AquaGen: Scaling generative models to molecular dynamics precision on thousands of atoms

Emmanuel Bengio[*1,2], **Sanjeev Raja**[*1,2,†], **Yui Tik Pang**[1,2], **Kerstin Klaeser**[1,2], **Cristian Gabellini**[1,2], **Nikhil Shenoy**[1,2], **Francesco Di Giovanni**[1,2], **Prudencio Tossou**[1,2]

[1]Valence Labs, [2]Recursion, [*]Equal Contribution

We present *AquaGen*, the first all-atom, explicit solvent, periodic-boundary-condition-aware generative model that produces molecular configurations from the Boltzmann distribution at a fraction of the cost of molecular dynamics (MD). This is in contrast with existing generative models that remove degrees of freedom by operating on coarse-grained, vacuum, or implicit solvent systems. Operating at this resolution allows for post-processing through force field energy evaluations and MD simulations, and enables the prediction of relevant properties in a *gray-box* manner (as ensemble averages of potential energy evaluations over generated samples). We demonstrate the utility of this paradigm on absolute hydration free energy (AHFE), producing estimates 4-10x faster and with comparable accuracy to standard GPU-based MD. By generating uncorrelated samples from alchemical Boltzmann distributions, we create more accurate, interpretable, and refinable ensemble predictions with calibrated uncertainty estimates, unlike regression methods which are entirely *black-box* predictors. Our approach also yields predictable benefits from increasing train- and test-time compute, realized by scaling model size and generating more samples, respectively. We believe that this approach demonstrates the utility of high-resolution ensemble generation for free energy estimation, with future potential to replace MD in tasks such as the prediction of lipophilicity, membrane permeability, or absolute binding free energy (ABFE)–whose grounding and interpretability may be critical for the development of new drugs and materials.

1 Introduction

Molecular dynamics (MD) is a foundational tool for studying atomistic systems (Frenkel & Smit, 2001). Given a force field describing interatomic interactions, MD simulates the time evolution of a system by numerically integrating the equations of motion at femtosecond resolution. Extracting relevant quantities, such as solvation or binding free energies (Gilson & Zhou, 2007), relies on converged sampling of the configurational phase space. In practice, this often requires trajectories spanning nanoseconds to milliseconds (Shaw et al., 2008; Lindorff-Larsen et al., 2011) , resulting in up to billions of inherently sequential integration steps. Consequently, even modern GPU-accelerated MD simulations can require substantial computational resources while only exploring a limited portion of configurational space. MD nevertheless remains the gold standard for estimating a broad range of structural, dynamical, and thermodynamic properties due to its physical fidelity and generality (Shaw et al., 2010; Wang et al., 2015; Shirts et al., 2007).

In this work, we show that using *generative models*, we can produce uncorrelated, approximately Boltzmann-weighted molecular configurations with accuracy on par with MD, and in a much faster (and embarrassingly parallel) way. We introduce *AquaGen*, a generative model which operates at the same resolution as industry-standard bio-molecular MD simulations. We model all atoms, including solvents, and the periodic boundary conditions under which these simulations operate. As a demonstration of the paradigm, we focus on sampling the conformational landscape of small, drug-like molecules solvated in water. The scale of these systems is of the order of 10^3 atoms. While existing models (Abramson et al., 2024; Passaro et al., 2025; Lewis et al., 2025; Kim et al., 2024) operate at a similar scale, they either do not account for water, use implicit water, or operate on a lower-dimensional representation (e.g. considering only backbone or heavy atoms). As such, the outputs of these models are either not immediately compatible with high-fidelity energy functions, or require significant post-processing (e.g., energy minimization) before being used in downstream calculations like free energy perturbations (FEP). By contrast, AquaGen is, to our knowledge, the first generative model whose

[†]UC Berkeley; work done during an internship at Valence Labs

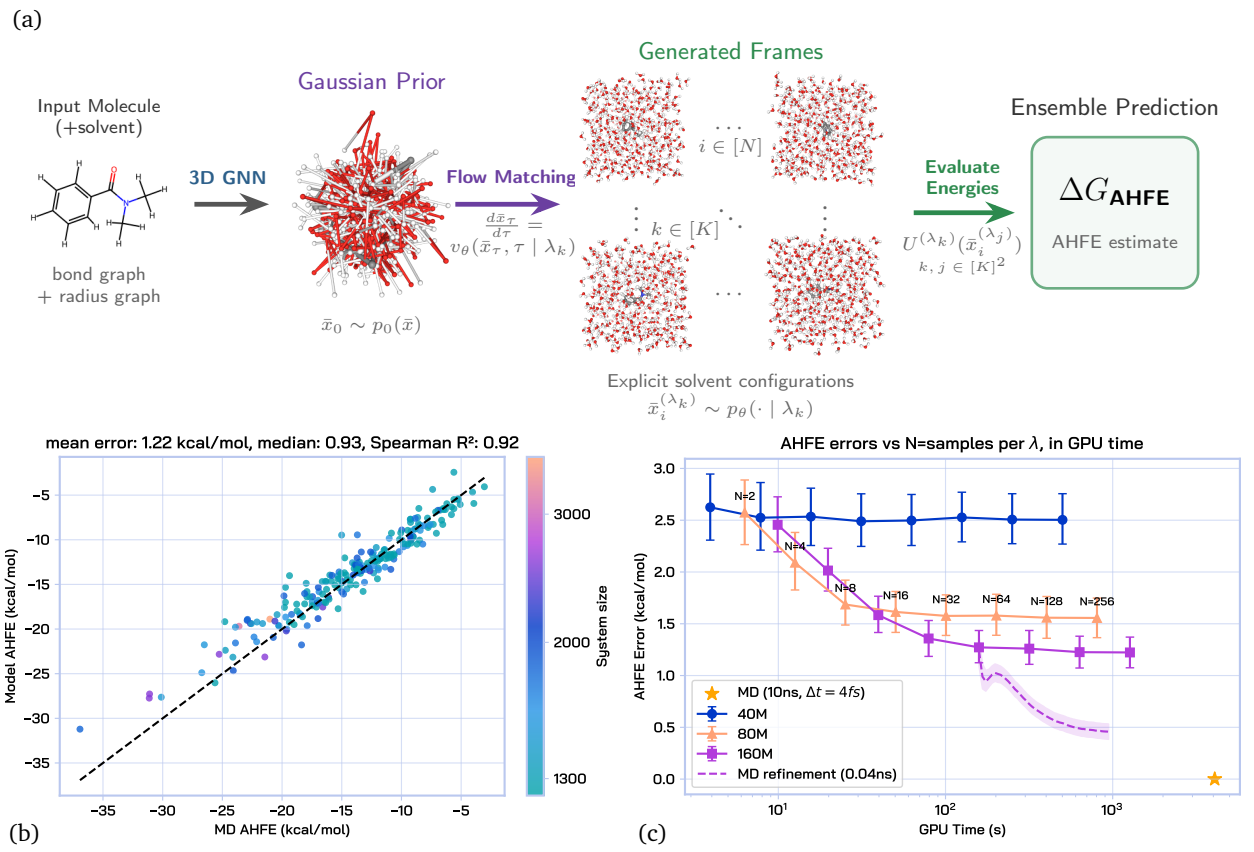

Figure 1: **Overview of AquaGen generative modeling framework and results on absolute hydration free energy (AHFE) estimation**. **(a)** A flow matching model is trained to generate uncorrelated, all-atom, explicit-solvent, PBC-aware configurations $\bar{x}_i^{(\lambda_k)} \sim p_\theta(\cdot \mid \lambda_k), i \in [1, N], k \in [1, K]$ which approximate samples from data distributions $p_{\text{data}}(\cdot \mid \lambda_k)$ along an alchemical pathway governed by λ_k. Potential energy evaluations $U^{(\lambda_k)}(\bar{x}_i^{(\lambda_j)})$ and MBAR are then used to estimate the absolute hydration free energy (AHFE). **(b)** Unrefined model samples yield an average and median error of 1.22 kcal/mol and 0.93 kcal/mol respectively on unseen systems, color-coded by number of atoms. **(c)** Our framework makes use of train- and test-time compute to improve AHFE prediction accuracy. Scaling model size from 40M to 160M parameters improves accuracy, while generating more samples at test-time achieves a similar effect. Operating at all-atom, explicit solvent resolution allows for MD refinement (green dashed line) starting from generated samples (green solid line), while remaining 4× faster than MD.

generated configurations are sufficiently geometrically accurate and diverse that they closely reproduce the energy distribution of reference MD simulations under an explicit-solvent, all-atom force field.

Faithfully reproducing the all-atom Boltzmann distribution unlocks the ability to compute pharmaceutically relevant quantities as ensemble averages of potential energy evaluations over generated samples. In this work, we demonstrate the value of this approach in predicting the *absolute hydration free energy (AHFE)*: the Gibbs free energy change associated with transferring a compound from vacuum to water solvent. We train AquaGen to sample conformations of water-solvated, drug-like compounds, conditioned on an alchemical order parameter λ which gradually attenuates interactions between the compound and the solvent. We then evaluate potential energies of the generated structures using an OpenFF 2.1.1 force field in conjunction with the TIP3P water model. Average potential energy differences $U^{(\lambda_i)} - U^{(\lambda_j)}$ determine the degree of overlap between samples at different points on the alchemical pathway. These energy differences, computed over many samples and many values of λ, yield the AHFE via the unbiased multistate Bennett acceptance ratio (MBAR) estimator (Shirts & Chodera, 2008). We show that AquaGen is accurate and efficient, producing AHFE estimates at 4-10× the speed of GPU-based MD, with approximately 1 kcal/mol error. We demonstrate

predictable gains from scaling train- and test-time compute, by increasing model capacity and the number of generated samples, respectively.

Since the potential energy is calculated with a physics-based force field, we deem it a *white-box* component of our pipeline. Meanwhile, we consider the generative model itself to be a *black-box* entity, as it is a flexibly parameterized, learnable function approximator. We thus characterize our overall AHFE prediction framework as *gray-box*, in contrast with the growing suite of methods on purely black-box chemical property prediction (Chithrananda et al., 2020; Ahmad et al., 2022; Heid et al., 2023; Passaro et al., 2025). Several advantages naturally arise from this gray-box characteristic. For instance, property predictions from AquaGen are *refinable*. Because our model generates physically valid atomistic conformations that are directly compatible with potential energy evaluations, they can serve as starting points for short MD refinement to improve accuracy. Additionally, uncertainty estimates derived from simple bootstrapping over generated samples are well-calibrated with AHFE prediction errors, which may be critical for trust and widespread adoption. If we push this gray-box framework to its logical conclusion, we may be able to scale to more complex atomic systems with $10^4 - 10^5$ atoms and compute quantities such as protein-ligand binding energy (Gilson & Zhou, 2007) in an inspectable, interpretable, refinable, and reliable manner.

