (identified by two hydrogens within the bonding distance) to that role. The complementary atoms are high-resolution. Both masks are supplied consistently at training and inference and are bound to the checkpoint, so no external routing input can disagree with the weights. They depend only on element identity and a local predicate, so they are static, interpretable and exactly reproducible chemical priors (Marrink et al., 2007; Kmiecik et al., 2016; Ryckaert et al., 1977; Hess et al., 1997; Hopkins et al., 2015).

Both configurations build on the E2Attention operator of E2Former-V2 (Huang et al., 2026), in which target atoms supply the queries, source atoms the keys and values, and cached edge features modulate each target update, with distinct query and key/value sets (Vaswani et al., 2017; Jaegle et al., 2021) in equivariant form (Liao & Smidt, 2022; Liao et al., 2023) (Appendix A). The two allocations differ in which rows a block reads and writes.

E2Former-V3-Standard. Every atom remains in one shared state $x \in \mathbb{R}^{N\times 9\times 256}$, partitioned into high- and low-resolution rows only for deciding which rows a block reads and writes. *All-to-High-resolution* (**A2H**) takes the high-resolution rows as queries against a key/value bank built from *both* role partitions and updates only the high-resolution rows. *High-resolution-to-All* (**H2A**) is its transpose: both role partitions supply queries against a high-resolution-only key/value bank, so every row is updated in the same 256-channel irreps space. Repeating $[\mathrm{A2H}, \mathrm{H2A}]$ twice gives four blocks: high-resolution atoms receive $4/4$ updates and low-resolution atoms $2/4$. Standard therefore allocates block participation only; it does not narrow any atom's persistent state.

E2Former-V3-Flash. Flash maintains separate persistent states $h \in \mathbb{R}^{N_{\mathrm{HR}}\times 9\times 256}$ and $\ell \in \mathbb{R}^{N_{\mathrm{LR}}\times 9\times 32}$. *A2H-flash* takes h as its queries, builds the key/value bank $[\,h; \mathrm{up}(\ell)\,]$, and updates h only. *A2L-flash* takes ℓ as its queries, builds $[\,\mathrm{down}(h); \ell\,]$, and updates ℓ only, where $[\,a; b\,]$ concatenates along the atom axis. The learned degree-wise projections $\mathrm{up} : \mathbb{R}^{9\times 32} \to \mathbb{R}^{9\times 256}$ and $\mathrm{down} : \mathbb{R}^{9\times 256} \to \mathbb{R}^{9\times 32}$ change channel width while commuting with rotations (Geiger & Smidt, 2022; Passaro & Zitnick, 2023). Repeating the three-block sequence twice gives six blocks: high-resolution atoms receive $4/6$ updates and low-resolution atoms $2/6$. Flash therefore allocates both persistent width and block participation, with bidirectional cross-role exchange in every repetition.

Both configurations apply role-specific atomic-energy heads to their routed rows: Standard reads the two role partitions of its shared state, while Flash reads its two persistent states. The selected atomic energies are merged by the router mask *before* any gradient is taken,

$$E \;=\; \sum_{i:\,\neg \mathrm{is_lr}(i)} \varepsilon_i^{H} \;+\; \sum_{i:\,\mathrm{is_lr}(i)} \varepsilon_i^{L}, \qquad \mathbf{F}_i \;=\; -\,\frac{\partial E}{\partial \mathbf{r}_i}, \tag{1}$$

where $\mathrm{is_lr}$ denotes the router bound to the configuration, and which is the uniform readout when nothing is low-resolution. Four invariants follow, separating this from the reduced-fidelity schemes of Section 2: no atom removed, no coordinate degree of freedom removed, cross-role interaction preserved both ways, and an all-atom energy plus per-atom conservative force as the output signa-

ture. Figure 1 sets both routed allocations beside the uniform one: the block stream of each, the two capacity axes, and the low-resolution-atom update count.

3.3 Neighbor-Union Filtering

In an A2H block only the rows a high-resolution atom can reach take part in the update: the queries are the high-resolution rows and the key/value bank need only cover their neighbour tables. In a solvated system that set is far smaller than the whole atom set, which is where most of the saving comes from. The All-to-High-resolution block's key/value side is also read only through the neighbour index, so it need not be built over every atom at all. Neighbor-union filtering is the inference-time rewrite that exploits both facts. It is most effective when the query atoms are few and spatially concentrated, as in A2H and A2H-flash under Flash; across the five measured systems it reduces latency by 23.1–27.3% (the Ablation Study paragraph of Section 4, Figure 6).

The optimization rewrites that bank: instead of projecting every node and its periodic images, the enabled path projects only the *union of the high-resolution-neighbour tables* some A2H query can read. On the solvated peptide of Section 4, the bank it projects is 322 rows against 10,546. The rewrite is algebraically lossless, because that side is consumed only through `neighbor_idx`, with no pooling across atoms and a row-wise `norm_1`, so restricting the bank removes rows no query can name and changes nothing any query reads. It is not bit-exact: the smaller row count changes the GEMM kernel the library selects, and the backward of `index_select` is an indeterminate `index_add`. The union is not the fraction of atoms routed high-resolution.

3.4 Implementation

The production Standard configuration uses four layers at a 6 Å cutoff. The production Flash configuration uses six layers at a 5 Å cutoff and carries its low-resolution stream at 32 channels per degree against 256 in the high-resolution stream. The router–architecture bindings above are training-time choices and are used unchanged at evaluation. The two configurations therefore differ in depth, cutoff, routing and state layout at once, so comparing them is not a single-variable ablation (Appendix I).

The implementation is compiled PyTorch (Ansel et al., 2024) with custom tile-level kernels (Tillet et al., 2019), configured through Hydra and trained data-parallel across GPUs in fp32, with TF32 disabled and compilation enabled at inference, which is the mode every timing in this paper was taken in. The high-resolution and low-resolution heads are trained by warm-starting on direct forces and then fine-tuning conservatively (Fu et al., 2025): the warm-start stage predicts the force directly, and the fine-tuning stage differentiates the merged energy of Equation 1, which is the weight set all accuracy results in this paper use. The routing mask is an input-format obligation and not only a model setting: it must be emitted on every inference path, because a harness covering only one format benchmarks Flash as Standard (Appendix A). Every accuracy row states its force mode (Bigi et al., 2024), and training optimises a weighted energy-and-force objective (Levine et al., 2025) on the merged scalar. Optimizer-state active-parameter counts for the production and other audited checkpoints are reported in Appendix E.1, and the eight numerical gates that certify the implementation — rotation equivariance, finite-difference agreement with the autograd force, size consistency, angular-order creation, empty-stream behaviour, gradient flow and no-op equivalence — are given in full in Appendix C.

4 Experiments

Experimental Setup. Accuracy is measured on the full 2762021-sample OMol25 composition validation split and on SPICE, with uniform-capacity E2Former-V2 as the primary comparison. Efficiency is measured on one H200 across five solvated systems and on an A100-PCIE-40GB waterbox ladder from 300 to 74001 atoms. No table mixes devices; out-of-memory outcomes are retained. Pure water is the all-low-resolution limit of Flash, in which the high-resolution stack is bypassed. Dataset, timing and aggregation protocols are given in Appendix B.

Accuracy and Efficiency. On OMol25, Standard reaches 8.146 meV/Å force MAE against E2Former-V2's published 8.38, a 2.8% reduction, whereas Flash reaches 11.272, a 34.5% in-

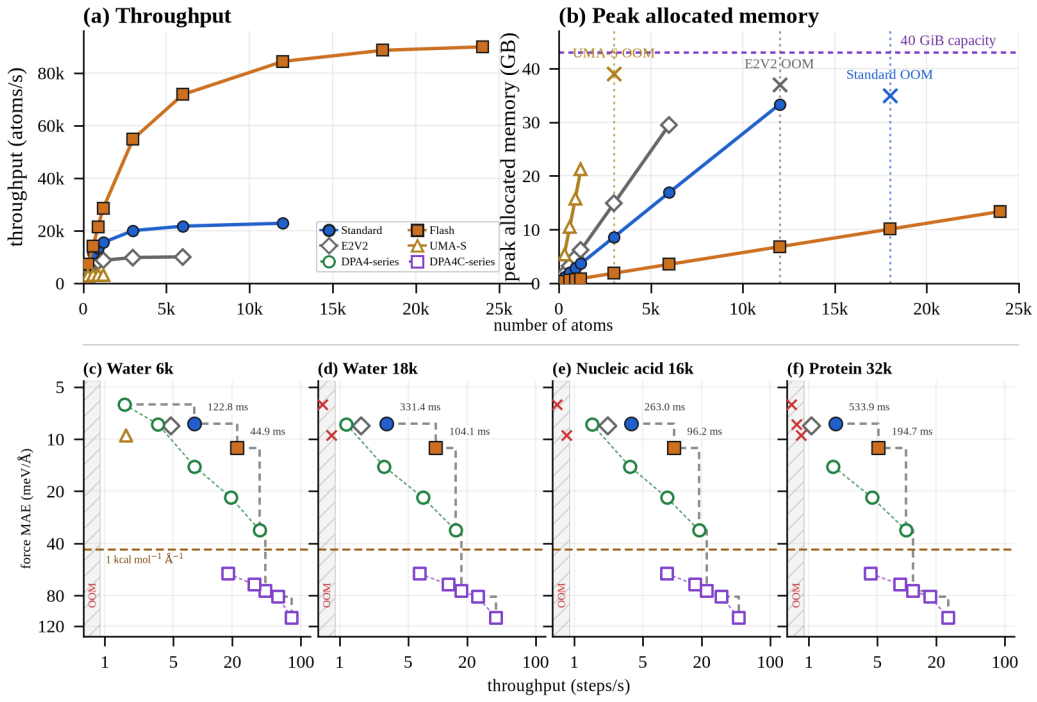


Figure 2: Efficiency under two device-specific protocols. **(a,b)** Single-step throughput and peak allocated memory on the A100 water-box ladder. **(c–f)** Force MAE on the **OMol25 composition validation split** (Table 1) against H200 throughput for benchmark systems (see Table 6 for details). The DPA4-series models include DPA4 Nano/Mini/Neo/Air/Plus; the DPA4C-series models include DPA4C Nano/Mini/Neo/Air/Plus. The A100 and H200 rows use different devices and timing protocols and are not pooled. Red crosses mark OOM outcomes, and gray dashed steps trace the empirical frontier.

crease that remains below the 1 kcal mol^{-1} Å^{-1} chemical-accuracy threshold (Chmiela et al., 2017) (Table 1). On SPICE, after retraining and evaluation at one common caliber, Standard reaches 5.89 meV/Å against E2Former-V2's 7.04, while Flash reaches 8.17 (Table 2); metric and aggregation details are in Appendix B.1. Energy is reported alongside force: on OMol25, Standard and Flash reach 2.190 and 3.941 meV/atom against E2Former-V2's 1.16, and on SPICE they reach 0.30 and 0.39 against 0.31.

On the frozen H200 campaign, Standard runs 1.75–1.84× and Flash 4.00–5.85× faster than E2Former-V2 across the five systems; both are non-dominated on every accuracy–throughput frontier (Figure 2, Table 6). On the 32437-atom protein, Flash uses 161.18 ms and 19.251 GB, against Standard's 477.38 ms and 82.899 GB (Table 4).

The A100 water-box ladder shows the same ordering at every shared size: Standard is 1.23–2.16× and Flash 1.40–7.15× faster than E2Former-V2 (Figure 2). E2Former-V2 and Standard run out of memory at 12000 and 18000 atoms, whereas Flash reaches 74001 atoms at 40.834 GB; the complete timing and memory records are in Appendix D.

High-resolution-role atoms comprise 49.1–49.6% of force components but carry 76–78% of absolute error in both routed configurations. Standard reduces participation only, whereas Flash reduces both persistent width and participation and routes more atoms low-resolution; their high-to-low-resolution error ratios nevertheless remain close, 3.634× and 3.324×. This consistency suggests that the imbalance more likely follows the chemistry of the roles than the precise capacity placement, although the two configurations are not a causal ablation (Figure 10, Appendix F).

Physical Reliability. Because averaged error does not rank usability (Fu et al., 2022; Liu et al., 2023; Raja et al., 2024), we gate the reduced-capacity water path on model-independent aqueous

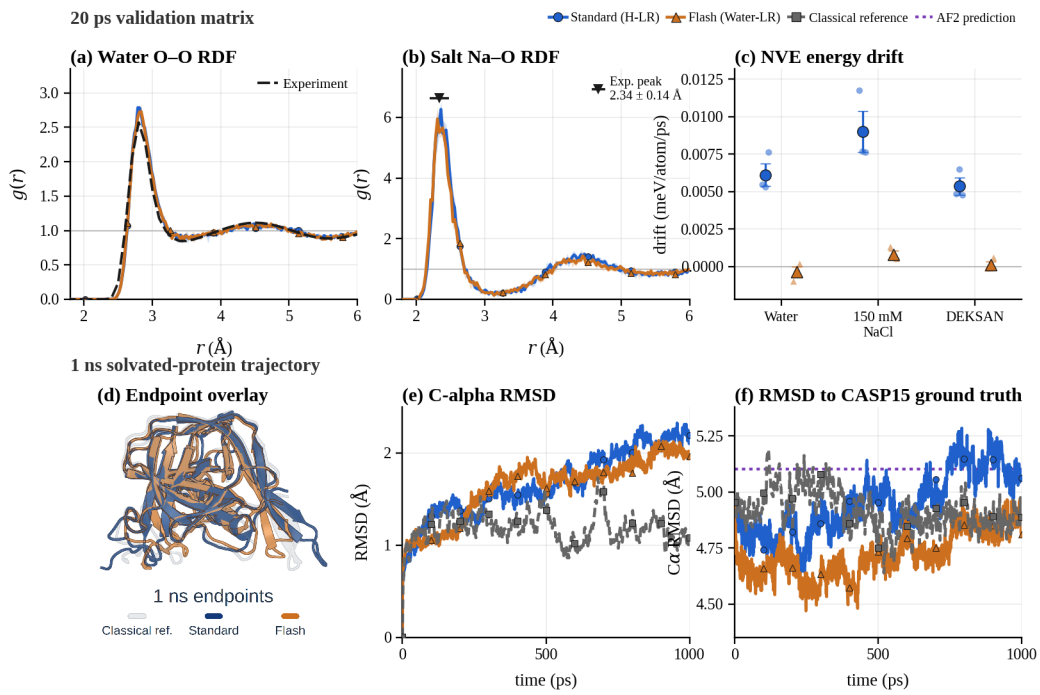


Figure 3: Physical reliability across two campaigns. **(a,b)** Water O–O and salt Na–O radial distribution functions from 20 ps NVT simulation results, three-seed means over 10–20 ps, with the seed range shaded. The water reference is the 295.1 K X-ray curve of Skinner et al. (2013); the salt marker is the 2.34 ± 0.14 Å Na–O peak of Mancinelli et al. (2007). **(c)** Signed total-energy drift fitted over 2–20 ps in NVE. The largest mean, 0.0090 meV/atom/ps, corresponds to 0.162 meV/atom over the 18 ps fit window. **(d–f)** Separate 1 ns, fixed-cell NVT trajectories of one 32437-atom solvated protein with one seed: endpoint overlay, Cα RMSD from the equilibrated start, and Cα RMSD to the CASP15 T1123-D1 experimental structure. Gray dashed curves in **(e,f)** are the AMBER ff14SB/TIP3P GROMACS control propagated from the same equilibrated start. The purple dotted line in **(f)** is the CASP15 AlphaFold2 submission scored against the same ground truth.

references (Appendix G) rather than on force error (Omranpour et al., 2024). Across water, a 150 mM salt solution and a solvated peptide, all 54 production legs complete without non-finite values, overlap or water fusion, and the three-seed mean signed constant-energy drift remains below 0.01 meV/atom/ps over the 2–20 ps fit window. The recorded water and ion radial-distribution landmarks remain within a few hundredths of an Ångstrom of experiment (Skinner et al., 2013; Mancinelli et al., 2007) (Figure 3; detailed curves are in Appendix G). Separate Standard and Flash checkpoints also complete 1000 ps on the 32437-atom solvated protein without water fusion, nonbonded clash or solute collapse (Figure 3, Table 20).