To summarize, the main contributions of this work are the following:

1. We introduce *AquaGen*, which is to our knowledge the first all-atom, explicit water molecular generative model of Boltzman-distributed atomistic conformations, scaling up to 4×10^3 atoms.
2. We achieve high energetic and structural accuracy relative to the true Boltzmann distribution of solvated compounds, obtained from over 1.5 billion frames of alchemical MD trajectories.
3. By conditioning AquaGen on an alchemical order parameter λ, we compute absolute hydration free energies (AHFE) of held-out compounds with approximately 1 kcal/mol error, with a 4-10x speedup relative to MD.
4. We demonstrate that AHFE predictions derived from AquaGen are refinable via short MD simulations to further reduce error to $<$ 0.5 kcal/mol, and that simple, zero-shot uncertainty quantification techniques are well-calibrated with model errors.

2 Molecular Dynamics Simulation and Free Energy Calculations

We briefly review the major technical concepts relevant to this work. In §2.1, we begin by reviewing concepts relevant to MD simulation. In §2.2, we review how alchemical MD simulations can be used to compute the *absolute hydration free energy* (AHFE) of a compound. The computational expense of these simulations motivates the development of surrogate generative models, which will be introduced in §3.

2.1 Molecular Dynamics Simulation

MD simulates the time-evolution of a molecular system under a potential energy function. Formally, denote the coordinates of an N-atom molecular system as $x \in \mathbb{R}^{N\times 3}$, and the potential energy function as $U(x) : \mathbb{R}^{N\times 3} \to \mathbb{R}$. The dynamics evolve according to Newton's equations of motion, where the force on each atom is given by the negative gradient of the potential energy:

$$m_i \frac{d^2 x_i}{dt^2} = F_i(x) = -\nabla_{x_i} U(x), \tag{1}$$

for atom i with mass m_i. In practice, these equations are numerically integrated using discrete timesteps to generate trajectories $\{x_t\}_{t=0}^{T}$. In this work, we consider MD of solvated, drug-like compounds obtained from the early stages of internal drug discovery campaigns. To reflect common experimental conditions, we perform constant-temperature, constant-pressure (NPT) MD simulations. In principle under an appropriate choice of thermostat and barostat, NPT MD simulation ergodically samples from the isothermal-isobaric Boltzmann distribution:

$$p_{\text{NPT}}(\bar{x}) = \frac{\exp(-\beta(U(\bar{x}) + PV(\bar{x})))}{Z}, \tag{2}$$

meaning that a time average of any observable $\hat{O} = \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} O(\bar{x}_t)$ over a sufficiently long simulation converges to an unbiased expectation over the Boltzmann distribution $\mathbb{E}_{\bar{x}\sim p_{\text{NPT}}} O(\bar{x})$. Here, $\bar{x} = \{x, c\}$, where

$c \in \mathbb{R}^{3\times3}$ represents the simulation box, $V : \mathbb{R}^{3\times3} \to \mathbb{R}$ is the instantaneous volume, $\beta = \frac{1}{k_B T}$ is the inverse temperature, P is the pressure, and Z is the partition function.

2.2 Absolute Hydration Free Energy

Many problems in drug discovery reduce to estimating free energy differences ΔG, which characterize the extent to which a molecular process, such as solvation or binding, is energetically favorable (Cournia et al., 2017; Muegge & Hu, 2023). The Gibbs free energy of a state can be expressed in terms of its partition function as $G = -\beta^{-1} \log Z$, such that the difference in free energy between two states A and B is

$$\Delta G_{A\to B} = G_B - G_A = -\beta^{-1} \ln \frac{Z_B}{Z_A}. \tag{3}$$

A particularly important example is the absolute hydration free energy (AHFE), which measures the free energy change associated with transferring a compound from vacuum into solvent (Mobley & Guthrie, 2014). AHFE is a key component of most downstream biomolecular binding calculations, which typically involve transfer of a compound from solution to a binding site. For this reason, AHFE prediction accuracy is often viewed as an approximate upper bound on accuracy for more complex binding free energy prediction tasks. The fact that free energy differences are *state functions*—that is, they depend only on the endpoints and not on the paths connecting them—can be exploited to compute ΔG_{AHFE} as

$$\Delta G_{\text{AHFE}} = \Delta G_{\text{vacuum}} - \Delta G_{\text{solvated}}, \tag{4}$$

where ΔG_{vacuum} and $\Delta G_{\text{solvated}}$ are the free energy changes associated with annihilating intermolecular interactions in vacuum and solvent respectively. The first term is typically cheap to compute, as it only involves a single drug-like molecule in vacuum. The second term is the computational bottleneck due to the presence of solvent and the resulting high-dimensional configurational space. Focusing on the solvated component, the free energy difference can formally be written in terms of a ratio of partition functions:

$$\Delta G_{\text{solvated}} = -\beta^{-1} \log \frac{Z_{\text{solvated, interacting}}}{Z_{\text{solvated, non-interacting}}}. \tag{5}$$

These partition functions are intractable, as they require integration over all configurational degrees of freedom of the compound and solvent. Instead, one can leverage MD to sample from the corresponding Boltzmann distributions. The Zwanzig formula

$$\Delta G_{\text{solvated}} = -\beta^{-1} \log \mathbb{E}_{\bar{x}\sim p_{\text{solvated, non-interacting}}} \left[\exp\left(-\beta\left(U_{\text{solvated, interacting}}(\bar{x}) - U_{\text{solvated, non-interacting}}(\bar{x})\right)\right)\right] \tag{6}$$

gives an unbiased estimator of the solvation free energy difference, with expectations approximated via samples from MD (Zwanzig, 1954). However, this estimator typically has impractically high variance due to the low overlap between the distributions $p_{\text{solvated, non-interacting}}$ and $p_{\text{solvated, interacting}}$. Converged free energy estimates would require potentially unfeasibly long MD simulations to adequately sample the extremely rare regions of overlap between these distributions. To alleviate this issue, one can introduce a sequence of overlapping intermediate alchemical distributions indexed by $\{\lambda_k\}_{k=1}^{K}$ (each with potential energy $U^{(\lambda_k)}$) that gradually transition from a fully interacting to a fully non-interacting system. The multistate Bennett acceptance ratio (MBAR) estimator (Shirts & Chodera, 2008) can be used to pool samples from all λ-states and compute an unbiased, minimum-variance estimate of the relative free energies. Thus, estimating $\Delta G_{\text{solvated}}$ reduces to running multiple λ-dependent MD simulations and computing MBAR on the samples.

Computational Cost. MD simulation offers a principled mechanism by which to sample from the Boltzmann distribution, and alchemical simulations can be used to accurately compute relevant free energy differences such as AHFE. However, due to presence of energy barriers separating relevant regions of the potential energy surface, MD requires long time horizons for relevant expectation values to converge. Considering the small time discretization necessary for numerical stability, along with the iterative and non-parallelizable nature of the algorithm, this makes alchemical MD simulation expensive in wall clock time–on the order of 1 GPU-hour per λ for a 1000-atom system. This is further exacerbated by the fact that the simulation length necessary for convergence can be a strong function of the system of interest, meaning that in practice we may need to run long simulations for all systems to ensure convergence.

3 AquaGen: All-Atom, Explicit-Water Generative Modeling

To alleviate the challenges posed by all-atom, explicit-water MD simulations, we turn to *generative modeling*.

3.1 Generative Modeling Framework

The central idea is to train a model which can generate Boltzmann-distributed configurations that are energetically and structurally indistinguishable from MD samples, without slow, iterative traversal of configurations. Formally, given a data distribution $p_{\text{data}} \approx p_{\text{NPT}}$ obtained from classical MD simulation, we wish to parameterize a generative model f_θ which maps a sample $z \sim p_0$ from a tractable prior distribution, p_0, to a sample $x = f_\theta(z) \sim p_\theta \approx p_{\text{data}}$ matching an i.i.d draw from p_{data}. The generative modeling framework straightforwardly extends to sampling from alchemical Boltzmann distributions $\{p_{\text{data}}(\cdot \mid \lambda)\}_{\lambda\in[0,1]}$ simply by conditioning the generative model on the alchemical parameter λ. Concretely, the λ-conditional model $f_\theta(\cdot|\lambda)$ learns to generate configurations $\bar{x} = \{x, c\}$ distributed according to $p_{\text{data}}(\cdot|\lambda)$, where the $[0, 1]$ interval is discretized into K lambda values $\{\lambda_k\}_{k=1}^K$. After training, we draw N samples $\{\bar{x}_i^{(\lambda_k)}\}_{i=1}^N \sim p_\theta(\cdot \mid \lambda_k)$ for $k \in [1, K]$ from the generative model and follow the rest of the AHFE estimation procedure described in §2.2 as if we had obtained samples from long, alchemical MD simulations: evaluate reduced energies $\{u^{(\lambda_k)}(\bar{x}_i)\}_{i=1}^N = \{\beta U^{(\lambda_k)}(\bar{x}_i)\}_{i=1}^N$, and pass these into MBAR to obtain an estimate of $\Delta G_{\text{solvated}}$. For the vacuum contribution ΔG_{vacuum}, we rely on classical MD estimates, as sampling in vacuum is computationally inexpensive relative to the solvated setting. The final AHFE is computed from equation 4.

Advantages of Generative AHFE Estimation. Unlike regression models which make *black-box* predictions, the generative route is *gray-box*; the AHFE prediction is derived via actual evaluations of potential energy functions $U^{(\lambda_k)}$ on the generated samples, which have interpretable physical meaning. Due to the use of MBAR, reasonable estimates of AHFE can only stem from reasonable overlaps in the energy distributions of adjacent λ values, and by extension from physical Boltzmann-weighted configurations. As we will show in §4.1, computing AHFE via generative model samples also introduces a natural axis of test-time compute scaling to increase prediction accuracy: the number of generated configurations per λ value. The paradigm also enables *refinability*, meaning it is possible to further increase accuracy by initiating very short MD simulations in parallel from the generated samples, and zero-shot *uncertainty quantification* via bootstrapped confidence intervals.

3.2 Generative Modeling Details

We use flow matching (Lipman et al., 2023), a class of generative models which learns a transport map between a tractable prior distribution p_0 and a target data distribution $p_{\text{data}}(\cdot \mid \lambda)$. Specifically, flow matching parameterizes a time-dependent "velocity" field $v_\theta(\bar{x}_\tau, \tau|\lambda)$ which defines a deterministic flow via the ordinary differential equation

$$\frac{d\bar{x}_\tau}{d\tau} = v_\theta(\bar{x}_\tau, \tau|\lambda), \quad \bar{x}_0 \sim p_0. \tag{7}$$

We emphasize that the flow matching velocity field is not a physical velocity, but rather is a rate of change of probability along the conditional path. We also use τ to denote the flow matching time in order to differentiate it from the notion of physical time t in MD simulations. The model is trained to match a target conditional velocity field $v^\star(\bar{x}_\tau, \tau|\lambda)$ induced by a prescribed conditional probability path between p_0 and $p_{\text{data}}(\cdot \mid \lambda)$, by minimizing

$$\mathcal{L}(\theta) = \mathbb{E}_{\tau,\lambda,\bar{x}_\tau}\left[\|v_\theta(\bar{x}_\tau, \tau|\lambda) - v^\star(\bar{x}_\tau, \tau|\lambda)\|^2\right]. \tag{8}$$

We perform flow matching jointly over Euclidean coordinates x and the simulation cell c. We use a Gaussian prior $p_0(x) = \mathcal{N}(0, 2I_{3N})$ for the Euclidean coordinates (Figure 1a). For the cell, we adopt a cubic Gaussian prior; that is, we sample a box length from $\ell \sim \mathcal{N}(\mu_c, \sigma_c^2)$ and construct a cell from mutually orthogonal lattice vectors of length ℓ. We adopt a simple, linear conditional probability path for training:

$$\bar{x}_\tau = \tau\bar{x}_1 + (1 - \tau)\bar{x}_0, \tag{9}$$

where $\bar{x}_1 \sim p_{\text{data}}(\cdot \mid \lambda)$ and $\bar{x}_0 \sim p_0$, and addition and multiplication are performed independently for the Euclidean coordinates and simulation box. The target conditional velocity for this choice of probability path is

given by

$$v^\star(\bar{x}_\tau, \tau|\lambda) = \bar{x}_1 - \bar{x}_0. \tag{10}$$

At inference time, we draw a prior sample from $p_0(\bar{x})$ and solve the ODE in equation 7 with Euler integration over $\tau \in [0, 1]$ to produce an approximate sample from $p_{\text{data}}(\bar{x} \mid \lambda)$. We find that the learned vector field has higher curvature near the prior, so we use a finer discretization near $\tau \approx 0$ (see §C.1 for more details).

Model architecture. We parameterize the velocity field v_θ with a standard graph neural network (GNN). The model's input coordinates are encoded as pairwise displacement vectors between nodes in the graph. While we account for all atoms in the input featurization and final velocity prediction head, in the core message-passing block we represent water O-H-H triplets with a single node to reduce overhead for large systems dominated by solvent molecules. We model periodic boundary conditions induced by the simulation box as additional virtual nodes. For more details on the model architecture, see §A.1.

3.3 Molecular Dynamics Training Data

We use a dataset comprising several thousand drug-like compounds derived from internal medicinal chemistry lead optimization. For each compound, we ran alchemical MD simulations with 20 discrete λ values with Hamiltonian replica exchange (HREX). The alchemical pathway uses a partial annihilation scheme in which electrostatic interactions are fully annihilated prior to Lennard-Jones interactions being decoupled (Alibay et al., 2026). We extracted uncorrelated samples from production trajectories of each λ value and each compound, yielding over 1 billion training frames. Hydration free energies were computed using the MBAR estimator as implemented in PyMBAR (Shirts & Chodera, 2008). See §B for more details about the training data.

4 Results

We first present the highest-level, end-to-end AHFE estimation results from Figure 1 in §4.1 . We then zoom in and analyze the energetic and structural accuracy of the generated samples at the fully-interacting endpoint (§4.2) and the intermediate alchemical distributions (§4.3). In §4.4, we discuss more nuanced aspects of AHFE estimation with AquaGen, such as uncertainty estimation and error cancellation. Finally, we perform various ablations in §4.5.

4.1 Absolute Hydration Free Energy Estimation

We use the generated samples at each λ value along the alchemical pathway to compute an unbiased estimate of $\Delta G_{\text{solvated}}$ via the MBAR estimator. This is combined with ΔG_{vacuum} from classical MD to produce a final AHFE prediction (equation 4). For a given compound, the AHFE absolute error (AE) is obtained as the absolute difference between the predicted AHFE, computed via MBAR on the generated samples, and the true AHFE, computed via MBAR on reference MD samples. Formally, given samples across K alchemical states, MBAR returns a matrix $\Delta\hat{G} \in \mathbb{R}^{K\times K}$ representing the estimated pairwise free energy differences between states. We compute $\Delta\hat{G}^{\text{model}}$ and $\Delta\hat{G}^{\text{MD}}$ from the generative model and MD samples respectively. The AHFE AE is computed as

$$\text{AHFE AE} = |\Delta\hat{G}^{\text{model}}_{1,K} - \Delta\hat{G}^{\text{MD}}_{1,K}| \tag{11}$$

Prediction accuracy. Figure 1b shows AHFE values resulting from MBAR on samples from AquaGen (y-axis) and ground truth alchemical MD simulations (x-axis) for 218 held-out compounds not seen during training. We achieve a median and mean AHFE AE of 0.93 kcal/mol and 1.22 kcal/mol relative to MD, respectively. More water-soluble compounds (i.e., those with a more negative AHFE) generally have higher positive errors. Given equation 4, this translates to an *under-estimation* of the change in solvation free energy $\Delta G_{\text{solvated}}$. For highly soluble compounds with energetically favorable electrostatic interactions between the compound and solvent, we hypothesize that due to suboptimal conditioning on the alchemical parameter λ, the model learns an "averaged" distribution over $\lambda_k, k \in [1, 5]$ which yields larger overlaps and a smaller $\Delta G_{\text{solvated}}$ (we substantiate this hypothesis further in §4.4 when we discuss *error cancellation*).

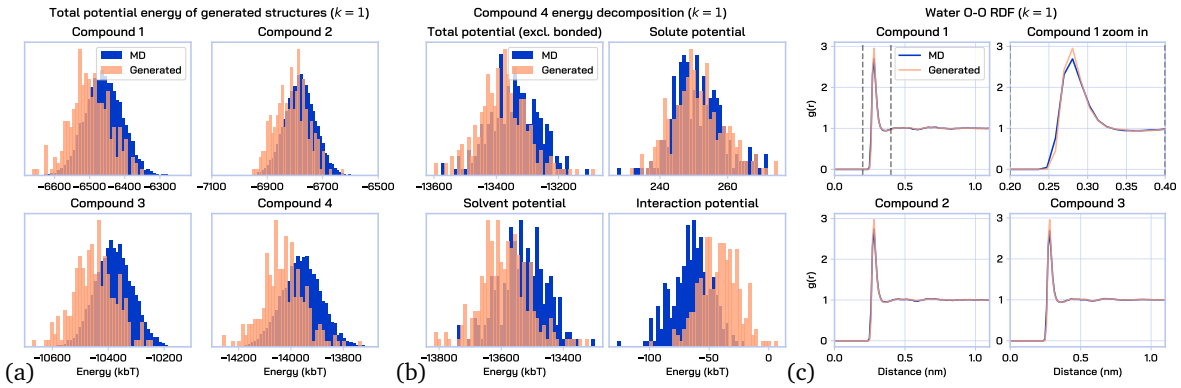

Figure 2: **Structural and energetic accuracy of samples generated by AquaGen at the fully interacting endpoint ($\lambda_k, k = 1$) (a)** True vs generated potential energy for 4 randomly chosen compounds. **(b)** Various decompositions of compound 4 energy. In clockwise order, starting from top left: total potential energy excluding solvent bonded terms (since the reference simulations are performed with rigid water), solute-only potential energy, solute-solvent interaction potential energy, solvent-only potential energy. All generated distributions are strongly overlapping with reference distributions, with slight overestimation of the solute-solvent interaction energy. **(c)** True and generated water O-O radial distribution functions (RDFs). The generated RDF is slightly overly ordered relative to the MD reference.

Train/test-time scaling and refinability. Our approach naturally admits train- and test-time compute scaling to improve AHFE estimation accuracy. Train-time compute is scaled by increasing model capacity, while test-time compute is scaled by increasing the number of generated samples from 2 to 256 at each λ_k. Figure 1c shows mean AHFE AE as a function of test-time GPU compute per λ value, per compound, with curves for varying model sizes. For any model size, generating more samples leads to improved results. The performance of smaller models plateaus more quickly, while larger models continue to improve up to 256 samples. Scaling model size from 40M to 160M parameters leads to improvements in AHFE at larger test-time compute budgets, while all models are relatively similar at smaller compute budgets. This suggests that larger models produce more diverse samples. By initiating very short (40 ps) MD refinements from the generated samples of the 160M parameter model, we achieve an AHFE error of 0.5 kcal/mol relative to MD, still with about 4x less GPU time (we use the same settings for the MD refinement as the data generation, see §B). Note that generative model sampling is trivially parallelizable across λ-windows, while MD incurs communication overhead for parallelization across λ due to the replica exchange, and most importantly is not parallelizable across time. Thus, given more inference hardware resources, in the future we can expect greater speedups in *wall clock time* from generative modeling relative to MD.

4.2 Energetic and Structural Accuracy of Fully-Interacting Endpoint

Having assessed the end-to-end AHFE estimation capabilities of AquaGen, we now zoom in and assess our ability to produce samples which closely align with the physical Boltzmann distribution produced by MD. To facilitate this understanding, in this section, we report results from a model trained only on frames from the fully interacting distribution (equation 2). Results are shown in Figure 2. The potential energy distribution of generated samples is highly overlapping with the reference distribution produced by MD (Figure 2a). No refinement or post-processing (e.g., energy minimization) was performed on the sampled frames. More granular energy decompositions (2b) reveal that the model slightly overestimates the solute-solvent interaction energy. While the ground truth MD data uses a rigid water assumption, AquaGen has no such constraint on the positions of generated water atoms. Thus, we have excluded bonded solvent energy contributions in Figure 2b on the top left (see §C for more detailed analysis). We also find that the bulk-water radial distribution functions (RDFs) produced by the model are quite accurate, though slightly overly ordered (Figure 2 c).