The short-range scan provides the main qualification: at 1.80 Å, both production high-resolution streams reproduce 99.2–100.1% of the DFT water-dimer repulsion wall, whereas Flash's reduced-capacity water path reproduces 49.7%. Its minimum remains at the reference position and depth, but neither averaged force error nor the radial distribution function reveals this short-range softening (Appendix G).

Ablation Study. Neighbor-union filtering is inference-only and leaves trained weights unchanged. It reduces Flash latency by 23.1–27.3% (1.30–$1.37\times$) across all five systems, while Standard's deltas remain within $\pm 0.4\%$ (Figure 6). All ten model–system pairs pass the frozen 10^{-3} eV energy and 10^{-3} eV/Å force consistency thresholds.

Table 1: Accuracy on OMol25 Val-Comp against uniform-capacity potentials. Best and second-best values are bold and underlined, respectively; all Standard/Flash results are conservative.

Dataset	Model	Biomolecules		Electrolytes		Metals		Organics		Total	
		E	F	E	F	E	F	E	F	E	F
All	eSEN-sm-cons (Fu et al., 2025)	0.59	**4.61**	**1.01**	8.08	2.30	28.86	**0.84**	**11.11**	1.27	8.25
All	UMA-S (Wood et al., 2025)	**0.53**	5.69	2.69	11.65	4.63	37.85	1.00	13.15	3.62	12.02
All	MACE-OMOL (Batatia et al., 2022)	1.76	6.05	2.15	10.87	**2.15**	24.51	2.31	23.70	1.84	10.10
All	E2Former-V2-Cons. (Huang et al., 2026)	1.01	4.96	1.52	10.11	2.18	**24.40**	1.26	19.69	**1.16**	8.38
All	E2Former-V3-Standard	2.01	5.10	2.27	**8.00**	2.47	25.56	1.89	17.07	2.19	**8.15**
All	E2Former-V3-Flash	5.00	7.68	3.92	11.45	2.86	28.40	2.45	20.07	3.94	11.27

Table 2: Accuracy on SPICE test set against uniform-capacity potentials. Best and second-best values are bold and underlined, respectively; all Standard/Flash results are conservative.

Model	PubChem		Monomers		Dimers		Dipeptides		Solv. Amino		Water		Qmugs		All	
	E	F	E	F	E	F	E	F	E	F	E	F	E	F	E	F
MACE Small	1.41	35.68	1.04	17.63	0.98	16.31	0.84	25.07	1.60	38.56	1.67	28.53	1.03	41.45	1.27	32.43
MACE Medium	0.91	20.57	0.63	9.36	0.58	9.02	0.52	14.27	1.21	23.26	0.76	15.27	0.69	23.58	0.80	18.63
MACE Large	0.88	14.75	0.59	6.58	0.54	6.62	0.42	10.19	0.98	19.43	0.83	13.57	0.45	16.93	0.77	13.39
E2Former-V1 33M	0.67	8.90	0.49	7.10	0.43	4.01	0.51	5.63	1.10	19.20	0.96	13.50	0.65	10.20	0.60	8.12
eSEN-Direct (Fu et al., 2025)							—								0.56	10.98
eSEN-Cons. (Fu et al., 2025)							—								**0.23**	6.36
E2Former-V2-Direct (Huang et al., 2026)	0.65	9.49	0.46	4.59	0.28	2.40	0.33	5.63	0.56	13.37	0.70	7.43	0.63	11.69	0.54	8.32
E2Former-V2-Cons. (Huang et al., 2026)	0.37	8.00	0.19	3.25	0.18	2.36	0.20	4.45	0.30	9.51	0.30	**5.26**	**0.25**	9.43	0.31	7.04
E2Former-V3-Standard (4L, 6 Å)	**0.36**	**6.68**	**0.18**	**2.62**	**0.17**	**2.01**	**0.14**	**3.79**	**0.22**	**7.94**	**0.24**	5.57	**0.25**	**7.98**	0.30	**5.89**
E2Former-V3-Flash (6L, 32 ch, 5 Å)	0.43	9.10	0.41	3.42	0.31	3.20	0.23	6.05	1.01	22.98	0.90	16.28	0.42	12.55	0.39	8.17

5 Conclusion

E2Former-V3 allocates capacity by atom role while keeping one all-atom conservative potential. **Atom-role-aware capacity allocation moves the speed–accuracy frontier.** On OMol25, Standard reaches 8.146 meV/Å force MAE against E2Former-V2's 8.38; Flash reaches 11.272 and remains within the chemical-accuracy threshold (Chmiela et al., 2017). On H200, they run 1.75–$1.84\times$ and 4.00–$5.85\times$ faster and are non-dominated on all five systems, while improving peak memory and feasible system size (Figure 2).

Both operating points are usable in molecular dynamics. Over water, a 150 mM salt solution and a solvated peptide, all 54 production legs finish without numerical or structural failure; RDF landmarks match experiment within a few hundredths of an Ångstrom, and the peptide's 196-bond graph remains intact. Both checkpoints also complete 1000 ps on the 32437-atom solvated protein without water fusion, nonbonded clash or solute collapse.

The role-stratified error imbalance persists under both capacity allocations. Standard changes participation only and has a $3.634\times$ HR/LR error ratio; Flash changes width and participation and has a $3.324\times$ ratio. In both, high-resolution-role atoms are 49.1–49.6% of force components but carry 76–78% of absolute error. This suggests a chemical origin rather than precise capacity placement, but is not a causal ablation (Appendix F).

Flash's short-range water surface is softer than the allocation alone would predict. Measured on the production configuration, its low-resolution stream reproduces 49.7% of a DFT reference repulsion wall at 1.80 Å, versus 100.1% and 99.2% for the two high-resolution streams. It still holds a bound state with a reference-matched minimum, but neither force MAE nor the RDF exposes the softening. This bounds use of the reduced-capacity water path where short-range repulsion or proton transfer matters (Appendix G).

Within the present scope—explicitly solvated biomolecules, single-seed evaluation and the bounds in Appendix I—atom-role-aware capacity allocation reduces cost while retaining practical accuracy and stability.

AI USE STATEMENT

This manuscript was produced with an LLM-driven multi-agent writing pipeline, and we disclose that use in full. AI assistance was used to draft prose in every section; to generate all figures from the experimental record by writing and running the plotting code; to locate, retrieve and cross-verify the cited literature against multiple bibliographic indices and to format the bibliography; to propose the document structure, section plan and title; and to transcribe measured results into the tables shown here. The research hypotheses were proposed and refined exclusively by the human authors. Generative AI did, however, participate in research execution: it was used to help design and provide feedback on the research methodology and experiments, and to assist with implementing the methods, including writing and revising research code. The authors evaluated and approved all AI-assisted methodological and experimental choices and reviewed and tested the AI-assisted implementations. The training runs and every accuracy, latency, memory, potential-energy-surface and molecular-dynamics measurement reported in this paper were conducted and validated by the authors, and every number in the paper originates in those measurements rather than in a model's output. AI was not used to generate synthetic data, to formulate or support mathematical claims, or to decide what the results mean. We reviewed all AI-assisted output against the experimental record and corrected it where the two disagreed — including one summary statement in our own records that its own underlying table contradicted, which is why the low-resolution-capacity spreads in the Physical Reliability paragraph of Section 4 are reported per metric. The authors take full responsibility for the final content, including all AI-assisted portions. There is no AI co-authorship.

ETHICS STATEMENT

This work involves no human subjects, no personal or sensitive data, and no data collection from people; the training corpus consists of quantum-chemical calculations. The dual-use surface is low: the models predict energies and forces for aqueous biomolecular systems and confer no capability that established simulation methods do not already provide. The honest risk is scientific rather than societal. An under-resolved solvent stream can reproduce a correct radial distribution function while failing short-range repulsion and proton transfer, so a user who validated only averaged error or only water structure could deploy a potential that is quantitatively wrong where it matters. This paper reports its physical-gate failures rather than leaving them out. It flags the two dynamics legs whose settings limit what they can show: a box whose cutoff fills 97% of its half-width, stated in Appendix G with its evidence beside it, and the pure-water column where the routing degenerates to its narrow stream alone, stated in the experimental setup. It states the measured short-range softening of its aggressive configuration in its conclusion, and gives both potential-energy-surface scans in full in Appendix G, including the proton-transfer gate, which the same configuration clears. It also states the coverage of its dynamics evidence rather than the horizon it would prefer to claim: 20 ps on three systems, and one system carried to 1000 ps, which does not transfer to the others. We also report the role-stratified result with its non-causal interpretation: similar HR/LR ratios under the two different capacity reductions suggest, but do not prove, that the imbalance is chemical rather than caused by the precise allocation.

REPRODUCIBILITY STATEMENT

The architecture, both schedules, the two static atom-role routers and the mask-merged readout of Equation 1 are specified in Section 3, with the two capacity allocations built out in Section 3.4 and implementation details in Appendix A. The training corpus, level of theory, the two validation calibers, the metric definitions, the timing protocol, both devices, the baseline-selection rule and every benchmark system with its atom count are given in Section 4. Numerical certification is given in full in Appendix C, and the extended cost, profiler and liquid-water tables are in Appendices D to H, with the boundary conditions collected in Appendix I. Two constraints on reproduction should be read with the above: every trained result is single-seed, so a replication should expect run-to-run variation we did not measure, and no code URL is given because this submission is anonymous.

REFERENCES

Noam Agmon. The grotthuss mechanism. *Chemical Physics Letters*, 1995. doi: 10.1016/0009-2614(95)00905-j.

Hans C Andersen. Rattle: A "velocity" version of the shake algorithm for molecular dynamics calculations. *Journal of Computational Physics*, 1983. doi: 10.1016/0021-9991(83)90014-1.

Jason Ansel, Edward Yang, Horace He, Natalia Gimelshein, Animesh Jain, Michael Voznesensky, Bin Bao, Peter Bell, David Berard, Evgeni Burovski, Geeta Chauhan, Anjali Chourdia, Will Constable, Alban Desmaison, Zachary DeVito, Elias Ellison, Will Feng, Jiong Gong, Michael Gschwind, Brian Hirsh, Sherlock Huang, Kshiteej Kalambarkar, Laurent Kirsch, Michael Lazos, Mario Lezcano, Yanbo Liang, Jason Liang, Yinghai Lu, C. K. Luk, Bert Maher, Yunjie Pan, Christian Puhrsch, Matthias Reso, Mark Saroufim, Marcos Yukio Siraichi, Helen Suk, Shunting Zhang, Michael Suo, Phil Tillet, Xu Zhao, Eikan Wang, Keren Zhou, Richard Zou, Xiaodong Wang, Ajit Mathews, William Wen, Gregory Chanan, Peng Wu, and Soumith Chintala. Pytorch 2: Faster machine learning through dynamic python bytecode transformation and graph compilation. *Proceedings of the 29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 2*, 2024. doi: 10.1145/3620665.3640366.

Ilyes Batatia, Dávid Péter Kovács, Gregor N. C. Simm, Christoph Ortner, and Gábor Csányi. Mace: Higher order equivariant message passing neural networks for fast and accurate force fields. *Advances in Neural Information Processing Systems*, 2022. doi: 10.52202/068431-0830.

Ilyes Batatia, Simon Batzner, Dávid Péter Kovács, Albert Musaelian, Gregor N. C. Simm, Ralf Drautz, Christoph Ortner, Boris Kozinsky, and Gábor Csányi. The design space of e(3)-equivariant atom-centred interatomic potentials. *Nature Machine Intelligence*, 2025. doi: 10.1038/s42256-024-00956-x.

Simon Batzner, Albert Musaelian, Lixin Sun, Mario Geiger, Jonathan P. Mailoa, Mordechai Kornbluth, Nicola Molinari, Tess E. Smidt, and Boris Kozinsky. E(3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials. *Nature Communications*, 2022. doi: 10.1038/s41467-022-29939-5.

Filippo Bigi, Marcel Langer, and Michele Ceriotti. The dark side of the forces: assessing non-conservative force models for atomistic machine learning. *arXiv preprint*, 2024. doi: 10.48550/arXiv.2412.11569.

Daniel Bolya, Cheng-Yang Fu, Xiaoliang Dai, Peizhao Zhang, Christoph Feichtenhofer, and Judy Hoffman. Token merging: Your vit but faster. *arXiv preprint*, 2022. doi: 10.48550/arXiv.2210.09461.

Lennard Böselt, Moritz Thürlemann, and Sereina Riniker. Machine learning in qm/mm molecular dynamics simulations of condensed-phase systems. *Journal of Chemical Theory and Computation*, 2021. doi: 10.1021/acs.jctc.0c01112.

Stefan Chmiela, Alexandre Tkatchenko, Huziel E. Sauceda, Igor Poltavsky, Kristof T. Schütt, and Klaus-Robert Müller. Machine learning of accurate energy-conserving molecular force fields. *Science Advances*, 3:e1603015, 2017. doi: 10.1126/sciadv.1603015.

Tri Dao, Daniel Y. Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. Flashattention: Fast and memory-efficient exact attention with io-awareness. *Advances in Neural Information Processing Systems*, 2022. doi: 10.52202/068431-1189.

Gabriel de Miranda Nascimento, Marc L. Descoteaux, Laura Zichi, Chuin Wei Tan, William C. Witt, Nicola Molinari, Sriteja Mantha, Daniil Kitchaev, Mordechai Kornbluth, Karim Gadelrab, Charles Tuffile, and Boris Kozinsky. Mixture of experts framework in machine learning interatomic potentials for atomistic simulations. *arXiv preprint*, 2026. doi: 10.48550/arXiv.2604.26143.

Peter Eastman, Pavan Kumar Behara, David L. Dotson, Raimondas Galvelis, John E. Herr, Josh T. Horton, Yuezhi Mao, John D. Chodera, Benjamin P. Pritchard, Yihan Wang, Gianni De Fabritiis, and Thomas E. Markland. SPICE, a dataset of drug-like molecules and peptides for training machine learning potentials. *Scientific Data*, 10:11, 2023. doi: 10.1038/s41597-022-01882-6.

Maha Elbayad, Jiatao Gu, Edouard Grave, and Michael Auli. Depth-adaptive transformer. *arXiv preprint*, 2019. doi: 10.48550/arXiv.1910.10073.

William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *arXiv preprint*, 2021. doi: 10.48550/arXiv.2101.03961.

Xiang Fu, Zhenghao Wu, Wujie Wang, Tian Xie, Sinan Keten, Rafael Gomez-Bombarelli, and Tommi Jaakkola. Forces are not enough: Benchmark and critical evaluation for machine learning force fields with molecular simulations. *arXiv preprint*, 2022. doi: 10.48550/arXiv.2210.07237.

Xiang Fu, Brandon M. Wood, Luis Barroso-Luque, Daniel S. Levine, Meng Gao, Misko Dzamba, and C. Lawrence Zitnick. Learning smooth and expressive interatomic potentials for physical property prediction. *arXiv preprint*, 2025. doi: 10.48550/arXiv.2502.12147.

Fabian B. Fuchs, Daniel E. Worrall, Volker Fischer, and Max Welling. Se(3)-transformers: 3d roto-translation equivariant attention networks. *arXiv preprint*, 2020. doi: 10.48550/arXiv.2006.10503.

Johannes Gasteiger, Janek Groß, and Stephan Günnemann. Directional message passing for molecular graphs. *arXiv preprint*, 2020. doi: 10.48550/arXiv.2003.03123.

Johannes Gasteiger, Florian Becker, and Stephan Günnemann. Gemnet: Universal directional graph neural networks for molecules. *arXiv preprint*, 2021. doi: 10.48550/arXiv.2106.08903.

Mario Geiger and Tess Smidt. e3nn: Euclidean neural networks. *arXiv preprint*, 2022. doi: 10.48550/arXiv.2207.09453.

Alex Graves. Adaptive computation time for recurrent neural networks. *arXiv preprint*, 2016. doi: 10.48550/arXiv.1603.08983.

Berk Hess, Henk Bekker, Herman J. C. Berendsen, and Johannes G. E. M. Fraaije. Lincs: A linear constraint solver for molecular simulations. *Journal of Computational Chemistry*, 1997. doi: 10.1002/(sici)1096-987x(199709)18:12⟨1463::aid-jcc4⟩3.0.co;2-h.

Chad W. Hopkins, Scott Le Grand, Ross C. Walker, and Adrian E. Roitberg. Long-time-step molecular dynamics through hydrogen mass repartitioning. *Journal of Chemical Theory and Computation*, 2015. doi: 10.1021/ct5010406.

Junbao Hu, Liyang Zhou, and Jian Jiang. Efficient machine learning force field for large-scale molecular simulations of organic systems. *CCS Chemistry*, 2025. doi: 10.31635/ccschem.024.202404785.

Lin Huang, Chengxiang Huang, Ziang Wang, Yiyue Du, Chu Wang, Haocheng Lu, Yunyang Li, Xiaoli Liu, Arthur Jiang, and Jia Zhang. E2former-v2: On-the-fly equivariant attention with linear activation memory. *arXiv preprint*, 2026. doi: 10.48550/arXiv.2601.16622.

David Immel, Ralf Drautz, and Godehard Sutmann. Adaptive-precision potentials for large-scale atomistic simulations. *The Journal of Chemical Physics*, 162(11):114119, 2025. doi: 10.1063/5.0245877.

InstaDeepAI. MLIP-tutorials: experimental O–O radial distribution function of liquid water at 295.1 K (digitised from skinner et al. 2013). Hugging Face repository, 2025. URL `https://huggingface.co/InstaDeepAI/MLIP-tutorials`. Accessed 2026-09-23.

Andrew Jaegle, Felix Gimeno, Andrew Brock, Andrew Zisserman, Oriol Vinyals, and Joao Carreira. Perceiver: General perception with iterative attention. *arXiv preprint*, 2021. doi: 10.48550/arXiv.2103.03206.

Leonid Kahle, Benoit Minisini, Tai Bui, Jeremy T. First, Corneliu Buda, Thomas Goldman, and Erich Wimmer. A dual-cutoff machine-learned potential for condensed organic systems obtained via uncertainty-guided active learning. *Physical Chemistry Chemical Physics*, 2024. doi: 10.1039/d4cp01980f.

Paul Katzberger and Sereina Riniker. Implicit solvent approach based on generalized born and transferable graph neural networks for molecular dynamics simulations. *The Journal of Chemical Physics*, 2023. doi: 10.1063/5.0147027.

Sebastian Kmiecik, Dominik Gront, Michal Kolinski, Lukasz Wieteska, Aleksandra Elzbieta Dawid, and Andrzej Kolinski. Coarse-grained protein models and their applications. *Chemical Reviews*, 2016. doi: 10.1021/acs.chemrev.6b00163.

Daniel S. Levine, Muhammed Shuaibi, Evan Walter Clark Spotte-Smith, Michael G. Taylor, Muhammad R. Hasyim, Kyle Michel, Ilyes Batatia, Gábor Csányi, Misko Dzamba, Peter Eastman, Nathan C. Frey, Xiang Fu, Vahe Gharakhanyan, Aditi S. Krishnapriyan, Joshua A. Rackers, Sanjeev Raja, Ammar Rizvi, Andrew S. Rosen, Zachary Ulissi, Santiago Vargas, C. Lawrence Zitnick, Samuel M. Blau, and Brandon M. Wood. The open molecules 2025 (omol25) dataset, evaluations, and models. *arXiv preprint*, 2025. doi: 10.48550/arXiv.2505.08762.

Tiancheng Li, Wentao Li, Anyang Peng, Jianming Xue, Linfeng Zhang, Duo Zhang, and Han Wang. Dpa4: Pushing the accuracy-cost frontier of interatomic potentials with emfa so(2) convolution. *arXiv preprint*, 2026a. doi: 10.48550/arXiv.2606.02419.

Tiancheng Li, Jianming Xue, Linfeng Zhang, Duo Zhang, and Han Wang. Universal machine-learning molecular dynamics at the speed of empirical potentials. *arXiv preprint*, 2026b. doi: 10.48550/arXiv.2608.19041.

Yunyang Li, Lin Huang, Zhihao Ding, Chu Wang, Xinran Wei, Han Yang, Zun Wang, Chang Liu, Yu Shi, Peiran Jin, Tao Qin, Mark Gerstein, and Jia Zhang. E2former: An efficient and equivariant transformer with linear-scaling tensor products. *Advances in Neural Information Processing Systems*, 2025. doi: 10.52202/085713-0797.

Yi-Lun Liao and Tess Smidt. Equiformer: Equivariant graph attention transformer for 3d atomistic graphs. *arXiv preprint*, 2022. doi: 10.48550/arXiv.2206.11990.

Yi-Lun Liao, Brandon Wood, Abhishek Das, and Tess Smidt. Equiformerv2: Improved equivariant transformer for scaling to higher-degree representations. *arXiv preprint*, 2023. doi: 10.48550/arXiv.2306.12059.

Yunsheng Liu, Xingfeng He, and Yifei Mo. Discrepancies and error evaluation metrics for machine learning interatomic potentials. *npj Computational Materials*, 2023. doi: 10.1038/s41524-023-01123-3.

Yuzhi Liu, Duo Zhang, Anyang Peng, Weinan E, Linfeng Zhang, and Han Wang. Mixture of experts architectures for machine learning interatomic potentials. *npj Artificial Intelligence*, 2026. doi: 10.1038/s44387-026-00144-9.

Shengjie Luo, Tianlang Chen, and Aditi S. Krishnapriyan. Enabling efficient equivariant operations in the fourier basis via gaunt tensor products. *arXiv preprint*, 2024. doi: 10.48550/arXiv.2401.10216.

R. Mancinelli, A. Botti, F. Bruni, M. A. Ricci, and A. K. Soper. Hydration of sodium, potassium, and chloride ions in solution and the concept of structure maker/breaker. *The Journal of Physical Chemistry B*, 111(48):13570–13577, 2007. doi: 10.1021/jp075913v.

Siewert J. Marrink, H. Jelger Risselada, Serge Yefimov, D. Peter Tieleman, and Alex H. de Vries. The martini force field: Coarse grained model for biomolecular simulations. *The Journal of Physical Chemistry B*, 2007. doi: 10.1021/jp071097f.

Shuichi Miyamoto and Peter A. Kollman. Settle: An analytical version of the shake and rattle algorithm for rigid water models. *Journal of Computational Chemistry*, 1992. doi: 10.1002/jcc.540130805.

Albert Musaelian, Simon Batzner, Anders Johansson, Lixin Sun, Cameron J. Owen, Mordechai Kornbluth, and Boris Kozinsky. Learning local equivariant representations for large-scale atomistic dynamics. *Nature Communications*, 2023. doi: 10.1038/s41467-023-36329-y.

Amir Omranpour, Pablo Montero De Hijes, Jörg Behler, and Christoph Dellago. Perspective: Atomistic simulations of water and aqueous systems with machine learning potentials. *The Journal of Chemical Physics*, 2024. doi: 10.1063/5.0201241.

Saro Passaro and C. Lawrence Zitnick. Reducing so(3) convolutions to so(2) for efficient equivariant gnns. *arXiv preprint*, 2023. doi: 10.48550/arXiv.2302.03655.

Helena L. Pi, Juan L. Aragones, Carlos Vega, Eva G. Noya, Jose L. F. Abascal, Miguel A. Gonzalez, and Carl McBride. Anomalies in water as obtained from computer simulations of the tip4p/2005 model: density maxima, and density, isothermal compressibility and heat capacity minima. *Molecular Physics*, 2009. doi: 10.1080/00268970902784926.

Eric Qu and Aditi S. Krishnapriyan. The importance of being scalable: Improving the speed and accuracy of neural network interatomic potentials across chemical domains. *Advances in Neural Information Processing Systems*, 2024. doi: 10.52202/079017-4412.

Sanjeev Raja, Ishan Amin, Fabian Pedregosa, and Aditi S. Krishnapriyan. Stability-aware training of machine learning force fields with differentiable boltzmann estimators. *arXiv preprint*, 2024. doi: 10.48550/arXiv.2402.13984.

Yongming Rao, Wenliang Zhao, Benlin Liu, Jiwen Lu, Jie Zhou, and Cho-Jui Hsieh. Dynamicvit: Efficient vision transformers with dynamic token sparsification. *arXiv preprint*, 2021. doi: 10.48550/arXiv.2106.02034.

David Raposo, Sam Ritter, Blake Richards, Timothy Lillicrap, Peter Conway Humphreys, and Adam Santoro. Mixture-of-depths: Dynamically allocating compute in transformer-based language models. *arXiv preprint*, 2024. doi: 10.48550/arXiv.2404.02258.

Jean-Paul Ryckaert, Giovanni Ciccotti, and Herman J.C Berendsen. Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *Journal of Computational Physics*, 1977. doi: 10.1016/0021-9991(77)90098-5.

Kristof T. Schütt, Pieter-Jan Kindermans, Huziel E. Sauceda, Stefan Chmiela, Alexandre Tkatchenko, and Klaus-Robert Müller. Schnet: A continuous-filter convolutional neural network for modeling quantum interactions. *Advances in Neural Information Processing Systems*, 2017. doi: 10.48550/arxiv.1706.08566.

Hans Martin Senn and Walter Thiel. Qm/mm methods for biomolecular systems. *Angewandte Chemie International Edition*, 2009. doi: 10.1002/anie.200802019.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint*, 2017. doi: 10.48550/arXiv.1701.06538.

Lawrie B. Skinner, Congcong Huang, Daniel Schlesinger, Lars G. M. Pettersson, Anders Nilsson, and Chris J. Benmore. Benchmark oxygen-oxygen pair-distribution function of ambient water from x-ray diffraction measurements with a wide q -range. *The Journal of Chemical Physics*, 2013. doi: 10.1063/1.4790861.

Chunhu Tan, Lijiang Yang, and Ray Luo. How well does poisson-boltzmann implicit solvent agree with explicit solvent? a quantitative analysis. *The Journal of Physical Chemistry B*, 2006. doi: 10.1021/jp063479b.

Nathaniel Thomas, Tess Smidt, Steven Kearnes, Lusann Yang, Li Li, Kai Kohlhoff, and Patrick Riley. Tensor field networks: Rotation- and translation-equivariant neural networks for 3d point clouds. *arXiv preprint*, 2018. doi: 10.48550/arXiv.1802.08219.

Philippe Tillet, H. T. Kung, and David Cox. Triton: an intermediate language and compiler for tiled neural network computations. *Proceedings of the 3rd ACM SIGPLAN International Workshop on Machine Learning and Programming Languages*, 2019. doi: 10.1145/3315508.3329973.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *arXiv preprint*, 2017. doi: 10.48550/arXiv.1706.03762.

W. Wagner and A. Pruß. The iapws formulation 1995 for the thermodynamic properties of ordinary water substance for general and scientific use. *Journal of Physical and Chemical Reference Data*, 2002. doi: 10.1063/1.1461829.

A. Warshel and M. Levitt. Theoretical studies of enzymic reactions: Dielectric, electrostatic and steric stabilization of the carbonium ion in the reaction of lysozyme. *Journal of Molecular Biology*, 1976. doi: 10.1016/0022-2836(76)90311-9.

Brandon M. Wood, Misko Dzamba, Xiang Fu, Meng Gao, Muhammed Shuaibi, Luis Barroso-Luque, Kareem Abdelmaqsoud, Vahe Gharakhanyan, John R. Kitchin, Daniel S. Levine, Kyle Michel, Anuroop Sriram, Taco Cohen, Abhishek Das, Ammar Rizvi, Sushree Jagriti Sahoo, Zachary W. Ulissi, and C. Lawrence Zitnick. Uma: A family of universal models for atoms. *Advances in Neural Information Processing Systems*, 2025. doi: 10.52202/085713-4310.

Yanqi Zhou, Tao Lei, Hanxiao Liu, Nan Du, Yanping Huang, Vincent Zhao, Andrew Dai, Zhifeng Chen, Quoc Le, and James Laudon. Mixture-of-experts with expert choice routing. *Advances in Neural Information Processing Systems*, 2022. doi: 10.52202/068431-0515.

A Architecture Overview

Figure 4 expands the E2Attention operator shared by the uniform-capacity E2Former backbone and the two atom-role-aware allocations. From atomic numbers and coordinates, neighbour construction supplies radial and directional edge features within the cutoff; the embedding initialises equivariant atom states; repeated interaction blocks refine them; and the energy readout produces atomic energies whose sum is differentiated with respect to the coordinates to obtain conservative forces. In E2Attention (Huang et al., 2026), target atoms supply queries, source atoms supply keys and values, and cached edge features modulate each target update. Panels (b) and (c) show the interleaved and interleaved_flash allocations used by Standard and Flash; Section 3 specifies their exact schedules and block semantics.

Both atom-role-aware allocations use separate high- and low-resolution energy heads: Standard applies them to the two role partitions of its shared state, whereas Flash reads them from its two persistent states; the selected atomic energies are summed before differentiation.

The routing mask is an input-format obligation and must be emitted on every inference path. The benchmark harness originally emitted the water-oxygen mask for only one input format, which would silently execute Flash checkpoints as Standard; after the fix, the control-system mask covered 1963 of 6085 atoms.

B Full Measurement Protocol

Accuracy is reported as energy MAE in meV/atom and force MAE in meV/Å, the latter a per-component MAE over the three Cartesian components of every valid atom, which must not be conflated with the per-atom force-vector L_2 error used as the training objective. Two independent accuracy-caliber distinctions are labeled per row and are not interchangeable: the full 2,762,021-sample validation split versus its 1M subsample, and the training-time batch mean (an unweighted mean over token-budget batches) versus the exact component-weighted MAE used for the role-stratified results. Efficiency is reported as single-step latency in ms, throughput in atoms/s and peak allocated device memory in GB, with out-of-memory outcomes preserved rather than dropped. Physical reliability is reported through trajectory completion and structural-health checks, temperature and constant-pressure acceptance, the coupled-coordinate proton-transfer barrier in kcal/mol, the water-dimer oxygen-oxygen wall energy at 1.80 Å in kcal/mol, radial-distribution peak positions and heights, first-shell coordination numbers, density in g/cm^3, isothermal compressibility, signed constant-energy total-energy drift rate in meV/atom/ps and, for the separate water-only analysis, final total-energy displacement in meV/atom. Unless a row states otherwise, all timings used fp32 with TF32 disabled, batch size 1, frozen weights and compilation enabled, with the first compilation iteration excluded and the median of the subsequent timed iterations reported. Two devices were used, an A100-PCIE-40GB and an H200, and their numbers never appear in one table. The Standard and Flash rows in Table 5 and Figure 2 were measured on architecturally identical checkpoints

from the direct-force warm-start stage; because weights are frozen during timing, latency and memory are independent of the learned weight values. Where timing points were repeated, repetitions ran in separate processes; the fixed-device A100 ladder used one fresh process per size, with the 74001-atom boundary repeated twice. Every trained result is single-seed. For Figure 2, every successful throughput value was measured locally on one H200 using the implementation or code repository accompanying the corresponding publication (this work's codebase for Standard and Flash), under the detailed campaign in Table 6. Standard and Flash use the same production checkpoints as their accuracy results; external-family points pair published accuracy with locally measured throughput and retain the published metric-definition caveat, while E2Former-V2 additionally uses an architecturally identical speed-only checkpoint whose weights differ from those behind its published accuracy.