4.3 Energetic and Structural Accuracy of Alchemical Annihilation

We next assess the extent to which AquaGen captures the essential physics of the solute-solvent interaction annihilation along the alchemical pathway $\{p_{\text{NPT}}(\bar{x} \mid \lambda_k)\}_{k\in[1,20]}$. Results in this section are reported on

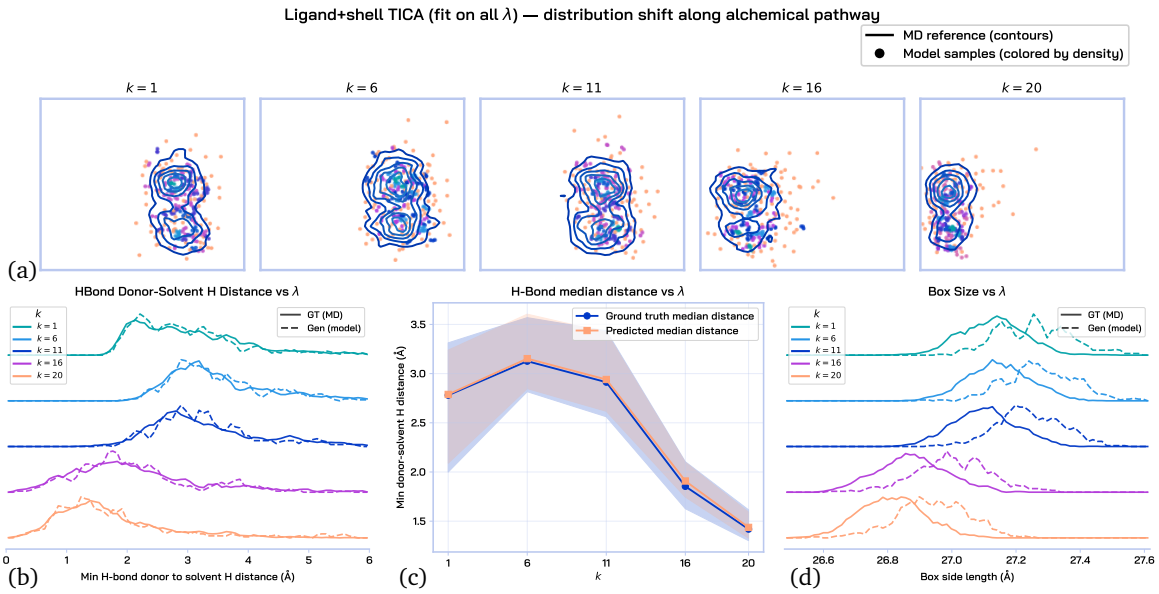

Figure 3: **Structural and energetic accuracy of all-atom, explicit water samples generated by AquaGen along the alchemical pathway ($\lambda_k, k \in [1, 20]$). (a)** Time-lagged independent component analysis (tICA) plots of samples from reference MD simulations (contours) and flow matching model (points - darker indicates higher density) across $\lambda_k, k \in \{1, 6, 11, 16, 20\}$. The model samples move across the first tICA component along with the reference contours, indicating sampling of the correct regions of conformational space. **(b)** Distribution of minimum distance between solvent hydrogen atom and electronegative atom (N, O, F) on the compound, at different lambdas, for a randomly selected compound. The generated samples closely match the non-monotonic trend along the alchemical pathway. **(c)** The mean across compounds of the minimum distance between solvent hydrogen atom and electronegative atom (N, O, F) on the compound, as a function of λ_k. The model samples capture the trend closely. **(d)** Distribution of simulation box lengths of generated and reference samples. While AquaGen captures the correct trend as a function of λ, it tends to overestimate the box size.

a model trained on all $\lambda_k, k \in [1, 20]$ (the same model which was used to produce the end-to-end AHFE results in Figure 1). Figure 3a shows a time-lagged Independent Component Analysis (tICA) plot of the reference MD and generated samples across λ values, showing that the model samples the correct the regions of conformational space (see §C for tICA plots of more compounds and details on the tICA featurization). We also analyze the distribution of distances between an electronegative atom (N, O, or F) on the compound and the closest hydrogen atom on a solvent water molecule, constituting a potential hydrogen-bond donor-acceptor pair. We expect the mean of this distribution to be a non-monotonic function of the alchemical parameter λ. In the electrostatic regime ($\lambda_k, k \in [1, 5]$), the annihilation of attractive electrostatic interactions between the compound and solvent should lead to an increase in the mean hydrogen-bond distance, while in the van der Waals regime ($\lambda_k, k \in [6, 20]$), the annihilation of repulsive, steric interactions should lead to a decrease in the mean hydrogen-bond distance. AquaGen captures this trend well, both for a single randomly chosen compound (Figure 3b) and across many compounds (Figure 3c). A salient and consistent error of AquaGen is the tendency to overestimate the simulation box length (Figure 3d). We observe a consistent overestimation of about 0.1 Å, which we attribute to our choice of prior (see §4.5 for details).

4.4 Additional Analysis of AHFE Predictions

We now discuss more nuanced aspects of estimating AHFE with AquaGen; notably, estimation of uncertainities, generalization away from the training set, error cancellation across the alchemical pathway, and a comparison to black-box prediction methods.

Calibrated uncertainties. As another demonstration of the utility of our gray-box framework for AHFE prediction, we can extract uncertainty estimates from AquaGen at test-time by computing the width of

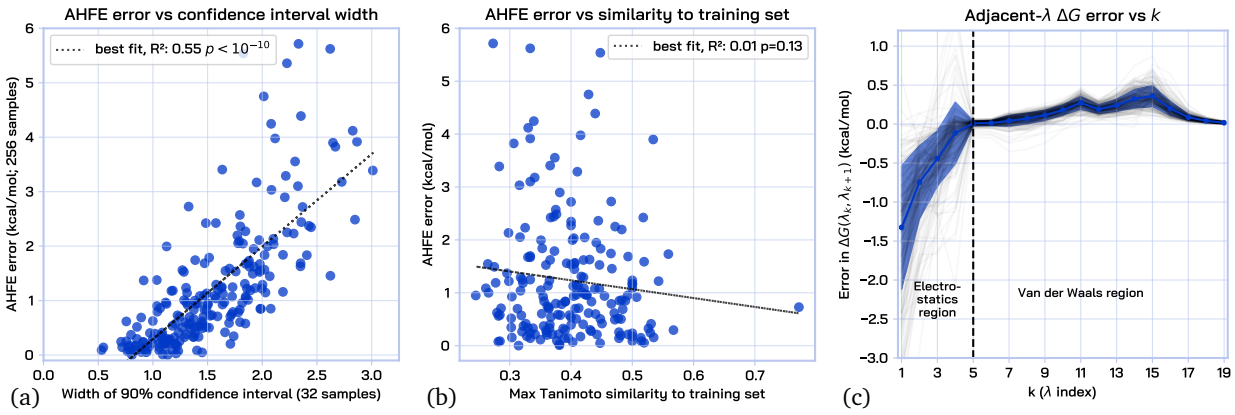

Figure 4: **Analysis of uncertainty estimation, generalization across compounds, and error cancellation when using AquaGen for AHFE prediction.** **(a)** The width of bootstrapped 90% confidence intervals, computed from random 32-sample subsets of 128 generated configurations, correlates strongly with the AHFE prediction error, indicating that AquaGen provides well-calibrated uncertainty estimates. **(b)** AHFE prediction error as a function of the maximum Tanimoto similarity between each test compound and any compound in the training set. The weak correlation suggests that AquaGen learns transferable representations as opposed to memorizing training compounds. **(c)** Errors in the first off-diagonal of the MBAR $\Delta\hat{G}$ matrix estimated from the reference MD and generated samples. The AHFE Absolute Error (AE) is the cumulative sum (i.e., area under the curve) of this plot. Since errors have opposite signs in different regions of the alchemical pathway, *error cancellation* affects the final AHFE AE. The Cumulative Absolute Error (CAE) metric is the cumulative sum of the *absolute value* of this plot, and is thus robust to error cancellation.

boostrapped confidence intervals on 32-sample subsets of the generated samples. This uncertainty estimate is well-calibrated, correlating strongly with the actual AHFE error (Figure 4a). This procedure can be thought of as analogous to estimating uncertainties by running multiple MD replicas, but with considerably lower computational cost. Such estimates can be used to guide downstream decisions, such as on which compounds to run MD refinement or perform physical experiments. Getting similar estimates from black-box predictive models is more challenging, typically requiring training expensive deep ensembles (Rahaman et al., 2021) or auxiliary confidence models (Jumper et al., 2021; Rhodes et al., 2025).

Compound generalization Figure 4b shows that there is no meaningful correlation between a compound's similarity to the training set (measured by maximum Tanimoto similarity to any compound in the training set) and the resulting AquaGen AHFE absolute error. We hypothesize that since the model is highly regularized by learning to imitate the conformational ensemble induced by classical force-field simulations, the bulk of its capacity is presumably spent modeling fine-grained details that generalize over compound structures.

Error cancellation. Inspection of the errors in the first off-diagonal of the MBAR free-energy matrix $\Delta\hat{G}$ (Figure 4c) shows that AquaGen tends to produce negative ΔG errors during the electrostatic annihilation stage ($\lambda_k, k \in [1, 5]$) and positive errors during the van der Waals (VDW) annihilation stage ($\lambda_k, k \in [6, 20]$). This highlights a limitation of the AE metric (equation 11): since the total free energy difference is computed by summing free-energy increments along the first off-diagonal of $\Delta\hat{G}$, a model can achieve a low AHFE AE despite substantial local errors if those errors cancel across the alchemical pathway. Consequently, AHFE AE alone does not distinguish between models that accurately capture the underlying physics and those that benefit from fortuitous cancellation. To address this issue, we introduce the cumulative absolute error (CAE),

$$\text{CAE} = \sum_{i=1}^{K-1} \left| \Delta\hat{G}^{\text{model}}_{i,i+1} - \Delta\hat{G}^{\text{MD}}_{i,i+1} \right|, \tag{12}$$

which measures the total deviation along the alchemical pathway and is robust to error cancellation. In § 4.5, we use both AE and CAE to evaluate the impact of various design choices. See §C.4 for a more detailed discussion of error cancellation.

Table 1: Comparing the AHFE mean AE of black-box regressors with AquaGen on various held-out test splits which assess both interpolation and extrapolation. Since AquaGen operates at all-atom resolution, we are able to apply MD refinement on the samples. AquaGen (160M) produces samples which, with short 40 ps refinement, achieve < 0.9 kcal/mol AHFE mean AE on various out-of-distribution data splits.