B.1 SPICE Accuracy Protocol

Table 2 evaluates the SPICE test set (Eastman et al., 2023) in the layout of E2Former-V2 Table 1 (Huang et al., 2026), with every row expressed at one aggregation caliber. Energy is in meV/atom; its *All* value is the structure-count-weighted mean of the seven subset MAEs, which reproduces the published MACE Large energy aggregate digit-for-digit. Force is the mean absolute error per Cartesian component in meV/Å; its *All* value is pooled over all force components. The source table does not use one basis: the MACE and E2Former-V1 rows are structure-weighted on both metrics, whereas E2Former-V2 is structure-weighted on energy but component-pooled on force. Table 2 therefore retains the published subset cells but recomputes every derivable *All* value on the stated common basis rather than transcribing a mixed aggregate column.

The force convention was verified, not assumed, from the test file's `MACE_forces` column. Its energy errors identify it as a MACE Large-quality model, which fixes its force reading to 1.08–1.19 per component rather than 2.15–2.39 as per-atom vector L_2. The two models from this work were retrained on SPICE from scratch and evaluated by pooling the full 50195-frame test file. Relative to the published MACE-OFF23 partition, that file contains 209 additional PubChem structures and 12 additional dimers, or 221 structures in total; the resulting aggregate difference is approximately 0.1%.

The published rows are transcribed rather than re-measured, and their training budgets are not matched. For E2Former-V2-Cons., component pooling gives 7.043, consistent with its printed 7.05, whereas structure weighting gives 6.28; its energy aggregate of 0.31 is structure-weighted and is retained. The corresponding recomputed E2Former-V2-Direct aggregate is 0.54/8.32. The eSEN rows publish only *All*, not the seven subset values; their pooling rule is unstated and they are 3.2M-parameter ablations, so they are context rather than commensurable rankings.

The transcription was checked cell by cell against E2Former-V2 Table 1. Its E2Former-V1 Water force cell is 13.50, while the original E2Former-V1 Table 3 prints 13.52; Table 2 follows E2Former-V2 Table 1 and therefore uses 13.50. Either value gives the same displayed component-pooled aggregate of 8.12. The remaining published subset cells agree with the source. The *All* column is 95.6% weighted on PubChem and dimers and is consequently nearly blind to the regimes targeted here: solvated amino acids and water carry only 0.10% and 0.17% of its energy weighting. On those two subsets Flash reads 22.98 and 16.28 in force MAE, against 7.94 and 5.57 for Standard. All trained results are single-seed.

C Numerical Correctness Gates in Full

Table 3 gives the eight gates that Section 3.4 names. All were run with random weights under eager execution with TF32 disabled, and none was re-run on the final conservative checkpoints, so they certify the implementation rather than the shipped models.

D Cost Measurements in Full

Table 4 gives the five-system H200 campaign whose values Section 4 quotes, and Figure 5 plots them. Figure 2 is in the main text, as is Figure 2, a third campaign that shares no row with Table 4

Table 3: Numerical correctness gates. Random weights, eager execution, TF32 disabled; not re-run on the final conservative checkpoints.

Gate	Result
No-op equivalence against non-dual-stream configurations	PASS. Identical parameter signature and counts; energy and force deltas equal to those from rerunning the same code revision.
Rotation equivariance	PASS. Relative force deviation 0.002% at each tested low-resolution-stream FFN grid resolution.
Finite-difference forces	PASS on all six probe systems. Relative deviation of the finite-difference force $-\mathrm{d}E/\mathrm{d}x$ from the autograd force was 3.6×10^{-4} to 1.5×10^{-3} at a 10^{-3} Å displacement.
Angular-order creation	PASS. Perturbing the low-resolution portion of the key/value bank moved the high-resolution attention output at all three degrees; a deliberately broken variant gave exactly zero at the highest degree.
Empty high-resolution stream	PASS. Forces finite; perturbing low-resolution blocks moved them, perturbing high-resolution blocks moved them by exactly zero, which is correct when no high-resolution atoms exist.
Empty low-resolution stream	PASS. Finite energy and forces on an all-oxygen cluster.
Gradients	PASS. Non-zero gradients in high-resolution blocks, low-resolution blocks, all up- and down-projections, both edge-degree embeddings and both energy heads.
Size consistency	PASS. Relative energy error of two copies 60 Å apart versus twice one copy: 3.7e-10 mixed, 7.3e-12 high-resolution-only, 2.8e-8 low-resolution-only.

or Table 5. Table 5 is its A100-40GB counterpart and contains no H200 number. It adds the high-resolution-routed fraction and the two out-of-memory outcomes, which are preserved rather than interpolated. The two devices never share a table. The separate frozen 17-slot accuracy–throughput campaign is separately frozen under its own protocol; Table 6 reports the two production rows together with the speed-only E2Former-V2 row measured in the same campaign, and no row is shared with Table 4. The two H200 campaigns have different timing boundaries and are kept separate rather than pooled. E2Former-V2 is the comparison object of Table 6 and of the accuracy claim of Section 4; it has *no* measurement under the protocol of Table 4 and Figure 5. Those two therefore report the production configurations only. The campaigns share no row, so their values are not interchangeable and neither may be quoted without its own table.

Table 4: Single-step latency and peak allocated memory on **H200** only. The water columns are marked * because pure water is *degenerate* under the Flash routing Flash uses: zero atoms route high-resolution and the high-resolution stack is bypassed, so they carry no headline efficiency claim. Frozen weights, compiled, fp32, TF32 off; the Pareto campaign of Table 6 is separately frozen under a different timing boundary, so the two may not share rows.

Model	Water* 6000	Water* 18000	Peptide 5992	Nucleic acid 15961	Protein 32437
Single-step latency (ms)					
Standard, 4L, 6 Å	114.45	294.95	114.86	236.84	477.38
Flash, 6L, 32 ch, 5 Å	39.77	86.34	45.41	82.39	**161.18**
Peak allocated memory (GB)					
Standard, 4L, 6 Å	16.929	49.856	17.217	40.052	82.899
Flash, 6L, 32 ch, 5 Å	3.569	10.136	3.794	8.964	**19.251**

Figure 7 draws the mechanism those numbers come from and Figure 8 attributes its saving on the A100-40GB, which is the device every row of both carries; the four-system kernel scan is a separate campaign from the H200 grid above and shares no number with it. The mechanism itself is specified in Section 3.3.

Table 5: Latency and peak memory on **A100-PCIE-40GB**; no H200 number appears here. N_{HR} is the number of atoms actually routed high-resolution. Out-of-memory outcomes are reported as measured. The nucleic-acid Standard row peaked at 40.006 GB on a 40 GB device. Both water rows have $N_{\mathrm{HR}} = 0$ under Flash and are degenerate: the entire high-resolution block stack is bypassed. Frozen weights, compiled, fp32, TF32 off. The Standard and Flash rows use architecturally identical direct-force warm-start checkpoints; with frozen weights, latency and memory are weight-independent.

		Standard, 4L, 6 Å			Flash, 6L, 32 ch, 5 Å		
System	Atoms	N_{HR}	ms	GB	N_{HR}	ms	GB
Water	6000	2000	270.75	16.876	0	84.03	3.516
Water	18000	6000	OOM	OOM	0	200.08	10.079
Peptide	5992	2017	278.57	17.162	85	98.81	3.739
Nucleic acid	15961	5627	620.42	40.006	608	188.54	8.909
Protein	32437	11415	OOM	OOM	1766	394.76	19.191

Table 6: Selected locally measured rows from the frozen H200 accuracy–throughput campaign used in Figure 2; each cell is latency in ms / peak allocated memory in GB. All rows share one timing boundary that includes graph construction and the conservative force path, and the aggregate is the median of process medians. The full campaign has $17 \times 5 = 85$ model–system points at three scheduled processes per point, and five high-spread points received three additional processes, giving 270 records in total: 225 successful, 30 out of memory and 15 blocked by checkpoint access. Standard is 1.75 to 1.84× faster than E2Former-V2 while allocating less peak memory than it on every system. †E2Former-V2 is speed-only: its latency and peak memory are measured locally under this protocol on an architecturally identical checkpoint, whereas its accuracy in Figure 2 is taken from the publication; the two quantities therefore come from different weights.

	Water 6000	Water 18000	Peptide 5992	Nucleic acid 15961	Protein 32437
Standard	122.839 / 16.907	331.438 / 49.832	124.404 / 17.196	262.964 / 40.024	533.866 / 82.855
Flash	44.915 / 3.553	104.070 / 10.118	55.295 / 3.771	96.197 / 8.927	194.719 / 19.215
E2Former-V2†	214.668 / 29.448	608.483 / 87.381	221.041 / 30.099	459.515 / 64.915	949.617 / 136.258

E Checkpoint Audit and External Families

E.1 Active Checkpoint Parameter Audit

Table 7 reports optimizer-state active-parameter counts for the audited checkpoints used in the OMol25 and physical-validation analyses. A parameter is active iff its Adam/AdamW optimizer state contains an `exp_avg` tensor, which requires that it received a training gradient; the count sums those tensor elements and excludes buffers and serialized inactive parameters.

F Role-Stratified Force Error in Full

Table 9 gives the two-configuration decomposition at the exact component-weighted caliber, and Table 10 resolves it by chemical domain. Each configuration covers 2762021 structures, 283298012 atoms and 849894036 force components. Because Standard and Flash route internally, their reported splits had to be *proven* to match their own partitions rather than assumed: three per-batch guards ran on every batch of the entire split, checking that the batch carried a genuine water mask, that it matched the evaluator's independently read mask, and that it agreed with the canonical routing returned by the model; routing was derived from each checkpoint's own configuration and never passed in. Two limits on what this decomposition can argue are stated here. These are absolute per-component MAEs, unnormalised by force magnitude, so a small low-resolution-role value is not on its own evidence that the narrow stream is adequate, and the normalised version was not computed. Standard and Flash also differ at once in architecture, training stage and routing, so no difference between them may be attributed to routing alone.

Table 7: Training-active parameters for the audited mixed-resolution checkpoints used in the OMol25 and physical-validation analyses. Counts are from checkpoint optimizer state, not model-definition or serialized-tensor totals.

Checkpoint	Active params (M)
Production, accuracy and warm-start checkpoints	
Standard production, 4L, 6 Å	22.223790
Standard non-production, 6L, 5.5 Å	33.619460
Standard earlier, 6L, 6 Å	33.147457
Flash production, 6L, 32 ch, 5 Å	22.875905
Flash non-production, 6L, 32 ch, 5.5 Å	22.875905
Standard direct warm-start, 4L / 6L	22.568880 / 34.997766
Flash direct warm-start, 6L, 32 ch	23.226501
Standard earlier conservative run, 4L	22.540779
Proton-transfer PES sweep	
Standard 4L / Standard 6L warm-start	22.568880 / 34.997766
Flash 4L, 32 / 64 ch	12.346065 / 12.576657
Flash 4L, 64 ch, attention 16×4 / 4×8	12.582849 / 12.536337
Flash 4L, 64 ch, attention 8×8	12.578721
Flash 6L, 64 ch, attention 16×4	23.155649
Water-wall and low-resolution-capacity checkpoints	
Standard replicate	22.540779
Flash, 64 scalar + 64 vector	22.301834
Flash, 128 scalar + 128 vector	22.560906
Flash, 64 scalar + 64 vector, two tails	22.513618

Table 8: More recent external families. These are on the *same* validation split as Table 1: their sources report the OMol25 composition validation split, which carries the same four domains and total as ours, and the two baselines these sources share with Huang et al. (2026) agree digit-for-digit after unit conversion (eSEN-sm-d 9.92 meV/Å $\leftrightarrow$ 0.23 kcal/mol/Å; eSEN-sm-cons 8.25 $\leftrightarrow$ 0.19). Force MAE is transcribed exactly at 43.3641 meV/Å per kcal/mol/Å. Energy is not transcribed because no source states an energy-referencing convention, and none states whether its force metric is per-component or per-vector, so the metric caveat still applies. Locally measured H200 throughput for these models is plotted in Figure 2; pairing it with these published accuracies still combines two measurement environments, which the figure marks as a separate class.

Family	Nano	Mini	Neo	Air	Plus
		force MAE (meV/Å)			
DPA4 (Li et al., 2026a)	33.65	21.77	14.40	8.20	6.33
DPA4C (Li et al., 2026b)	107.75	81.58	75.06	69.25	59.97
UMA-S-1.1, as re-transcribed in Li et al. (2026a): 9.54 meV/Å					

The two configurations apply different strengths of HR/LR capacity reduction: Standard changes participation only, whereas Flash changes both participation and persistent width and assigns a larger fraction of the system to the low-resolution role. Their HR/LR error ratios nevertheless remain close, 3.634 against 3.324. This consistency suggests that the imbalance more likely follows the chemistry of the two roles than the precise capacity placement. It is not a causal ablation: the configurations differ simultaneously in architecture, training stage and routing.

G Extended Physical Validation

Table 11 and Figure 11 give the full measured record of the coupled-coordinate proton-transfer gate. Section 4 retains only the short-range repulsion result on the production checkpoints and points here for the complete physical-gate record. Table 12 and Figure 12 report the oxygen–oxygen repulsion scans. Tables 13, 14, 15 and 16, together with Figures 13, 14 and 15, give the liquid-structure and dynamics evidence. Tables 17, 18 and 19 give the temperature, chemical-integrity and energy-conservation evidence for the 54-leg matrix. Table 20 gives the separate production 1 ns NVT campaign that extends the duration question past this matrix's 20 ps on one system. Two statements are repeated here because they bound what these tables can be used for, and they are *not* the same statement. The 64-molecule Standard leg is a protocol defect rather than a model defect, its interaction

Table 9: Role-stratified per-component force MAE (meV/Å) on the full 2762021-sample split at the *exact component-weighted* caliber. Each row uses the router bound to that configuration; HR and LR denote its high- and low-resolution roles. Every ratio is the quotient of the two preceding columns in its own row. Single-seed.

Model	HR role	LR role	Overall	HR / LR	HR share of components
Exact component-weighted caliber					
E2Former-V3-Standard, 4L, 6 Å	12.839	3.533	8.146	3.634	49.57%
E2Former-V3-Flash, 6L, 32 ch, 5 Å	17.495	5.264	11.272	3.324	49.12%

Table 10: The same evaluation resolved by chemical domain, force-component MAE in meV/Å, formatted *high-resolution / low-resolution / all*, with each configuration evaluated under its bound router. Domain atom counts are 91063984 (32.1%), 171462133 (60.5%), 13709053 (4.8%) and 7062842 (2.5%). Exact caliber, single-seed. The last column provides the published all-atom comparison: E2Former-V2's four domain values and its total, transcribed from Table 2 of Huang et al. (2026), which reports the same four domains and total on the same OMol25 composition validation split, in the source's own printed *Total* box (no per-subset high-resolution/low-resolution split is published, so this column has no such cell). It is *not* on our caliber: the source's force metric convention is unstated, whereas our three columns are component-weighted, so it is read as a level and not as a per-domain tie-break, exactly as Table 1 reads its total. Transcribed, not re-measured: E2Former-V2 has no checkpoint that is accuracy-valid for the role-stratified evaluator, so it could not be re-scored here even with unmatched weights.

Domain	Standard	Flash	E2Former-V2 (published all)
Biomolecules	7.478 / 2.716 / 5.104	11.075 / 4.353 / 7.685	— / — / 4.96
Electrolytes	12.595 / 3.635 / 8.002	17.831 / 5.476 / 11.446	— / — / 10.11
Metals	41.764 / 6.825 / 25.555	46.121 / 7.950 / 28.396	— / — / 24.40
Organics	26.485 / 5.759 / 17.072	30.971 / 7.099 / 20.065	— / — / 19.69
Total	12.839 / 3.533 / 8.146	17.495 / 5.264 / 11.272	— / — / 8.38

cutoff filling 97% of that box's half-width: that makes the leg uninformative on the thermodynamic axes, not invalid, and its structure, integrity and energy-conservation evidence still stands. The Flash pure-water column is a *degenerate limit* of the routing rather than a failed measurement — pure water has no high-resolution atoms under it, so only the narrow stream executes — which is why it cannot evidence cross-role exchange or carry an efficiency headline, and not a reason to discard what it does show. The body states the degenerate limit in its experimental setup (Section 4); the 64-molecule protocol defect is stated here only, with its evidence beside it.