Model	**AHFE target split** extrapolation ↓ (kcal/mol)	**AHFE target split** interpolation ↓ (kcal/mol)	**FreeSolv** ↓ (kcal/mol)	**CombiSolv** ↓ (kcal/mol)
Random Forest (ECFP Fingerprint)	8.29 ± 0.82	2.76 ± 0.66	4.56 ± 0.27	5.25 ± 0.30
GNN (Vacuum Structure)	2.15 ± 0.36	1.34 ± 0.34	3.30 ± 0.28	2.75 ± 0.26
GNN (Solvated Structure)	2.05 ± 0.32	1.00 ± 0.32	3.23 ± 0.16	2.60 ± 0.18
AquaGen (160M)	2.39 ± 0.65	1.24 ± 0.54	3.92 ± 0.55	3.06 ± 0.62
AquaGen (160M) w/ MD refinement (40ps)	0.87 ± 0.26	0.65 ± 0.25	0.40 ± 0.19	0.62 ± 0.17

Table 2: Ablation study on model architecture and sampling choices. Variants to AquaGen Base (40M) are sorted by ascending AHFE Mean AE. Due to error cancellation, AHFE AE correlates only weakly with Overall CAE. Some variants (e.g., centering) improve AE, but at the expense of CAE, indicating less physical samples.

Model Variant	**Electrostatic CAE** ↓ (kcal/mol)	**van der Waals CAE** ↓ (kcal/mol)	**Overall CAE** ↓ (kcal/mol)	**AHFE AE** ↓ (kcal/mol)
AquaGen Base (160M)	2.82 ± 0.42	2.24 ± 0.13	5.04 ± 0.48	1.22 ± 0.31
AquaGen Base (160M) + MD (40ps)	0.43 ± 0.14	0.42 ± 0.12	0.88 ± 0.26	0.46 ± 0.15
AquaGen Base (40M)	3.41 ± 0.46	1.38 ± 0.13	4.76 ± 0.54	2.21 ± 0.45
Autoguidance	2.55 ± 0.38	2.28 ± 0.21	4.74 ± 0.42	1.44 ± 0.30
Compound Centering	5.03 ± 0.61	4.47 ± 0.26	9.45 ± 0.70	1.64 ± 0.44
$\mathcal{N}(0, 1)$ prior	3.36 ± 0.46	1.70 ± 0.19	4.99 ± 0.60	2.50 ± 0.48
Auto-x_0-variance prior	5.63 ± 0.73	1.67 ± 0.19	7.19 ± 0.79	4.17 ± 0.70
Linear timestep integration	2.89 ± 0.80	4.23 ± 1.00	6.92 ± 1.58	6.75 ± 1.63

Comparison to Black-Box Regressors. We compare AquaGen to baselines that produce AHFE estimates in a *black-box* fashion, i.e., directly as the output of a regression task. We consider two regression variants: 1) a random forest (RF) model trained on RDKit Extended Connectivity fingerprints (ECFP), a geometry-agnostic representation of the solvated compound, and 2) a GNN trained on the energy-minimized (vacuum or solvated) configuration of the system of interest. We evaluate models on various data splits designed to test interpolation and extrapolation capabilities based on the target AHFE value (see §C for more details). A 160M AquaGen model outperforms RF and performs comparably with the GNN baselines, despite not being explicitly trained to predict AHFE as the baselines were. Due to the refinability of the predictions, we can run very short (40 ps) MD simulations from the generated samples and obtain significantly better AHFE predictions (< 0.9 kcal/mol error on all splits). We also find that relative to AquaGen (Figure 4b), the vacuum GNN and random forest baselines exhibit a stronger negative correlation between Tanimoto similarity to the training set and AHFE absolute error (Figure 8), suggesting slightly poorer generalization.

4.5 Ablations

We explore the tradeoff between AE and CAE caused by error cancellation, by running various ablations. To keep training costs reasonable, we perform all ablations on the 40M parameter base model, rather than the 160M parameter model on which we reported previous results. A summary of all ablations is provided in Table 2, with variants to the base run sorted by ascending AHFE AE (we still report the unmodified 160M parameter model results in the Table for completeness). We find that AHFE AE correlates only loosely with Overall CAE, suggesting that error cancellation plays a significant role in the final results.

Compound centering. Translating the compound center-of-mass to the center of the simulation box in all MD frames used for training leads to a significant degradation in the CAE of both alchemical legs. We hypothesize that centering the compound creates a strong bias for the model, resulting in an over-repulsion of the solvents around the origin. While this run achieves better AHFE AE than **Base (40M)** due to error cancellation, we opt not to use it due to the very high CAE.

Gaussian prior variance. The Auto-x_0-variance prior is an attempt to reduce the discrepancy between the variance of the prior and target positions x_0 and x_1. We set the prior simulation box length ℓ based on an empirically fitted linear relationship with the square root of number of atoms, $\ell = 0.41\sqrt{N} + 8.59$ (such that the prior simulation box itself is $c_0 = \ell I_3$), and the variance of the prior positions x_0 such that 95% of the probability density is contained within c_0 prior to periodic wrapping. Surprisingly, this yields worse results across the board, with a particularly negative effect on electrostatic-leg CAE. Setting the prior variance position variance too small (i.e. $\mathcal{N}(0, 1)$) creates extreme path crossing and curvature near $t = 0$ during training and integration. We find that a better middle ground of $\sigma^2 = 2$ yields the best results (AquaGen Base). The downside of this choice is that due to the expansion in volume from x_0 to x_1, we induce an overestimation of the box size. See §C.1 for more discussion.

Uniform timestep integration. Replacing the exponential time discretization (§C.1) with a uniform discretization over $t \in [0, 1]$ at inference time leads to a significant degradation in van der Waals CAE, and ultimately a degradation in AHFE AE to 6.75 kcal/mol. This is consistent with our observation that the curvature of the learned marginal velocity field is considerably higher near $t \approx 0$ (Figure 5). We hypothesize that this effect is substantially more pronounced for the van der Waals leg because the model must coordinate the global spatial organization of many solvent molecules extremely early in the integration trajectory, making accurate resolution of the small-t regime particularly important. In contrast, errors in the electrostatic leg appear to arise more locally from imperfect solvent orientation and hydrogen-bond alignment, which accumulate gradually throughout the integration and are less sensitive to the timestep schedule.

Autoguidance. Autoguidance (Karras et al., 2024) replaces the unconditional model typically used in classifier-free-guidance (Ho & Salimans, 2022) with a λ-conditional model from an earlier checkpoint in training. This yields a significant improvement in Electrostatic CAE from 2.58 to 2.00 kcal/mol, but worsens the van der Waals and overall CAE. Despite an improved AHFE AE of 1.44 kcal/mol from error cancellation, we don't observe similar cancellation when applying autoguidance to the 160M parameter model, so we elect not to use autoguidance for the final results. We believe that this is because the 160M parameter model is already estimating electrostatics fairly accurately, making autoguidance somewhat redundant. However, this points to the potential of exploring guidance strategies to counteract certain model biases, and the importance of improving the expressivity of λ conditioning.

5 Related Work

Biomolecular generative models. Folding models (Jumper et al., 2021; Abramson et al., 2024) have induced a paradigm shift in protein structure prediction. More recently, generative modeling of atomistic Boltzmann distributions has seen rapid progress through normalizing flows (Noé et al., 2019; Kim et al., 2024), diffusion models (Jing et al., 2024b; Lewis et al., 2025), and flow matching approaches (Lipman et al., 2023; Hassan et al., 2024). However, many existing approaches operate on reduced-dimensionality manifolds (e.g., torsional angles, internal coordinates, or coarse-grained frames) or on discrete states of a simplified dynamical representation such as a Markov State Model (Kapuśniak et al., 2026). Although solvent effects are typically incorporated in the underlying training data, to our knowledge, no existing model explicitly includes solvent atoms in its generated configurations. This would be a crucial limitation in our gray-box approach of computing AHFE via potential energy evaluations, since the energetic precision of implicit solvent force fields still lags behind that of explicit solvent (Tan et al., 2006), especially for macromolecules (Katkova et al., 2017; Robinson et al., 2016). Hence, explicit water models, such as TIP3P (Price & Brooks, 2004), remain the standard practice in industrial MD workflows (Schindler et al., 2020; Thaler et al., 2026). We contrast the characteristics of recent models with this work (AquaGen) in Table 3.

Traditional free energy methods. Free energy estimation is a classical problem in molecular simulation, with a long history of methods including free energy perturbation (FEP) (Zwanzig, 1954), thermodynamic integration (TI) (Kirkwood, 1935; Frenkel & Smit, 2001), umbrella sampling (Torrie & Valleau, 1977), replica

Table 3: Comparison of recent biomolecular generative models. Our model, AquaGen, is the first to model the Boltzmann conformational ensemble at all-atom resolution, including hydrogens and solvent atoms.

Model	Output Resolution	Objective	Generates Solvent	Systems
Scalable Flows (Kim et al., 2024)	All heavy-atoms	Conformational ensemble	No	Proteins
ESM-Flow (Jing et al., 2024a)	β-Carbon atoms	Conformational ensemble	No	Proteins
AlphaFold 3 (Abramson et al., 2024)	All heavy atoms	Structure prediction	No	Proteins, ligands, nucleic acids
Bio-Emu (Lewis et al., 2025)	Backbone heavy atoms	Conformational ensemble	No	Proteins
Boltz-2 (Passaro et al., 2025)	All heavy atoms	Structure + affinity prediction	No	Proteins, ligands, nucleic acids
MarS-FM (Kapuśniak et al., 2026)	All heavy atoms	Conformational ensemble	No	Proteins
ATMOS (Shi et al., 2026)	All heavy atoms	Dynamics	No	Proteins, protein–ligand complexes
AquaGen (this work)	**All-atom**	**Conformational ensemble**	**Yes**	**Solvated compounds**

exchange (Hukushima & Nemoto, 1996), Bennett acceptance ratio (BAR) (Bennett, 1976), and multistate Bennett acceptance ratio (MBAR) (Shirts & Chodera, 2008). These approaches are statistically principled and remain standard in computational chemistry, but they rely on extensive MD sampling, often across many intermediate alchemical or thermodynamic states. As a result, their practical cost is dominated by the need to generate sufficiently decorrelated samples with adequate phase-space overlap between neighboring states.