G.1 Temperature, Chemical Integrity and Energy Conservation

The main text claims that the production trajectories preserve chemical structure, hold temperature stable and match experimental liquid structure. This subsection gives the first two of those in full, on the same 54 legs, and states where the evidence stops.

Temperature. Table 17 reports the measured kinetic temperature over the 5 to 20 ps window, target 300 K, timestep 0.5 fs. The first 5 ps are excluded because they carry the start-up transient from the initial velocity assignment, which dominates the block means over that stretch. Over the remaining window the 54 leg means average 299.1 K; every one of the 18 system-model-ensemble cells sits within 3.4 K of the target, and 17 of 18 within 2.4 K. There is no secular drift: per leg the second half minus the first half stays within 7.1 K, and the largest fitted slope magnitude of any leg is 0.78 K/ps; both are dominated by the shortest, most per-atom-fluctuating legs.

The six constant-energy cells — eighteen of the 54 legs — are the exception, and the table must be read with the protocol in mind. Because an NVE leg is continued from the completed matching NVT leg rather than re-thermalized, it inherits that endpoint's instantaneous kinetic energy as its conserved total energy, so an NVE leg's mean temperature stays wherever its parent NVT leg happened to end. That is why the two water NVE cells, the smallest box at 216 molecules and therefore the largest per-

Table 11: Coupled-coordinate proton-transfer barrier gate (Agmon, 1995), kcal/mol on a rigid 234-geometry slice zeroed at the uncompressed reactant. The stated acceptance threshold is approximately 9.8 kcal/mol; the applied boundary fell between 9.65 and 12.21 and is therefore not a sharp cut, and the two routings describe different surfaces, so only the ordering within a routing is meaningful. The last row stays *unfilled for this scan*: the production Flash configuration was never run on this 234-geometry slice and no other checkpoint is substituted for it. It is not unmeasured — it was scanned on the separate onset-frame surface of the foreground panel of Figure 11, where it passes — but the two slices are different surfaces and may not share a row. The two Standard rows used architecturally identical warm-start checkpoints rather than the production conservative ones. Single-seed.

Checkpoint	Routing	Layers	Step	Coupled barrier	Compression + barrier	Gate
Standard, 6 Å	Standard	4	740 000	**40.33**	4.4 + 36.0	pass
Standard, 5.5 Å	Standard	6	550 000	**35.25**	5.2 + 30.1	pass
Flash, 64 ch, 16×4 attention	Flash	4	200 000	14.70	12.3 + 2.4	pass
Flash, 64 ch, 16×4 attention	Flash	6	25 000	13.21	11.8 + 1.4	pass
Flash, 64 ch	Flash	4	200 000	12.21	11.1 + 1.2	pass
Flash, 64 ch, 8×8 attention	Flash	4	200 000	9.65	8.0 + 1.6	**FAIL**
Flash, 32 ch	Flash	4	200 000	9.48	5.3 + 4.1	**FAIL**
Flash, 64 ch, 4×8 attention	Flash	4	200 000	6.57	5.0 + 1.6	**FAIL**
Flash, 32 ch (production)	Flash	6	—	—	—	not run on this slice

Table 12: Water-dimer oxygen-oxygen repulsive wall at 1.80 Å, against a high-level DFT reference of 140.38 kcal/mol. The production block is the paper's two reported checkpoints, scanned 2026-09-23 on the same 44-point grid and energy convention as the sweep below it, so the two blocks are directly comparable; both checkpoints' SHA-256 values were re-verified and the scan is a deterministic single-point evaluation of frozen weights. In that block the routing column names the oxygen routing actually applied, and only the routing a model was trained under is that model's own surface — the scan's oxygen-LR flag is not gated on the checkpoint's training config, so Standard's LR row is a routing those weights never saw. All rows of the *sweep* block are trained 200k-step checkpoints of a 4M-sample subset, *not* production models; there the two nominally equivalent Standard runs differ by 1.27 kcal/mol on the LR stream and 0.56 kcal/mol on the HR stream. The sweep's 64-scalar Flash row is the only checkpoint measured that softened its *high-resolution* stream.

Checkpoint	Oxygen routing	LR-stream E (kcal/mol)	% of reference	HR-stream E (kcal/mol)	% of reference
Production checkpoints					
E2Former-V3-Standard, 4L, 6 Å (Standard)	LR	103.02	73.4	140.58	100.1
E2Former-V3-Flash, 6L, 32 ch, 5 Å (Flash)	LR	69.81	**49.7**	139.22	99.2
200k-step configuration sweep (4M-sample subset), not production					
Full-resolution uniform, no routing	—	139.16	99.1	139.16	99.1
Standard, run A	LR	32.30	23.0	140.47	100.1
Standard, run B	LR	33.57	23.9	139.91	99.7
Flash, 64 scalar	LR	58.57	41.7	**128.62**	**91.6**
Flash, 64 scalar + 64 vector	LR	54.16	38.6	139.41	99.3
Flash, 128 scalar + 128 vector	LR	37.06	26.4	138.12	98.4
Flash, 64 scalar, two tail blocks	LR	45.76	32.6	138.12	98.4

atom fluctuations, span 290.83 to 310.33 K, while every NVE cell on the larger NaCl and DEKSAN boxes lies within 3.2 K of target. In none of them does the temperature walk: the offset is set at the start and held. The honest statement is therefore that the thermostatted ensembles hold 300 K and the constant-energy legs conserve energy at whatever temperature they started from, not that all 54 legs sit at 300 K.

Chemical integrity. Table 18 reports the covalent-bond audit. Only the DEKSAN peptide carries it: the water and NaCl boxes are held together by the same inferred covalent graph, but the audit's per-leg record stores the solute bond statistics only for the peptide, so for those two systems chemical integrity rests on the weaker zero-overlap and zero-fusion evidence below rather than on a bond-length measurement. Across the 18 peptide legs the inferred graph is 196 bonds in every single leg, matching the reference exactly, and the worst stretch of any bond in any frame of any leg is a ratio of 1.17548 against a gate limit of 1.5. The peptide stays folded, with mean RMSD 0.91 to 1.20 Å and mean radius of gyration 5.89 to 6.02 Å in all 18 legs.

Table 13: Liquid-water structure over the twelve measured 20 ps *water-only* legs (two box sizes × two routings × three ensembles; the precedence for the paper's liquid-structure evidence is Table 14), against the X-ray benchmark. This is not the 54-leg matrix: it predates it and covers water alone. Cells marked "not reported" are an outstanding analysis gap: they are not zero and no value is substituted for them. Peak positions agree to within 0.05 Å while peak heights sit 11 to 43% above experiment, consistent with the amplitudes panel of Figure 13.

Structural quantity	Flash	Standard	Experiment
First O-O peak position	2.82 Å	not reported	2.80 Å
First O-O peak height	2.70	2.77	2.57
Second shell position	4.45 Å	4.61 Å	4.50 Å
First minimum	not reported	3.44 Å	3.45 Å
First-shell coordination number	4.78	4.97	4.3 to 4.7

Table 14: First-shell structure on the **54-leg production matrix**, from the stored curves rather than from tabulated scalars, and only on the NVT legs, which are the ones for which radial distribution functions were recorded. This is the table the paper's liquid-structure claim rests on; Table 13 is the water-only precedence. Peak values are read off the three-seed mean curve and first-shell coordination is the committed cumulative count integrated over the range the comparable experiment states, so it need not reproduce the per-leg-peak convention of an earlier summary. Salt O–O is included because the salt box carries water; Na–O and Cl–O are the ion-hydration shells. All curves are three seeds × 50 frames, $t = 10$ to 20 ps, single trajectory per seed.
Coord. is the committed cumulative count integrated over **the range the comparable experiment states**, so it can be read against a published integral; on the narrower integration to each curve's own first minimum the same curves give 5.07/4.97 (water O–O), 5.21/4.78 (salt O–O), 5.30/5.30 (Na–O) and 6.66/7.03 (Cl–O), which are not comparable to a published number and are not used for comparison here.

Curve	Model	Peak r (Å)	Peak g	First min. r (Å)	Coord.
Water O–O	Standard	2.788	2.786	3.438	4.44
Water O–O	Flash	2.838	2.727	3.413	4.40
Salt O–O	Standard	2.812	2.749	3.513	4.32
Salt O–O	Flash	2.812	2.753	3.413	4.28
Salt Na–O	Standard	2.362	6.263	3.138	5.52
Salt Na–O	Flash	2.312	5.945	3.138	5.34
Salt Cl–O	Standard	3.163	3.197	3.888	6.96
Salt Cl–O	Flash	3.213	2.838	3.987	6.38

The weaker evidence does cover all 54 legs: no leg recorded any anomaly, and specifically none recorded an oxygen-oxygen contact below 0.65 Å or an oxygen-fusion event below 1.8 Å. That is absence of atomic overlap, not bond preservation, and the distinction matters for the water-only boxes, where no intramolecular coordinate was tracked at all.

Energy conservation. Table 19 gives the constant-energy drift per leg. Flash sits below Standard on all three systems, but by an order of magnitude only on the NaCl box. On water and DEKSAN the Flash drift is indistinguishable from zero and changes sign between replicas (-0.00101, -0.00020, $+0.00016$ and -0.00012, -0.00011, $+0.00058$ meV/atom/ps), so those two ratios are ratios of noise and the ranking there is not resolvable at this caliber. The drift is small in absolute terms either way, at most $+0.0117$ meV/atom/ps.

G.2 Duration Beyond the Matrix: The Production 1 ns NVT Campaign

Duration beyond 20 ps. The matrix above stops at 40000 steps by design, and its 20 ps window is shorter than every failure this work has recorded elsewhere. A separate campaign ran the two *production* checkpoints on the 32437-atom solvated protein for 2000000 steps at 0.5 fs = 1000 ps, in NVT at 300 K, one seed, cells fixed (Table 20). It answers the duration question for that one system and nothing else. Its bounds are load-bearing: one system, one ensemble, one seed; fixed-cell NVT makes density constant, so no density, radial distribution function or thermodynamic quan-

Table 15: Constant-pressure legs. Densities are 3.6 to 5.9% too high and compressibilities roughly threefold too stiff. The 64-molecule Standard leg is a protocol defect rather than a model defect: its interaction cutoff filled 97% of that box's half-width, and its compressibility is twelvefold inconsistent with its own larger box. The experimental row is the IAPWS-95 reference at 298 K and 1 bar (Wagner & Pruß, 2002; Pi et al., 2009).

Leg	Density (g/cm^3)	Deviation	Compressibility (1e-5 /bar)	Acceptance
512 molecules, Flash	1.0560 ± 0.0016	+5.9%	1.45 ± 0.11	0.83
512 molecules, Standard	1.0332 ± 0.0015	+3.6%	1.35 ± 0.13	0.83
64 molecules, Flash	1.0403 ± 0.0063	—	2.24 ± 0.35	0.86
64 molecules, Standard	0.9718 ± 0.0209	—	$\mathbf{16.41 \pm 0.45}$	0.86
Experiment[1]	0.997	—	4.52	—

Table 16: Constant-energy legs on the twelve water-only legs. The Flash rows are the routing's *degenerate limit*: pure water has no high-resolution atoms under it, so only the narrow scalar stream executed. They are a valid measurement of that stream and are not evidence about cross-role exchange. The 64-molecule Standard leg is the protocol defect described above, which affects its thermodynamic readings rather than its conservation.

Leg	Final drift (meV/atom)	RMS	Level shift, first to last 4 ps	Shift / RMS
512 molecules, Flash	−0.013	0.008	+0.001	0.12
64 molecules, Flash	+0.043	0.023	−0.002	−0.10
512 molecules, Standard	+0.041	0.021	**+0.041**	**1.95**
64 molecules, Standard	+0.130	0.093	**+0.189**	**2.03**

tity follows from it; and it does not retire the low-resolution-capacity failure onsets at 40, 58, 60 and 511.9 ps, which are different systems and different configurations. It shows that the production checkpoints can carry this system past the horizon at which every failure so far was observed, without a structural failure.

Classical control. The same equilibrated starting state was also propagated for 1000 ps in NVT with an AMBER ff14SB/TIP3P GROMACS control. Its Cα displacement from its own start is 1.0799 Å; Flash's is 1.964 Å, and the Flash and classical 1000 ps endpoints are separated by 1.8273 Å after a common Cα alignment. The classical protein's mass-weighted R_g, evaluated by `gmx gyrate` over the Protein group, spans 16.881–17.268 Å and ends at 17.043 Å. Figure 3 reports both full classical displacement time series alongside Standard and Flash in panel (e). Panel (f) instead reports Cα RMSD to the CASP15 T1123-D1 experimental structure, using the 214 resolved Cα atoms shared by the target and CASP15 AlphaFold2 submission T1123TS270_1; that AF2 submission is shown as a horizontal reference at 5.102 Å. Over 1 ns, the final ground-truth RMSDs are 5.064 Å (Standard), 4.809 Å (Flash), and 4.889 Å (classical). The comparison covers one system, one ensemble and one seed, so these curves do not establish a general accuracy ranking.

H Low-Resolution-Capacity Variants and Warm-Start Accuracy

Table 21 carries three retired non-production configurations evaluated on the full split at the training-time batch-mean caliber; they were not re-scored at the exact component-weighted caliber.

Table 22 is the evidence behind the statement in Appendix I that averaged error does not rank dynamical stability: the spreads are 0.4% on validation force, 0.4% on held-out force, 1.2% on validation energy and 1.8% on the held-out energy metric, while the dynamics outcomes are not comparable at all. Table 23 gives the barrier profile of the same variants and controls on one rigid proton-transfer coordinate; for the Flash rows, pure water is the all-low-resolution routing limit, for which there is no absolute reference, so only the ordering across checkpoints is meaningful; a barrier of 0.0 means the two basins merged rather than that the measurement failed. Table 24 records the direct-force warm-start checkpoints, which are validated on the 1M subsample and are therefore not comparable to the full-split rows of Table 1.

Table 17: Kinetic temperature over the 5 to 20 ps window, target 300 K, over the 54-leg production matrix. "Mean" is the mean of the three leg means, "SEM" its three-replica standard error, "Leg range" the span of the three leg means, and "Frame range" the extremes reached by any saved frame. The first 5 ps are excluded as start-up transient. The NVE rows carry a protocol caveat rather than a modelling one: each NVE leg is continued from its matching NVT endpoint and conserves the energy it inherits, so its mean temperature reflects where that endpoint stopped and is not thermostatted. All values measured, none extrapolated.

System	Model	Ensemble	Mean (K)	SEM	Leg range (K)	Frame range (K)
water	Standard	NVT	300.06	1.34	298.18–302.66	274.00–333.31
water	Standard	NPT	297.68	0.36	297.02–298.26	270.52–320.71
water	Standard	NVE	299.47	4.32	290.83–303.84	270.49–325.17
water	Flash	NVT	298.24	0.52	297.20–298.82	272.51–325.03
water	Flash	NPT	299.83	0.48	299.28–300.79	269.62–326.18
water	Flash	NVE	303.34	3.59	298.38–310.33	281.49–329.67
NaCl	Standard	NVT	298.54	0.29	298.00–298.99	292.38–304.31
NaCl	Standard	NPT	298.75	0.26	298.23–299.02	291.89–305.28
NaCl	Standard	NVE	298.60	0.93	296.93–300.16	292.96–304.98
NaCl	Flash	NVT	298.99	0.14	298.72–299.20	292.81–305.44
NaCl	Flash	NPT	299.28	0.14	299.10–299.57	292.54–309.01
NaCl	Flash	NVE	299.18	0.47	298.33–299.97	294.34–304.11
DEKSAN	Standard	NVT	298.40	0.21	298.08–298.81	288.90–307.34
DEKSAN	Standard	NPT	298.94	0.04	298.86–299.00	289.17–307.67
DEKSAN	Standard	NVE	299.26	0.76	297.97–300.59	291.30–307.00
DEKSAN	Flash	NVT	298.84	0.57	298.22–299.97	288.24–313.28
DEKSAN	Flash	NPT	298.44	0.73	297.38–299.85	288.04–307.55
DEKSAN	Flash	NVE	298.26	1.07	297.09–300.40	288.68–306.44
All 54 legs			299.12	—	290.83–310.33	—

Table 18: DEKSAN covalent-bond audit over the 18 peptide legs, gate limit 1.5 on the stretch ratio. The bond count is inferred from covalent radii and is 196 in every leg, matching the reference graph. "Max stretch" is the worst ratio over all bonds and all frames of all legs of that model, and "Max bond" the longest absolute bond. RMSD and R_g are leg means, so the ranges are across legs rather than across frames. No value is reported for water or NaCl because the per-leg record stores solute bond statistics for the peptide only.

Model	Covalent bonds	Max stretch	Max bond (Å)	RMSD mean (Å)	R_g mean (Å)
Standard	196	1.17148	1.6819	0.931–1.198	5.888–5.998
Flash	196	1.17548	1.7116	0.913–1.203	5.912–6.024
Reference	196	1.00000	—	—	—

A separate ensemble-and-box-size scan of one model recorded failure onsets from 16.2 to 189.0 ps, an 11.7-fold range. Those runs are not the row-wise capacity comparison in Table 22. Likewise, the 680 ps capacity-sweep trajectory and the separately located 511.9 ps onset below are distinct trajectories and are not competing timestamps for one run.

I Boundary Conditions and Outstanding Measurements

I.1 Bounds That Travel With the Frontier Result

The force claim concerns per-component MAE. It is not a claim of comparable accuracy across all metrics. The energy MAE, 2.459 meV/atom for Standard and 4.474 for Flash, is above the published uniform-capacity conservative figure of 1.16. The energy column is printed beside the force column at equal prominence and is not argued in our favour (Table 1). Every trained result is single-seed,

Table 19: Constant-energy drift in meV/atom/ps, signed, one row per leg. Negative values indicate the energy falling. The three-replica means are $+0.00612$, $+0.00900$ and $+0.00535$ for Standard water, NaCl and DEKSAN, against -0.00035, $+0.00081$ and $+0.00011$ for Flash. Measured over an 18 ps window (t = 2 to 20 ps), so a drift this small bounds the short-time behaviour only and is not a long-time conservation claim.

System	Model	Seed 104729	Seed 130363	Seed 155921	Mean
water	Standard	$+0.00545$	$+0.00529$	$+0.00761$	$+0.00612$
water	Flash	-0.00101	-0.00020	$+0.00016$	-0.00035
NaCl	Standard	$+0.01174$	$+0.00767$	$+0.00761$	$+0.00900$
NaCl	Flash	$+0.00129$	$+0.00065$	$+0.00050$	$+0.00081$
DEKSAN	Standard	$+0.00483$	$+0.00647$	$+0.00475$	$+0.00535$
DEKSAN	Flash	-0.00012	-0.00011	$+0.00058$	$+0.00011$

Table 20: The completed production 1 ns NVT campaign on the 32437-atom solvated protein (one system, one ensemble, one seed). Both legs reached 2000000 steps / 1000 ps. The structural pass exists because the runner's own scalar log can read healthy while water dissociates; all quantities are PBC-aware with intra-residue covalent contacts excluded. Uncertainties are not reported: one seed cannot give them.

	Standard	Flash
Status	**complete, 1000 ps**	**complete, 1000 ps**
Non-finite values	none	none
Cα RMSD from start, final / max (Å)	2.192 / 2.326	1.964 / 2.136
Radius of gyration, final / range (Å)	17.340 / 16.71–17.47	16.120 / 16.06–17.06
Min. nonbonded distance over run (Å)	1.242	1.239
Contacts below 1.2 Å	0	0
Water O–H, mean / max (Å)	0.972 / 1.143	0.973 / 1.142
Endpoint total-energy change / 1000 ps (meV/atom/ps)	-0.00938	-0.00994
Final temperature (K)	298.8	298.5

Final readout: 2026-09-25 15:59 UTC. Source: experimental log Section 2.23.

so no statistical claim attaches to any accuracy number. Timing, by contrast, was repeated across processes.

The two configurations should be read as operating points. They differ at once in depth, cutoff, routing, persistent-state layout and attention implementation, so their comparison is not a single-variable ablation. No state-layout variant was measured for accuracy. That axis is therefore unmeasured, and the existing results do not separate routing from the remaining axes. No uniform-capacity model is retained in the role-stratified comparison, so the chemistry interpretation rests on consistency across the two routed configurations rather than a no-routing control.

The numerical gates were run with random weights and were not re-run on the final conservative checkpoints. They certify the implementation rather than the shipped models (Appendix C). The role split is an absolute per-component MAE unnormalised by force magnitude. A small low-resolution-role value therefore does not on its own establish that the narrow stream is adequate (Appendix F). The external families of Table 8 do not share our measurement environment: their accuracy is published, their latency was measured here, and their metric convention is unstated.

I.2 The Dynamics Evidence Has a Horizon and a Thermodynamic Bias

The 20 ps matrix on three systems is shorter than the water-box failure onsets this work recorded elsewhere: near 40 ps, 58 ps and 60 ps, one located onset at 511.9 ps, and one at 680 ps. Completion of all 54 legs therefore supports short-time stability and local structure over the measured 20 ps window; it does not establish longer-term stability. The separate 1 ns campaign covers one system, one ensemble and one seed and does not transfer to the other three (Section G.2).

Table 21: Retired non-production configurations of this work, on the full 2762021-sample validation split at the training-time batch-mean caliber. Force MAE is per-component and energy MAE is at equal prominence. Both Flash configurations are worse on force than every Standard configuration. Single-seed.

Model	Routing	Force MAE (meV/Å)	Energy MAE (meV/atom)	Val. split	Force mode	Provenance
E2Former-V3-Standard, 6L, 5.5 Å	Standard	7.200	1.999	full	conserv.	this work
E2Former-V3-Standard, 6L, 6 Å	Standard	7.651	1.475	full	conserv.	this work
E2Former-V3-Flash, 6L, 32 ch, 5.5 Å	Flash	10.506	3.921	full	conserv.	this work

Table 22: Low-resolution-capacity variants: indistinguishable averaged error, divergent dynamics. Trained 200k-step checkpoints. Both E/F columns report energy MAE in meV/atom and per-component force MAE in meV/Å. Cells marked "not measured" are not zero and no value is substituted. The variant with the lowest held-out energy and force MAEs nevertheless developed a frozen ion pair from 60 ps.

Variant	Validation E / F	Held-out E / F	Dynamics outcome
64 scalar + 64 vector low-resolution	13.91 / 18.91	0.406 / 8.354	clean 1 ns; clean 200 ps on the molecular set
128 scalar + 128 vector low-resolution	14.08 / 18.93	0.401 / 8.382	fused water pairs from 40 ps, 968 of 1000 pairs fused, minimum O-O 2.033 Å; diverged at 2.5 to 3.0 ps on the molecular set
64 scalar + 64 vector, two tail blocks	14.05 / 18.99	**0.399 / 8.346**	frozen ion pair from 60 ps; clean 200 ps on the molecular set
64 scalar low-resolution (Flash reference)	not measured	not measured	fused at 680 ps with two-proton defects
Full-resolution and Standard controls	not measured	not measured	clean

Averaged error does not rank these dynamics outcomes. Three low-resolution-capacity variants sat within 0.4% of one another on validation force, 0.4% on held-out force, 1.2% on validation energy and 1.8% on held-out energy. Their dynamics ranged from a clean nanosecond to 968 of 1000 water pairs fused from 40 ps at a minimum oxygen-oxygen distance of 2.033 Å. The variant with the *best* held-out energy had the worst dynamics. A separate scan across ensembles and box sizes placed one variant's onset between 16.2 and 189.0 ps, an 11.7-fold range (Appendix H). The constant-pressure densities sit near 1.07 g/cm^3 against an experimental 0.997 and their compressibilities are roughly threefold too stiff, so a correct radial distribution function is not by itself evidence of a usable potential. Neighbor-union filtering has maximum absolute energy and force-component differences of 2.2888×10^{-5} eV and 6.3676×10^{-4} eV/Å, respectively, below the frozen 1×10^{-3} thresholds in their respective units across all ten model–system pairs (Section 4).

I.3 Outstanding Measurements and Negative Claims

Within the paper's scope — aqueous explicit-solvent biomolecular systems — several measurements remain outstanding. These are hydrogen-pair observables, ion residence and exchange rates, the $\geq$ 200 ps screening tier and the 1 ns tier of the 54-leg matrix. The study also has neither a recipe-matched full-resolution control nor a trained same-low-resolution-fraction random-routing control. The claim that atom-role-aware allocation beats a chemistry-agnostic allocation of the same low-resolution fraction therefore remains an explicit *conjecture* (Section 2). We claim no reduced angular resolution, no learned or dynamic routing, no optimal partition, no zero-shot transfer of the routing policy to another solvent, and no demonstrated validity in reactive regimes. Beyond potentials, whether nodes in structured scientific data need uniform computation at all remains open.

Table 23: Barrier profile of the same variants and controls on one rigid proton-transfer coordinate, in kcal/mol. The Flash rows are evaluated in pure water, its all-low-resolution routing limit; LR denotes the low-resolution role. There is no absolute reference for this coordinate, so only the ordering across checkpoints is meaningful, and a barrier of 0.0 means the two basins merged rather than that the measurement failed.

Checkpoint	Routing	At 2.80 Å	At 2.31 Å	Best coupled	Observed MD outcome
64 scalar + 64 vector	Flash limit	23.9	0.1	10.5	survived 1 ns
64 scalar	Flash limit	26.2	4.6	9.8	failed at 511.9 ps
64 scalar, two tail	Flash limit	24.2	2.5	9.7	not run
128 scalar + 128 vector	Flash limit	27.5	**0.0**	7.2	failed at 40 ps
Standard, 4 layers	Standard	48.6	26.3	36.0	never failed
Full-resolution uniform	uniform	46.0	25.9	34.5	never failed

Table 24: Direct-force warm-start checkpoints, validated on the **1M subsample** and therefore not comparable with the full-split rows of Table 1; the subsample reads about 6.5% optimistic. Force mode is direct except where stated, and direct-force rows must not be read as conservative-force accuracy. The last two rows are the published direct-force variants of the baselines in Table 1, on their own unstated split. Single-seed.

Model	Routing	Force MAE (meV/Å)	Energy MAE (meV/atom)	Force mode
Standard, 4L, 6 Å	Standard	10.310	1.390	direct
Standard, 6L, 5.5 Å	Standard	8.668	1.250	direct
Flash, 6L, 32 ch, 5 Å	Flash	12.808	1.783	direct
Standard, 4L, 6 Å (earlier run)	Standard	8.147	1.200	conservative
E2Former-V2 (Huang et al., 2026)	uniform	9.55	1.24	direct
eSEN (Fu et al., 2025)	uniform	9.92	1.49	direct

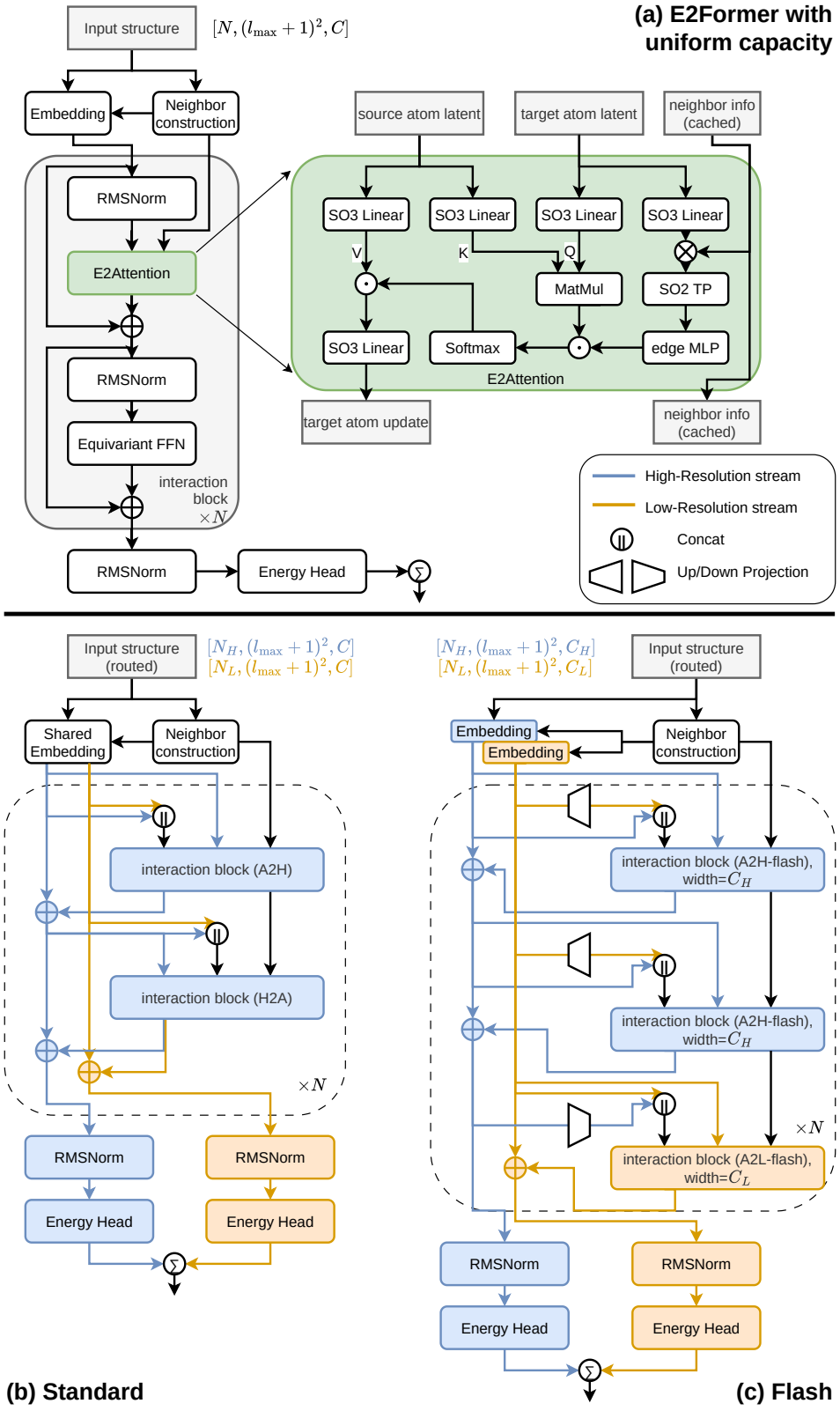


Figure 4: **The same chain under three capacity allocations.** **(a)** Uniform capacity: E2Attention (Huang et al., 2026) at the interaction block's core. **(b)** Interleaved, used by Standard, allocates the update-count axis; **(c)** interleaved_flash, used by Flash, allocates both axes.