Learned free energy estimation. A separate line of work has explored using generative models and learned transport maps to accelerate free energy calculations. Neural Thermodynamic Integration (Mate et al., 2024) replaces a hand-designed alchemical path with a trainable neural Hamiltonian, enabling thermodynamic integration through learned intermediate ensembles, with follow-up work applying this idea to solvation free energy estimation (Máté et al., 2025). FEAT (He et al., 2025) is another complementary direction, using adaptive learned transports to construct free energy estimators based on non-equilibrium identities such as the escorted Jarzynski equality and controlled Crooks relations. In general, the scalability of these methods has not been demonstrated beyond low-dimensional settings. Recent biomolecular models such as BioEmu (Lewis et al., 2025), AlphaFold3 (Abramson et al., 2024), and Boltz-2 (Passaro et al., 2025) have also been extended toward conformational ensemble generation and binding affinity prediction. However, these approaches generally do not produce explicit-solvent Boltzmann samples suitable for direct potential-energy evaluation and MBAR reweighting.

6 Conclusions

In this work, we introduced AquaGen, an atomistic generative model that efficiently samples from the Boltzmann distribution of solvated drug compounds, in the all-atom ($\sim 10^3$ atoms), explicit-solvent, PBC-aware setting. This enables the computation of absolute hydration free energy in a *gray-box* fashion. Although the samples are drawn from a *black-box* ML model, they are inspectable, and their energies are computed with physical force fields, making the ensemble AHFE predictions of this framework grounded in the underlying physics. While there is still room for improvement, particularly in the expressivity of λ conditioning and the accuracy of box volume estimation, we have demonstrated benefits from self-consistent error cancellation and train- and test-time computation to provide accurate free energy estimates, and the ability to easily extract calibrated uncertainties from the model.

Future work. In future work, we hope to further demonstrate the benefits of our gray-box approach and of modeling biophysical problems at a level of resolution immediately compatible with potential energies used in standard MD. To this end, we plan to improve our modeling of rigid water and improve the expressivity of conditioning on the alchemical parameter λ. Finetuning or test-time search strategies (Domingo-Enrich et al., 2024; Singhal et al., 2025) aimed at regions of high distributional overlap or towards alignment with experimentally obtained free energies (Lewis et al., 2025) could also be fruitful directions. Ultimately, we hope to leverage the proposed framework to tackle more problems for which MD is a classical solution in real-world industrial settings. Solving problems such as the prediction of lipophilicity (logP), membrane permeability, or absolute binding free energy (ABFE) –which are significantly more challenging due to their scale, system heterogeneity, and pipeline complexity–would be transformative for drug discovery.

References

Josh Abramson, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, Lindsay Willmore, Andrew J Ballard, Joshua Bambrick, et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature*, 630(8016):493–500, 2024.

Walid Ahmad, Elana Simon, Seyone Chithrananda, Gabriel Grand, and Bharath Ramsundar. ChemBERTa-2: Towards chemical foundation models. *arXiv preprint arXiv:2209.01712*, 2022.

Irfan Alibay, Richard J Gowers, David W H Swenson, Michael M Henry, Benjamin Ries, Hannah M Baumann, James R B Eastwood, Ashley Mitchell, David Dotson, Joshua T Horton, Matthew Thompson, and Alyssa Travitz. The open free energy library, 2026.

Charles H Bennett. Efficient estimation of free energy differences from Monte Carlo data. *Journal of Computational Physics*, 22(2):245–268, 1976.

Seyone Chithrananda, Gabriel Grand, and Bharath Ramsundar. ChemBERTa: Large-scale self-supervised pretraining for molecular property prediction. *arXiv preprint arXiv:2010.09885*, 2020.

Zoe Cournia, Bryce Allen, and Woody Sherman. Relative binding free energy calculations in drug discovery: Recent advances and practical considerations. *Journal of chemical information and modeling*, 57(12): 2911–2937, 2017.

Carles Domingo-Enrich, Michal Drozdzal, Brian Karrer, and Ricky TQ Chen. Adjoint matching: Fine-tuning flow and diffusion generative models with memoryless stochastic optimal control. *arXiv preprint arXiv:2409.08861*, 2024.

Daan Frenkel and Berend Smit. *Understanding Molecular Simulation: From Algorithms to Applications*. Academic Press, 2001.

Michael K Gilson and Huan-Xiang Zhou. Calculation of protein-ligand binding affinities. *Annu. Rev. Biophys. Biomol. Struct.*, 36(1):21–42, 2007.

Laura M Grubbs, Mariam Saifullah, Nohelli E De La Rosa, Shulin Ye, Sai S Achi, William E Acree, Jr, and Michael H Abraham. Mathematical correlations for describing solute transfer into functionalized alkane solvents containing hydroxyl, ether, ester or ketone solvents. *Fluid Phase Equilib.*, 298(1):48–53, November 2010.

Majdi Hassan, Nikhil Shenoy, Jungyoon Lee, Hannes Stärk, Stephan Thaler, and Dominique Beaini. ET-Flow: Equivariant flow-matching for molecular conformer generation. *Advances in Neural Information Processing Systems*, 37:128798–128824, 2024.

Jiajun He, Yuanqi Du, Francisco Vargas, Yuanqing Wang, Carla P Gomes, José Miguel Hernández-Lobato, and Eric Vanden-Eijnden. FEAT: Free energy estimators with adaptive transport. *arXiv preprint arXiv:2504.11516*, 2025.

Esther Heid, Kevin P Greenman, Yunsie Chung, Shih-Cheng Li, David E Graff, Florence H Vermeire, Haoyang Wu, William H Green, and Charles J McGill. Chemprop: A machine learning package for chemical property prediction. *Journal of chemical information and modeling*, 64(1):9–17, 2023.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

Koji Hukushima and Koji Nemoto. Exchange Monte Carlo method and application to spin glass simulations. *Journal of the Physical Society of Japan*, 65(6):1604–1608, 1996.

Bowen Jing, Bonnie Berger, and Tommi Jaakkola. AlphaFold meets flow matching for generating protein ensembles. *arXiv preprint arXiv:2402.04845*, 2024a.

Bowen Jing, Hannes Stärk, Tommi Jaakkola, and Bonnie Berger. Generative modeling of molecular dynamics trajectories. *Advances in Neural Information Processing Systems*, 37:40534–40564, 2024b.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with AlphaFold. *nature*, 596(7873):583–589, 2021.

Kacper Kapuśniak, Cristian Gabellini, Michael Bronstein, Prudencio Tossou, and Francesco Di Giovanni. MarS-FM: Generative modeling of molecular dynamics via markov state models. In *International Conference on Learning Representations (ICLR)*, 2026.

Tero Karras, Miika Aittala, Tuomas Kynkäänniemi, Jaakko Lehtinen, Timo Aila, and Samuli Laine. Guiding a diffusion model with a bad version of itself. *Advances in Neural Information Processing Systems*, 37: 52996–53021, 2024.

E V Katkova, A V Onufriev, B Aguilar, and V B Sulimov. Accuracy comparison of several common implicit solvent models and their implementations in the context of protein-ligand binding. *J. Mol. Graph. Model.*, 72:70–80, March 2017.

Joseph C Kim, David Bloore, Karan Kapoor, Jun Feng, Ming-Hong Hao, and Mengdi Wang. Scalable normalizing flows enable boltzmann generators for macromolecules. *arXiv preprint arXiv:2401.04246*, 2024.

John G Kirkwood. Statistical mechanics of fluid mixtures. *The Journal of chemical physics*, 3(5):300–313, 1935.

Sarah Lewis, Tim Hempel, José Jiménez-Luna, Michael Gastegger, Yu Xie, Andrew YK Foong, Victor García Satorras, Osama Abdin, Bastiaan S Veeling, Iryna Zaporozhets, et al. Scalable emulation of protein equilibrium ensembles with generative deep learning. *Science*, 389(6761):eadv9817, 2025.

Kresten Lindorff-Larsen, Stefano Piana, Ron O Dror, and David E Shaw. How fast-folding proteins fold. *Science*, 334(6055):517–520, 2011.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *International Conference on Learning Representations*, 2023.

Aleksandr V Marenich, Casey P Kelly, Jason D Thompson, Gregory D Hawkins, Candy C Chambers, David J Giesen, Paul Winget, Christopher J Cramer, and Donald G Truhlar. Minnesota solvation database (MNSOL) version 2012, 2020.

Balint Mate, Francois Fleuret, and Tristan Bereau. Neural thermodynamic integration: Free energies from energy-based diffusion models. *The Journal of Physical Chemistry Letters*, 15(45):11395–11404, 2024.

Bálint Máté, François Fleuret, and Tristan Bereau. Solvation free energies from neural thermodynamic integration. *The Journal of Chemical Physics*, 162(12), 2025.

David L Mobley and J Peter Guthrie. FreeSolv: A database of experimental and calculated hydration free energies, with input files. *Journal of computer-aided molecular design*, 28(7):711–720, 2014.

Edouard Moine, Romain Privat, Baptiste Sirjean, and Jean-Noël Jaubert. Estimation of solvation quantities from experimental thermodynamic data: Development of the comprehensive CompSol databank for pure and mixed solutes. *J. Phys. Chem. Ref. Data*, 46(3):033102, September 2017.

Ingo Muegge and Yuan Hu. Recent advances in alchemical binding free energy calculations for drug discovery. *ACS Medicinal Chemistry Letters*, 14(3):244–250, 2023.

Frank Noé, Simon Olsson, Jonas Köhler, and Hao Wu. Boltzmann generators: Sampling equilibrium states of many-body systems with deep learning. *Science*, 365(6457):eaaw1147, 2019.

Saro Passaro, Gabriele Corso, Jeremy Wohlwend, Mateo Reveiz, Stephan Thaler, Vignesh Ram Somnath, Noah Getz, Tally Portnoi, Julien Roy, Hannes Stark, et al. Boltz-2: Towards accurate and efficient binding affinity prediction. *BioRxiv*, 2025.

Daniel J Price and Charles L Brooks, 3rd. A modified TIP3P water potential for simulation with ewald summation. *J. Chem. Phys.*, 121(20):10096–10103, November 2004.

Rahul Rahaman et al. Uncertainty quantification and deep ensembles. *Advances in neural information processing systems*, 34:20063–20075, 2021.

Benjamin Rhodes, Sander Vandenhaute, Vaidotas Šimkus, James Gin, Jonathan Godwin, Tim Duignan, and Mark Neumann. Orb-v3: Atomistic simulation at scale. *arXiv preprint arXiv:2504.06231*, 2025.

Melina K Robinson, Jacob I Monroe, and M Scott Shell. Are AMBER force fields and implicit solvation models additive? a folding study with a balanced peptide test set. *J. Chem. Theory Comput.*, 12(11):5631–5642, November 2016.