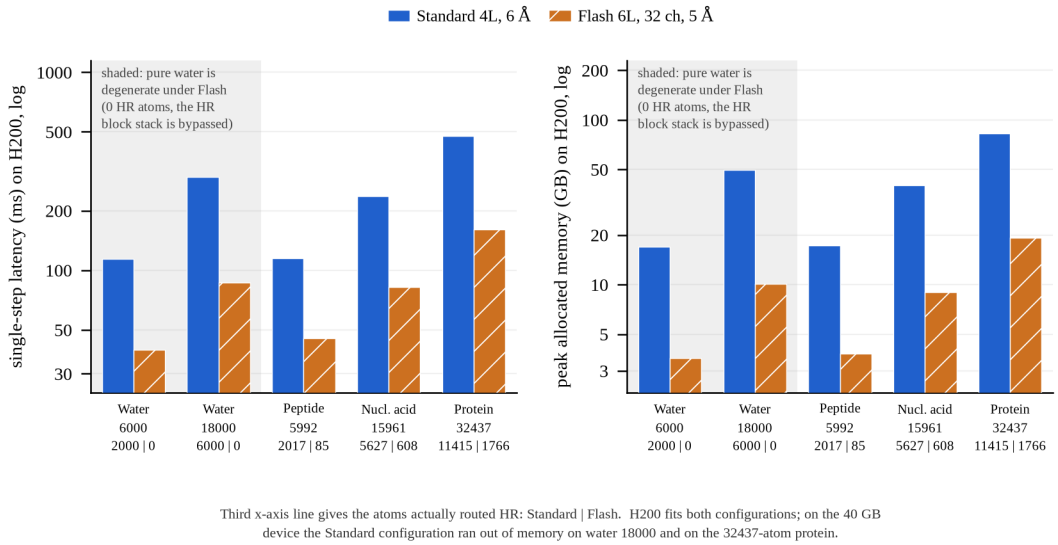
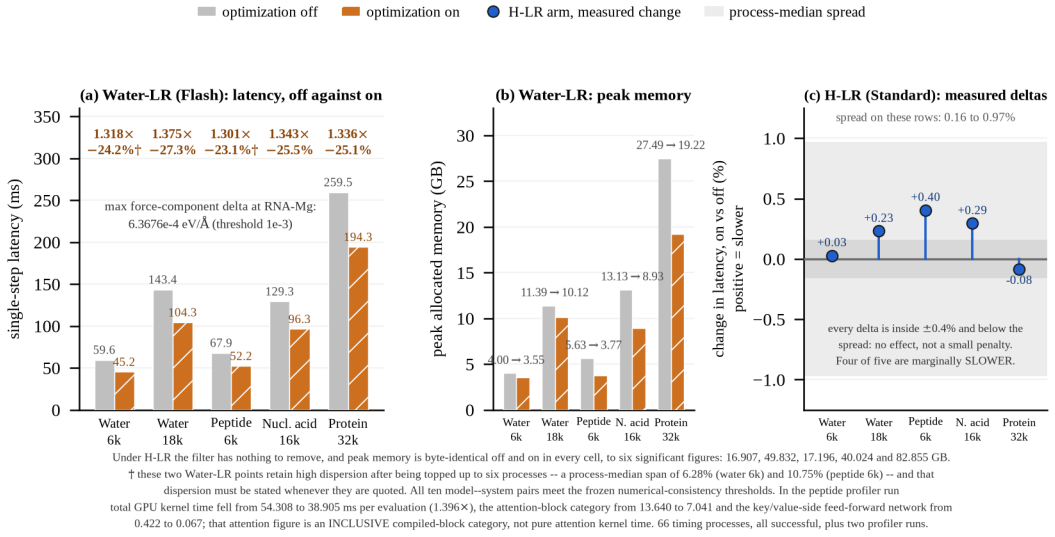


Figure 5: Single-step latency and peak allocated memory across five benchmark systems, H200 only, on log-scaled axes, for the two production configurations. On the 32437-atom solvated protein Flash runs at 161.18 ms and 19.251 GB against Standard at 477.38 ms and 82.899 GB. The third x-axis line gives the atoms actually routed high-resolution under each routing. The two shaded pure-water groups are degenerate under Flash: zero atoms route high-resolution, so the high-resolution block stack is bypassed and only the narrow stream executes; those groups carry no headline efficiency claim. The systems that exercise both streams are the peptide with 85 high-resolution atoms, the nucleic acid with 608 and the protein with 1766. Every configuration shown fits the H200; on the 40 GB device the Standard configuration ran out of memory on water 18000 and on the 32437-atom protein, and those outcomes are stated rather than interpolated.



Under H-LR the filter has nothing to remove, and peak memory is byte-identical off and on in every cell, to six significant figures: 16.907, 49.832, 17.196, 40.024 and 82.855 GB.
† these two Water-LR points retain high dispersion after being topped up to six processes -- a process-median span of 6.28% (water 6k) and 10.75% (peptide 6k) -- and that dispersion must be stated whenever they are quoted. All ten model--system pairs meet the frozen numerical-consistency thresholds. In the peptide profiler run total GPU kernel time fell from 54.308 to 38.905 ms per evaluation (1.396×), the attention-block category from 13.640 to 7.041 and the key/value-side feed-forward network from 0.422 to 0.067; that attention figure is an INCLUSIVE compiled-block category, not pure attention kernel time. 66 timing processes, all successful, plus two profiler runs.

Figure 6: Inference-time neighbor-union filtering, optimization off against on, over 66 timing processes and two profiler runs. The Flash configuration gains 1.30 to 1.37× across all five systems with peak memory falling in every case; the Standard arm gains nothing, its five deltas lying inside $\pm 0.4\%$ — four of the five marginally *slower* — against a 0.16 to 0.97% process-median spread, with peak memory unchanged to six significant figures. Two points retain high dispersion after being topped up to six processes (6.28% and 10.75% process-median span) and that dispersion travels with them. Across all ten model–system pairs, the maximum absolute energy difference is 2.2888×10^{-5} eV and the maximum force-component difference is 6.3676×10^{-4} eV/Å, both below the frozen 1×10^{-3} thresholds in their respective units. The profiler's attention-block figure is an inclusive compiled-block category, not pure attention kernel time.

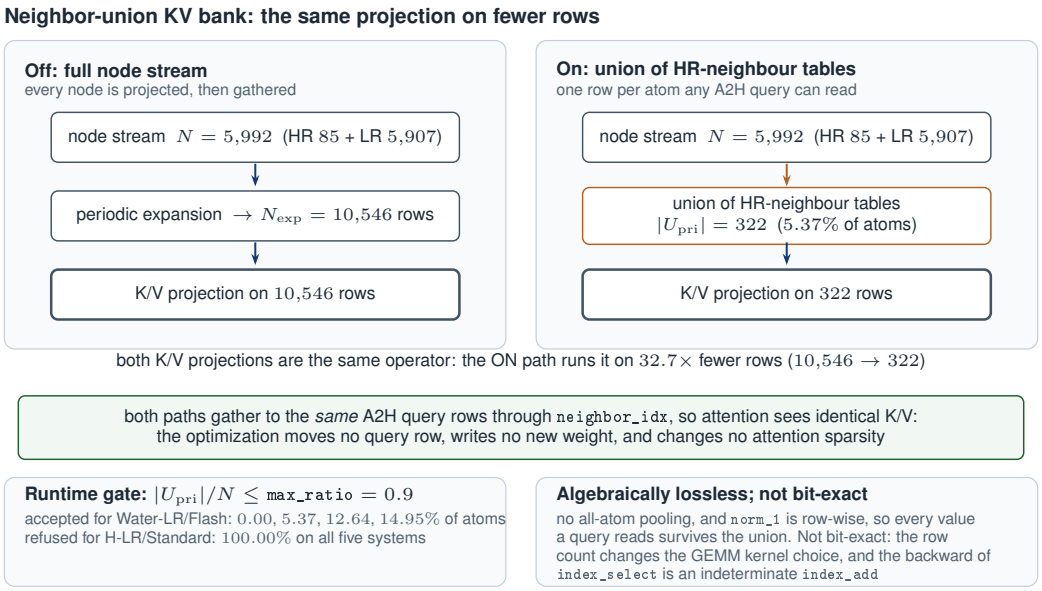
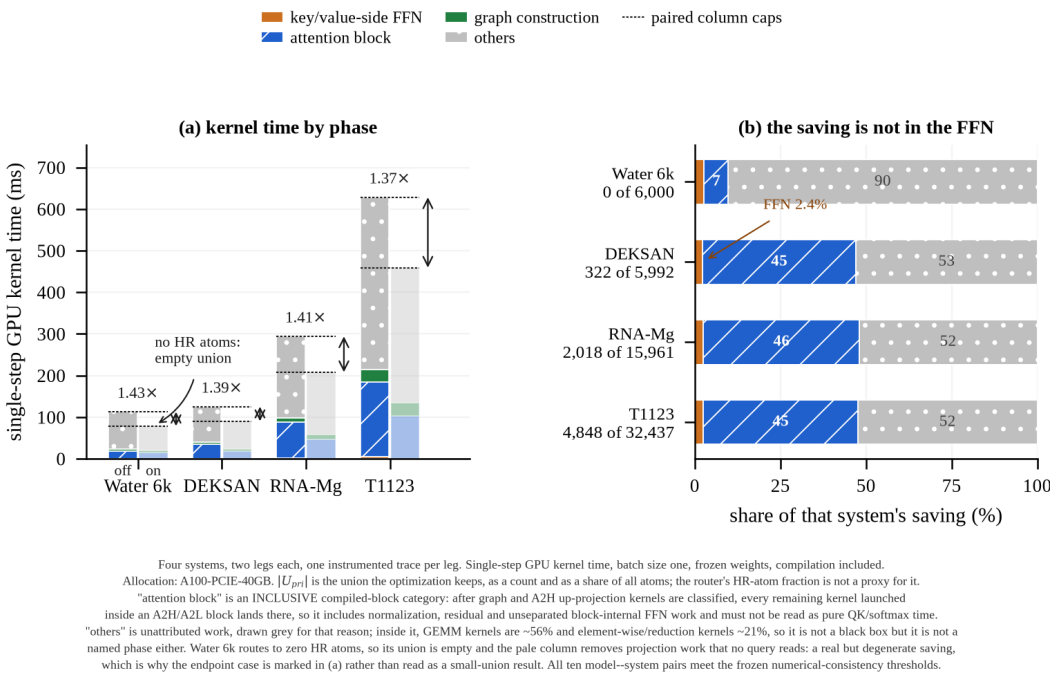


Figure 7: The neighbor-union KV bank, drawn on the solvated peptide, the smallest system that exercises both streams. Both paths are the same model on the same system, so the only thing that moves between the panes is a row count. OFF projects every node and expands the periodic images to $10{,}546$ rows; ON projects only the 322 rows (5.37% of atoms) that some A2H query can reach through `neighbor_idx`, running the identical projection operator on $32.7\times$ fewer rows. Both then gather to the same A2H query rows, so attention sees identical keys and values: no query row moves, no weight is written and no attention sparsity changes. That equivalence is what makes the rewrite algebraically lossless — the all-to-all side is read only through a gather, with no pooling across atoms and a row-wise `norm_1` — and it is also all it is: the path is not bit-exact, because the row count changes the GEMM kernel choice and the backward of `index_select` is an indeterminate `index_add`. The union row count is also not the fraction of atoms routed high-resolution, and the runtime gate refuses it above 0.9 of atoms, which is the case for Standard at 100.00% on all five systems; the percentages in the accepted row are union sizes, not speedups. All three qualifications that travel with the measurements this mechanism underlies are unchanged: the attention-block figure is an inclusive compiled-block category rather than pure attention kernel time, all ten model–system pairs meet the numerical-consistency thresholds, and the two high-dispersion Flash timing points retain their 6.28% and 10.75% process-median spans.



Four systems, two legs each, one instrumented trace per leg. Single-step GPU kernel time, batch size one, frozen weights, compilation included. Allocation: A100-PCIE-40GB. $|U_{pri}|$ is the union the optimization keeps, as a count and as a share of all atoms; the router's HR-atom fraction is not a proxy for it. "attention block" is an INCLUSIVE compiled-block category: after graph and A2H up-projection kernels are classified, every remaining kernel launched inside an A2H/A2L block lands there, so it includes normalization, residual and unseparated block-internal FFN work and must not be read as pure QK/softmax time. "others" is unattributed work, drawn grey for that reason; inside it, GEMM kernels are ~56% and element-wise/reduction kernels ~21%, so it is not a black box but it is not a named phase either. Water 6k routes to zero HR atoms, so its union is empty and the pale column removes projection work that no query reads: a real but degenerate saving, which is why the endpoint case is marked in (a) rather than read as a small-union result. All ten model--system pairs meet the frozen numerical-consistency thresholds.

Figure 8: Where the neighbor-union saving actually lands. Panel (a) stacks single-step GPU kernel time into four non-overlapping categories, optimization off against on, for four systems spanning 6.0k to 32.4k atoms; the dashed caps mark each column's total and the arrow the reduction between them. Panel (b) decomposes each system's own reduction, in the same colours. The component the optimization is named for, the key/value-side FFN, carries 2.35 to 2.72% of the saving on every system: a figure plotting it alone would attribute a 34.90 ms effect on the peptide system to a 0.82 ms component. The inclusive attention block carries 44.7 to 45.7% on the three non-degenerate systems, remarkably stable across a 5.4× span in atom count, and 52.0 to 52.9% lands in *others*, the unattributed category, drawn grey for that reason; within it GEMM kernels are about 56% and element-wise and reduction kernels about 20%, so it is not a black box, but it is not a named phase either and no claim rests on it. The leftmost system is an endpoint case, marked in (a): Flash routes pure water to *zero* high-resolution atoms, so its union is empty and the pale ON column deletes projection work no query can read — a real but degenerate saving, which is why it is not pooled with the others. All three qualifications that travel with the neighbor-union measurements hold here unchanged: the attention-block figure is an inclusive compiled-block category, not pure attention kernel time; all ten model–system pairs meet the numerical-consistency thresholds; and the two high-dispersion Flash timing points retain their 6.28% and 10.75% process-median spans. Allocation: A100-PCIE-40GB, one device for every row, so no number here is comparable with the H200 cost campaign in Table 4.

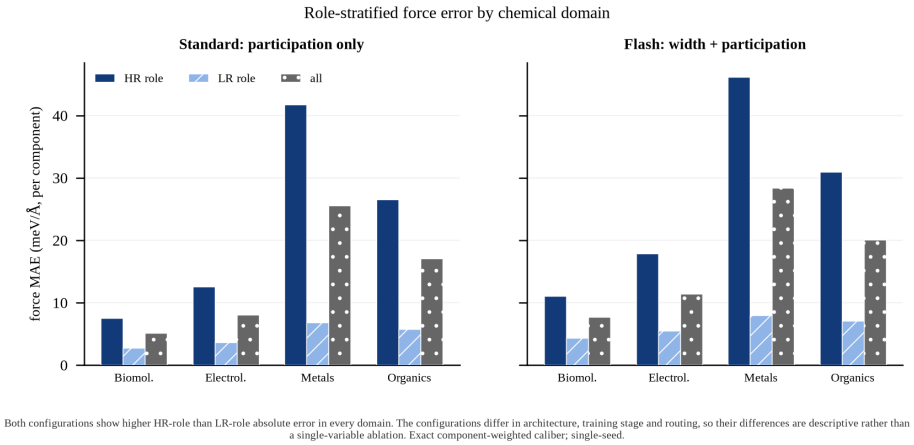


Figure 9: Role-stratified force error by chemical domain. Each panel shows HR-role, LR-role and all-atom per-component MAE for one production configuration under its bound router. The HR/LR imbalance appears in all four domains under both the participation-only Standard allocation and the stronger width-plus-participation Flash allocation. Exact component-weighted caliber, full split; the two configurations differ at once in architecture, training stage and routing, so their differences are descriptive rather than a single-variable ablation. Single-seed.

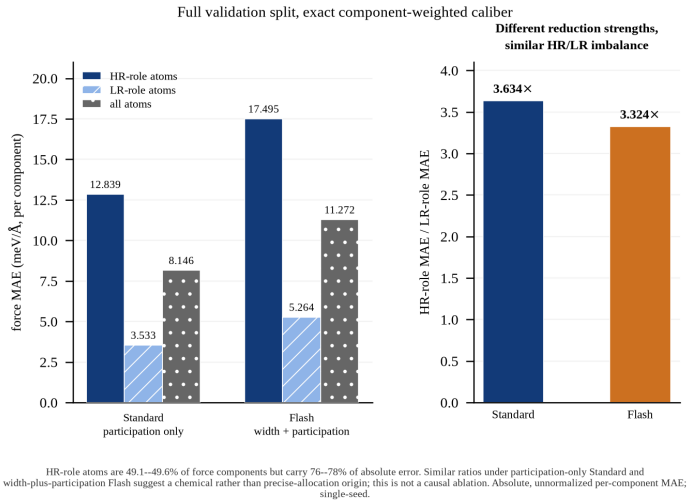
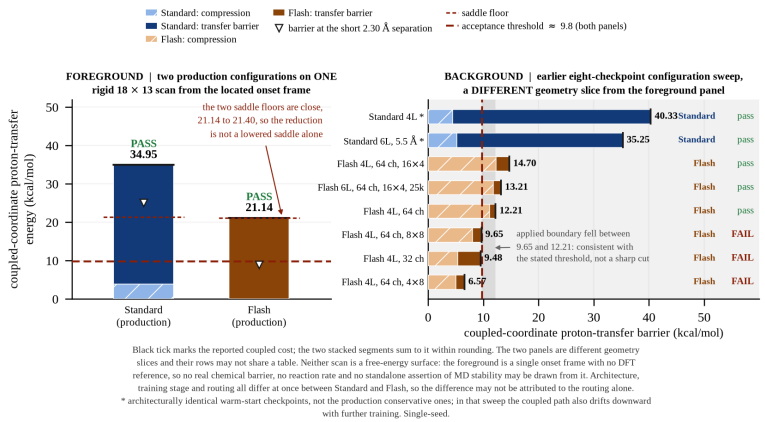
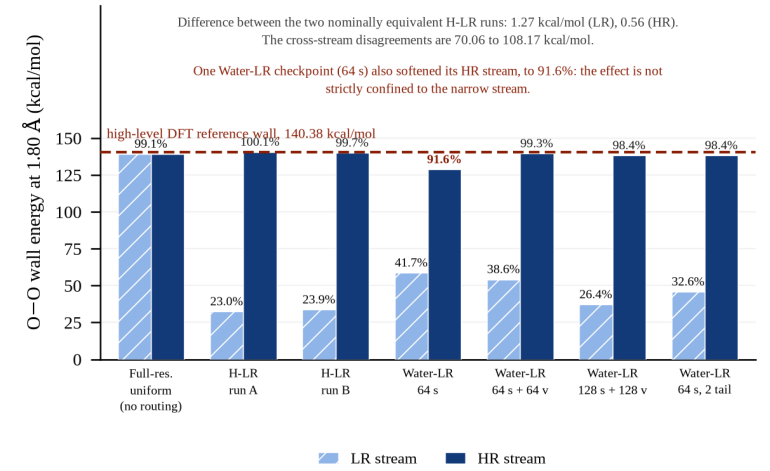


Figure 10: Role-stratified per-component force MAE, full 2762021-sample split, exact component-weighted caliber. Standard changes participation only; Flash changes persistent width and participation and routes more atoms low-resolution. Even under these different reduction strengths, their HR/LR ratios remain close, 3.634× and 3.324×. HR-role atoms are 49.1 to 49.6% of components but carry 76 to 78% of the absolute error. This consistency suggests that the imbalance is more likely chemical than a consequence of the precise capacity placement, but the two configurations are not a causal ablation. Bounds in the text above and Appendix I. Single-seed.



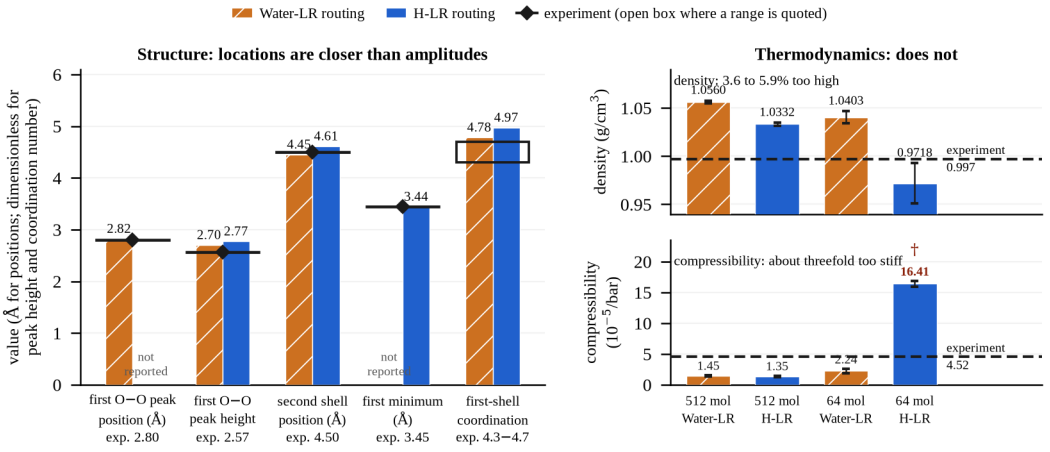
Black tick marks the reported coupled cost; the two stacked segments sum to it within rounding. The two panels are different geometry slices and their rows may not share a table. Neither scan is a free-energy surface: the foreground is a single onset frame with no DFT reference, so no real chemical barrier, no reaction rate and no standalone assertion of MD stability may be drawn from it. Architecture, training stage and routing all differ at once between Standard and Flash, so the difference may not be attributed to the routing alone. * architecturally identical warm-start checkpoints, not the production conservative ones; in that sweep the coupled path also drifts downward with further training. Single-seed.

Figure 11: Coupled-coordinate proton-transfer energetics, two panels on *different geometry slices* whose bars may not be read against each other. **Foreground**: the two production configurations on one identical rigid 18×13 scan from the located 511.855 ps failure-onset frame. Both pass the approximately 9.8 kcal/mol threshold, at coupled costs of 34.95 (Standard) and 21.14 (Flash); the marked saddle floors are close, 21.14 to 21.40, so Flash's reduction lives as much in compression and the product basin as in the barrier. **Background**: the earlier eight-checkpoint configuration sweep on a rigid 234-geometry slice, routing class marked on every bar so a Flash number is never read against a Standard one. There the two Standard configurations clear the threshold threefold to fourfold, at 40.33 and 35.25, while three of six Flash configurations fail, worst at 6.57. Its qualifications stand: the applied boundary fell between 9.65 and 12.21 rather than sharply, the two routings describe different surfaces, the coupled path drifts downward with further training, and both Standard rows are warm-start rather than production conservative checkpoints. All rows single-seed.



Bar labels give each value as a percentage of the reference wall. All rows are trained 200k-step checkpoints; ount scalar and vector LR channels. LR denotes the low-resolution role and HR the high-resolution one. The full-resolution uniform ro stream is plotted in both positions. LR-versus-HR disagreement on the same geometry was 70.06 to 108.17 kcal/mol for every mixed-resolution checkpoint. Single-seed.

Figure 12: Water-dimer oxygen-oxygen repulsive wall at 1.80 Å separation, LR stream against HR stream, per checkpoint, with each bar labelled as a percentage of the high-level DFT reference wall of 140.38 kcal/mol. LR and HR denote the low-resolution and high-resolution roles. The reduced-capacity stream systematically under-repels at short range, reproducing only 23.0 to 41.7% of the reference, while the HR streams sit at 98.4 to 100.1%; LR-versus-HR disagreement on the same geometry was 70.06 to 108.17 kcal/mol for every mixed-resolution checkpoint. The two nominally equivalent Standard runs differ by 1.27 kcal/mol on the LR stream and 0.56 kcal/mol on the HR stream. One Flash checkpoint also softened its HR stream, to 91.6%, so the effect is not strictly confined to the narrow stream. The full-resolution uniform row applies no routing, so its single stream is plotted in both positions. All rows are trained 200k-step checkpoints and single-seed.



Twelve measured 20 ps legs; the thermodynamic panels are the constant-pressure legs. The structural panel plots the recorded peak positions, heights, minima and coordination numbers: no full radial-distribution curves exist for these runs. Error bars are the reported uncertainties. † The 64-molecule H-LR leg is a protocol defect rather than a model defect: its interaction cutoff filled 97% of that box's half-width, and it is twelvefold inconsistent with its own larger box. The Water-LR constant-energy column is void as a test of the mixed-resolution model, because pure water has no HR atoms under that routing and only the narrow scalar stream executes. A 20 ps window is shorter than failures recorded at 40, 60, 511.9 and 680 ps in separate runs. Single-seed.

Figure 13: Liquid water: some structural locations are close, but amplitudes and thermodynamics are not uniformly reproduced. On the twelve water-only legs, the first oxygen-oxygen peak sits at 2.82 Å against 2.80 experimental, and a first minimum at 3.44 Å against 3.45 (the two are drawn from different routings, as the table's *not reported* cells record); peak heights are 2.70 and 2.77 against 2.57, and first-shell coordination is 4.78 and 4.97 against an experimental 4.3 to 4.7. The constant-pressure legs give densities 3.6 to 5.9% too high and compressibilities roughly threefold too stiff. A correct radial distribution function is therefore not evidence of a usable potential, and a 20 ps window is shorter than the tabulated failures near 40 and 60 ps and the separate failures at 511.9 and 680 ps. This panel plots scalar summaries against their experimental values, so its y-axis mixes units, Ångstroms for positions and dimensionless values for peak height and coordination number; for the full curves and for the three-seed values on the 54-leg matrix, see Figure 14 and Table 14. Error bars are the reported uncertainties. Two leg classes must be read with care. The 64-molecule Standard leg, marked with a dagger, is a protocol defect rather than a model defect: its interaction cutoff filled 97% of that box's half-width and the leg is twelvefold inconsistent with its own larger box, which invalidates its thermodynamic readings but not its structure or energy conservation. And the Flash pure-water column is the routing's *degenerate limit* rather than a failed test: pure water has no high-resolution atoms under that routing, so only the narrow scalar stream executes, and it cannot evidence cross-role exchange or carry an efficiency headline. All results are single-seed.

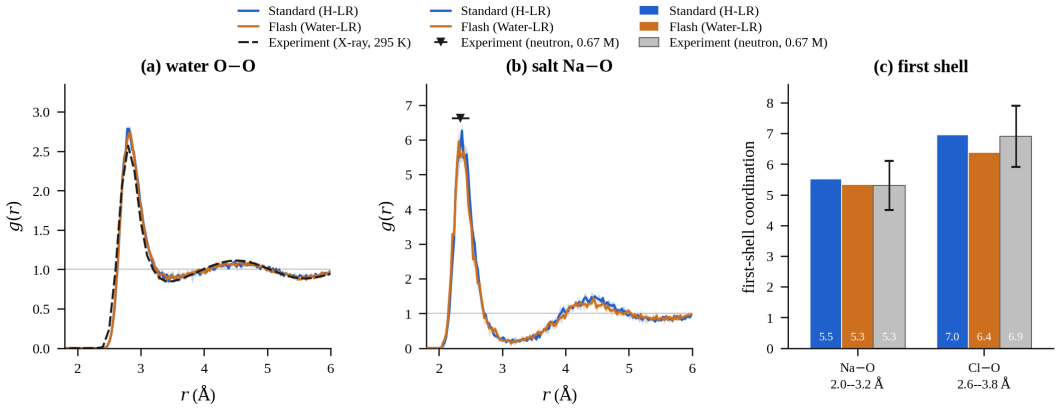


Figure 14: The 54-leg production matrix's radial distribution functions, on the NVT legs only, which are the legs for which curves were recorded. Each model curve is the three-seed mean with the seed-to-seed span shaded; every curve is 50 frames over $t = 10$ to 20 ps, one trajectory per seed. **(a)** Water O–O against the X-ray measurement of Skinner et al. (2013): the two routings and the experiment nearly coincide, the peak at 2.788 Å (Standard) and 2.838 Å (Flash) against 2.80, with the same first minimum at about 3.4 Å and the same second shell. The reference curve is a digitised rendering of that publication's figure rather than its authors' own table (InstaDeepAI, 2025), but it reproduces the published scalars to the last digit. **(b)** Salt-solution Na–O, where Standard and Flash again coincide to 0.05 Å; the marker is the measured first-peak position of Mancinelli et al. (2007), and no reference *curve* is drawn because that source publishes its distributions as figures without a table and no numeric first-peak height for Na–O or Cl–O was located anywhere in the literature. **(c)** First-shell coordination integrated over *exactly* each pair's published range (Na–O 2.0–3.2 Å, Cl–O 2.6–3.8 Å), so these numbers are the like-for-like counterparts of the published 5.3(0.8) and 6.9(1.0); integrating to each curve's own first minimum instead gives smaller values that are not comparable to a published integral (Table 14), which is why the two conventions are not mixed within the panel. Four bounds travel with the figure. The experimental ion values are at 0.67 M, the most dilute measurement that exists, against this box's 150 mM, so they are not a like-for-like comparison in concentration, and Mancinelli et al. (2007) reports the Na–O peak position as concentration-independent, which is what makes the dilute limit usable as a reference at all. The salt panel's evidence is weaker than the water panel's by construction, a position rather than a curve, and must not be read as equivalent. The box's Na–Cl pair distribution *is* measured but is not converged — 11 Na^+ and 11 Cl^- over 50 frames, with per-seed peak heights from 0.0 (no contact at all in 20 ps) to 18.2, so the three-seed means carry uncertainty larger than the value — and is therefore not drawn. And a correct radial distribution function is not evidence of a usable potential (Figure 13). One trajectory was run per seed; the three seeds give the shaded dispersion.

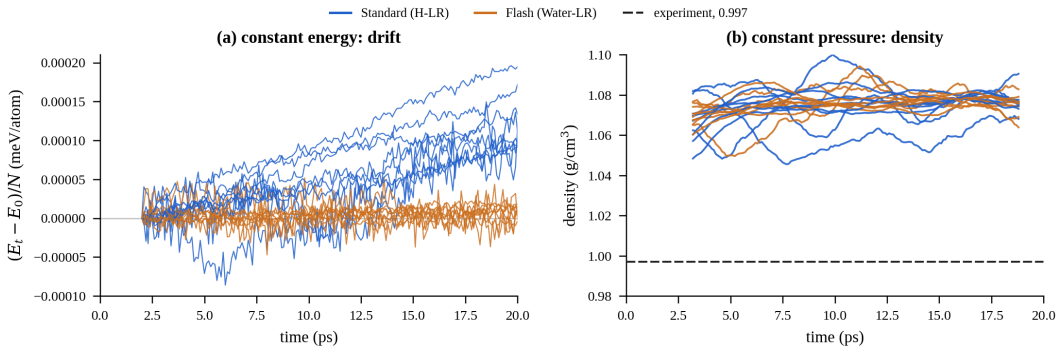


Figure 15: The measured 20 ps production dynamics matrix, drawn from the per-frame traces rather than from summary scalars: two models × three systems × three ensembles × three seeds, 54 legs of 40000 steps, all 54 completing with no failure, no non-finite value, no overlap and no oxygen-fusion anomaly, every constant-pressure leg's Monte-Carlo acceptance inside $[0.30, 0.40]$. One thin line per seed. **(a)** Constant-energy drift, $(E_t - E_0)/N$ per atom, which is the quantity a conservative potential is supposed to hold: the Flash traces are flat and centred on zero on all three systems, while three of the Standard traces show the visible upward slope whose slopes are tabulated in Table 19. **(b)** Constant-pressure density, rolled over 25 frames because the per-frame trace is dominated by the box's instantaneous fluctuation. Both models sit near 1.07 g/cm^3 against an experimental 0.997, and the overshoot is drawn rather than implied because it indicates a probable pressure and equation-of-state bias: stable, unfused, energy-conserving trajectories must not be rewritten as quantitatively accurate thermodynamics. **The duration bound is load-bearing**: 20 ps is shorter than the tabulated failures near 40 and 60 ps and the separate failures at 511.9 and 680 ps, so these legs support short-time stability and local structure only and are not a stability claim. The 1 ns tier of this 54-leg matrix was not run; the separate 1 ns NVT campaign on the two production checkpoints reaches 1000 ps on one system and is reported in Table 20. First-shell coordination, which this figure used to carry, is now read off the stored curves in Figure 14 and Table 14. Hydrogen-pair structural observables, ion residence and exchange rates, and the 200 ps screening tier remain unmeasured. One trajectory was run per seed; the three seeds give the spread visible at each time in both panels.