David Rogers and Mathew Hahn. Extended-connectivity fingerprints. *Journal of chemical information and modeling*, 50(5):742–754, 2010.

Christina E M Schindler, Hannah Baumann, Andreas Blum, Dietrich Böse, Hans-Peter Buchstaller, Lars Burgdorf, Daniel Cappel, Eugene Chekler, Paul Czodrowski, Dieter Dorsch, Merveille K I Eguida, Bruce Follows, Thomas Fuchß, Ulrich Grädler, Jakub Gunera, Theresa Johnson, Catherine Jorand Lebrun, Srinivasa Karra, Markus Klein, Tim Knehans, Lisa Koetzner, Mireille Krier, Matthias Leiendecker, Birgitta Leuthner, Liwei Li, Igor Mochalkin, Djordje Musil, Constantin Neagu, Friedrich Rippmann, Kai Schiemann, Robert Schulz, Thomas Steinbrecher, Eva-Maria Tanzer, Andrea Unzue Lopez, Ariele Viacava Follis, Ansgar Wegener, and Daniel Kuhn. Large-scale assessment of binding free energy calculations in active drug discovery projects. *J. Chem. Inf. Model.*, 60(11):5457–5474, November 2020.

David E Shaw, Martin M Deneroff, Ron O Dror, Jeffrey S Kuskin, Richard H Larson, John K Salmon, Cliff Young, Brannon Batson, Kevin J Bowers, Jack C Chao, et al. Anton, a special-purpose machine for molecular dynamics simulation. *Communications of the ACM*, 51(7):91–97, 2008.

David E Shaw, Paul Maragakis, Kresten Lindorff-Larsen, Stefano Piana, Ron O Dror, Michael P Eastwood, Joseph A Bank, John M Jumper, John K Salmon, Yibing Shan, et al. Atomic-level characterization of the structural dynamics of proteins. *Science*, 330(6002):341–346, 2010.

Liang Shi, Jiarui Lu, Junqi Liu, Chence Shi, Zhi Yang, and Jian Tang. Atomic trajectory modeling with state space models for biomolecular dynamics. *arXiv preprint arXiv:2603.17633*, 2026.

Michael R Shirts and John D Chodera. Statistically optimal analysis of samples from multiple equilibrium states. *The Journal of Chemical Physics*, 129(12):124105, 2008.

Michael R Shirts, David L Mobley, and John D Chodera. Alchemical free energy calculations: Ready for prime time? *Annual reports in computational chemistry*, 3:41–59, 2007.

Raghav Singhal, Zachary Horvitz, Ryan Teehan, Mengye Ren, Zhou Yu, Kathleen McKeown, and Rajesh Ranganath. A general framework for inference-time scaling and steering of diffusion models. *arXiv preprint arXiv:2501.06848*, 2025.

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*, 110(37):18680–18687, 2006.

Stephan Thaler, Zhiyi Wu, William G Glass, Richard T Bradshaw, Gail Bartlett, Prudencio Tossou, and Geoffrey PF Wood. Boltz-ABFE: Free energy perturbation without crystal structures. *Journal of Chemical Theory and Computation*, 22(4):1823–1833, 2026.

Glenn M Torrie and John P Valleau. Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling. *Journal of computational physics*, 23(2):187–199, 1977.

Florence H Vermeire and William H Green. Transfer learning for solvation free energies: From quantum chemistry to experiments. *Chemical Engineering Journal*, 418:129307, 2021.

Lingle Wang, Yujie Wu, Yuqing Deng, Byungchan Kim, Levi Pierce, Goran Krilov, Dmitry Lupyan, Shaughnessy Robinson, Markus K Dahlgren, Jeremy Greenwood, et al. Accurate and reliable prediction of relative ligand binding potency in prospective drug discovery by way of a modern free-energy calculation protocol and force field. *Journal of the American Chemical Society*, 137(7):2695–2703, 2015.

Robert W Zwanzig. High-temperature equation of state by a perturbation method. I. nonpolar gases. *The Journal of Chemical Physics*, 22(8):1420–1426, 1954.

A Model Details

A.1 Model Architecture

We parameterize the flow-matching velocity field $v_\theta(\bar{x}_\tau, \tau|\lambda)$ with a graph neural network (GNN) operating on the joint system state $\bar{x}_\tau = \{x_\tau, c_\tau\}$, where $x_\tau \in \mathbb{R}^{N\times 3}$ denotes atomic coordinates and c_τ denotes the periodic simulation cell represented via virtual nodes. Given atomic coordinates x_τ, we construct a graph

$$\mathcal{G} = (\mathcal{V}, \mathcal{E}_{\text{bond}} \cup \mathcal{E}_{\text{dist}}), \tag{13}$$

where $\mathcal{E}_{\text{bond}}$ contains covalent bond edges and $\mathcal{E}_{\text{dist}}$ contains geometric k-nearest-neighbor edges computed under periodic boundary conditions using the minimum-image convention. Node features include standard chemical descriptors such as atom type, charge, and force-field features. Edge features include bond order and relative positional information. The model operates on relative displacement vectors

$$\Delta x_{ij} = x_{\tau,j} - x_{\tau,i}, \tag{14}$$

rather than absolute coordinates. The network is conditioned on both the flow-matching time τ and alchemical parameter λ.

Periodic boundary condition representation. To represent the periodic simulation cell c_τ, we introduce virtual nodes corresponding to the shape of the simulation box, which together fully specify the box geometry. These virtual nodes' velocities are jointly predicted as part of the learned vector field $v_\theta(\bar{x}_\tau, \tau|\lambda)$.

In the Gaussian prior distribution $p_0(\bar{x})$, the virtual node coordinates are initialized such that the corresponding simulation box contains approximately 95% of the atomic coordinates sampled from the coordinate prior. For our chosen coordinate prior $x_0 \sim \mathcal{N}(0, 2I_{3N})$, this corresponds to a mean box length of $\mu_c = 7.84$ and a variance of $\sigma_c^2 = 0.05$.

Water compression. To reduce the cost of message passing in solvent-dominated systems, we compress each water molecule into a single latent node during the GNN backbone. We retain the oxygen atom as the representative message-passing node and remove the two associated hydrogen nodes from the graph. The local water geometry is encoded, using the two O–H displacement vectors. and the resulting feature h_{water} is attached as an additional latent feature to the retained oxygen node. Message passing is then performed on the compressed graph, in which non-water atoms and water oxygens remain explicit, but water hydrogens are omitted.

Let z_i^{cmp} denote the output latent representation produced by the message-passing backbone for node i in the compressed graph. For a retained water oxygen O, the corresponding compressed-graph output z_O^{cmp} is used both as the oxygen representation and to reconstruct latent representations for the two omitted hydrogens. In particular, we apply learned projection networks to produce

$$\left(z_{H_1}^{\text{exp}}, z_{H_2}^{\text{exp}}\right) = \phi_{\text{proj}}\left(z_O^{\text{cmp}}\right), \tag{15}$$

and set $z_O^{\text{exp}} = z_O^{\text{cmp}}$. Here z_i^{exp} denotes the reconstructed explicit-atom latent representation after decompressing the water molecules. The reconstructed hydrogen representations are then inserted back into the original atom ordering, yielding an explicit latent representation $z^{\text{exp}} \in \mathbb{N} \times$ over all N atoms. The prediction head is applied to these latent features to produce per-atom velocities.

Since the majority of atoms in solvated systems are water hydrogens, this reduces the number of solvent nodes passed through the backbone from three per water molecule to one per water molecule. This corresponds to an overall FLOP reduction of roughly $\frac{2}{3}$ in water-dominated systems, with minimal loss in accuracy for sufficiently large models.

Model scaling. We generally find that scaling model width, depth, and message-passing horizon produce comparable improvements, provided that nodes have sufficient receptive field. In practice, increasing width is typically the most computationally efficient scaling direction.

B Molecular Dynamics Dataset

Internal molecules in the training dataset were selected as products of medicinal-chemistry lead optimization, and only uncharged compounds were retained to avoid complications associated with ionization-state and

finite-size corrections. To ensure methodological consistency across datasets, the reference AHFE values for both FreeSolv and CombiSolv were recomputed using the same protocol applied to the internal compounds.

All calculations were performed with the OpenFE absolute solvation workflow, which implements an absolute hydration free energy thermodynamic cycle in which the ligand is transformed in both solvent and vacuum. In this protocol, electrostatic interactions are fully annihilated first, followed by Lennard-Jones decoupling with intramolecular Lennard-Jones interactions preserved while removing intermolecular van der Waals interactions using Gapsys's soft-core potential. Amino acids were parameterized using AMBER ff14SB force field. Small molecules were parameterized using the OpenFF 2.1.1 force field with AM1-BCC charges. For the solvent leg, we used the TIP3P water model. We used a Middle Langevin integrator and performed equilibration and production in the NPT ensemble with a Monte Carlo barostat.

The alchemical path was discretized into 20 λ values with schedules:

$$\lambda_{\text{elec}} = [0.0, 0.25, 0.5, 0.75, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0] \tag{16}$$

$$\lambda_{\text{vdW}} = [0, 0, 0, 0, 0, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.65, 0.7, 0.75, 0.8, 0.85, 0.9, 0.95, 1.0] \tag{17}$$

Electrostatics were turned off over the first five windows, after which van der Waals interactions were gradually decoupled across the remaining windows, consistent with the OpenFE AHFE strategy. Sampling was enhanced using Hamiltonian Replica Exchange (HREX). For each compound, we ran 3 independent replicas, each for 10 ns, using HREX at 298.15 K and 1 bar. A 4 fs timestep was used, enabled by hydrogen mass repartitioning (3.0 amu). Using a sampling frequency of 1 ps, 10,000 frames were extracted from production trajectories across each compound, replica, and λ window. We trained on over 2000 compounds, yielding approximately over 1 billion training frames. Hydration free energies, ΔG_{AHFE}, were obtained by applying the MBAR estimator as implemented in PyMBAR 4.0.

C Additional Results and Details

C.1 Curvature of learned velocity field.

We find that our choice of Gaussian prior leads to non-uniform curvature of the learned velocity field over the flow matching time interval $\tau \in [0, 1]$. As a proxy for curvature, we measure the discretized *angular velocity* between two adjacent timesteps during sampling via the arccos of the cosine similarity between consecutive $v_\theta(x, \tau|\lambda)$:

$$v_t = v_\theta(x_t, t|\lambda) \tag{18}$$

$$\omega_\tau = \frac{v_\tau \cdot v_{\tau+\Delta\tau}}{\|v_\tau\|\|v_{\tau+\Delta\tau}\|\Delta\tau} \tag{19}$$

As shown in Figure 5, the angular velocity is considerably higher near $t = 0$ for the Gaussian priors with a fixed variance of either 1 or 2. We hypothesize that this is due to the change in scale from the prior to the target distribution. This motivates the use of an exponential timestep schedule during sampling which employs a finer discretization near $t = 0$. Specifically, we define a normalized auxiliary variable $u \in [0, 1]$ with uniform spacing and map it to the integration schedule

$$t(u) = \frac{\exp(\alpha u) - 1}{\exp(\alpha) - 1}, \tag{20}$$

where $\alpha > 0$ controls the concentration of timesteps near $t = 0$. Larger values of α allocate more integration steps near the prior distribution, where the learned vector field exhibits higher curvature. In all experiments, we use $\alpha = 4$.

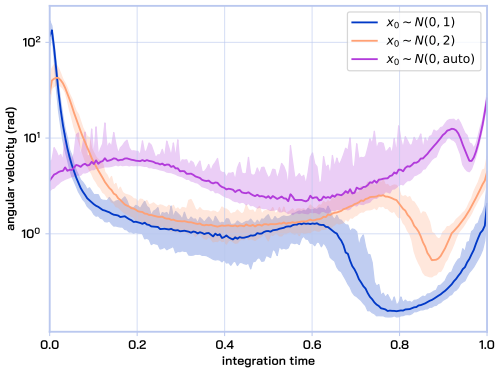

Figure 5: **Angular velocity of** $v_\theta(x, \tau|\lambda)$ during integration. The large velocity near $t = 0$ can be counteracted by an exponential integration schedule for the $N(0, 1)$ and $N(0, 2)$ models. The auto-variance prior reduces the angular velocity near $t = 0$, but leads to higher angular velocities near $t = 1$ and higher overall CAE values.

The auto-variance prior has lower angular velocity at $t = 0$ but higher angular velocity at $t = 1$. We tried counteracting this with various integration schedules (e.g., a reverse exponential which allocates more timesteps near $t = 1$), but were unable to achieve competitive results with that of the base model which uses a Gaussian prior with a fixed variance of 2.

C.2 Additional tICA plots.

We show tICA plots for more randomly selected compounds from the test set in Figure 6. The reference tICA is fit on MD samples across all λ values, using as features all pairwise Euclidean distances (in Å) among the ligand heavy atoms, in addition to the 20 water oxygens nearest to the ligand.

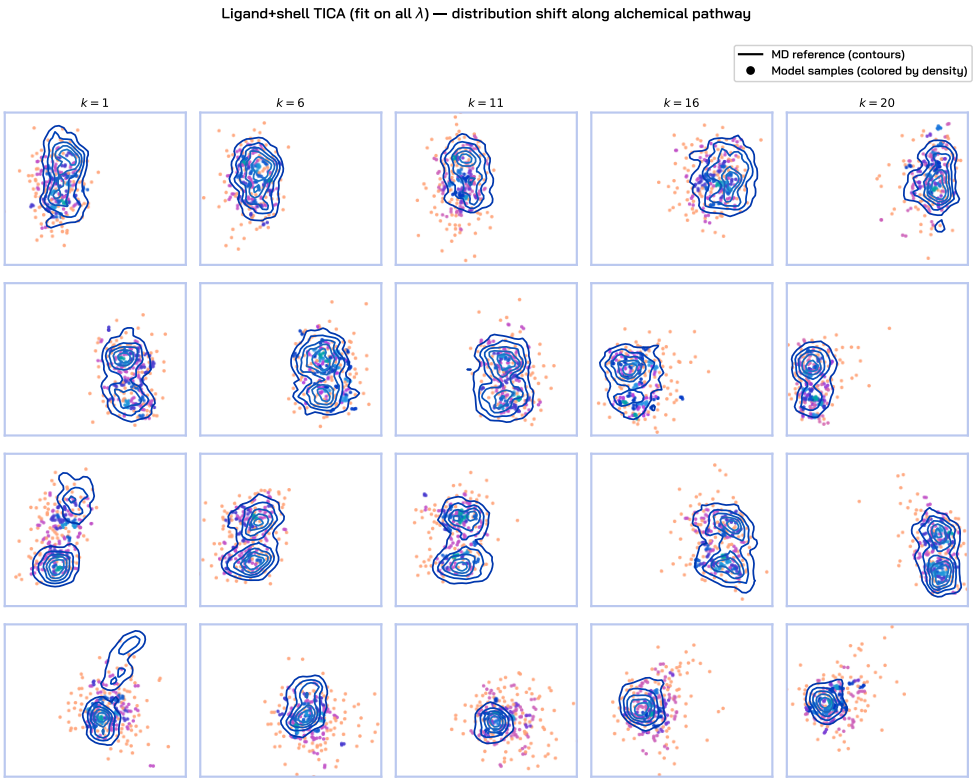

Figure 6: **tICA decomposition** over the alchemical trajectories. AquaGen samples track the change in principal components as λ evolves.

C.3 Additional energy decompositions at fully-interacting distribution.

In Figure 7, we show more granular energy decompositions of samples generated by a 40M and 160M parameter AquaGen model trained only the fully-interacting endpoint ($\lambda_k, k = 1$). Due to the lack of a rigid water assumption in our generative models, we overestimate the solvent bonded energy contribution (second row from bottom) relative to the reference MD samples. We also tend to underestimate the solute bonded contribution, suggesting minor mode collapse which causes excessive stiffness around average bond lengths. In general, scaling the model from 40M to 160M parameters has the largest positive effect on solvent nonbonded energies, which are not modulated by λ. This suggests that future work should prioritize allocating model capacity towards parts of the system which are more consequential for the downstream application, i.e., AHFE estimation.

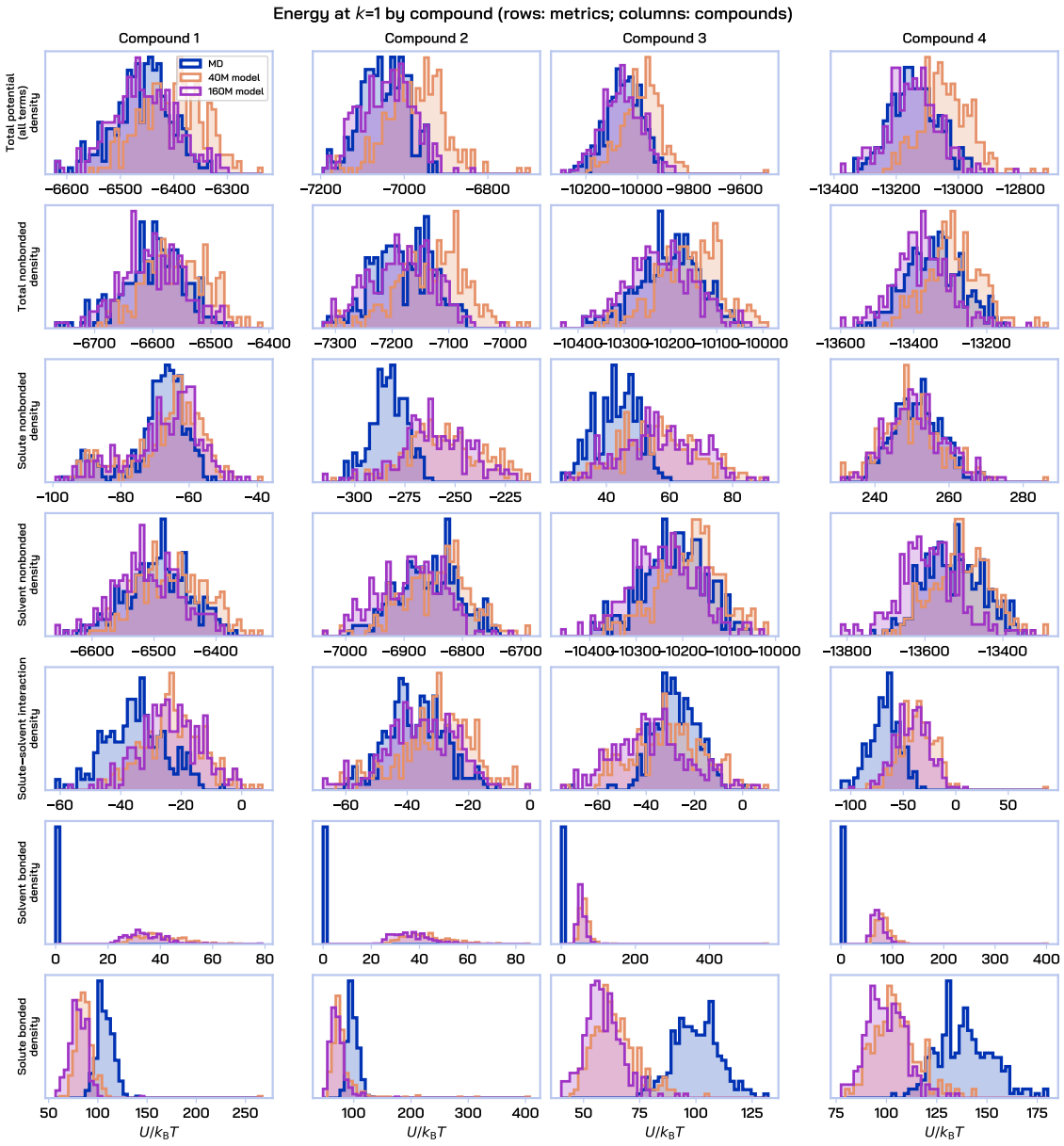

Figure 7: **Energy decompositions at the fully-interacting endpoint ($\lambda_k, k = 1$).** For each compound and energy component, we plot the reference MD distribution (gray), the distribution from the 40M AquaGen model (red), and the distributon from the 160M AquaGen model (blue).

C.4 Error cancellation along the alchemical pathway

The free energy difference between the alchemical endpoints can be written as a sum over the first off-diagonal of the MBAR free-energy matrix:

$$\Delta\hat{G}_{0,K} = \sum_{i=1}^{K-1} \Delta\hat{G}_{i,i+1}. \tag{21}$$

Figure 4c provides additional insight into model behavior along this pathway. The region to the left of the dashed line corresponds to electrostatic annihilation ($\lambda_k, k \in [1, 5]$), while the region to the right corresponds to van der Waals annihilation ($\lambda_k, k \in [6, 20]$). We observe a consistent trend in which the model produces negative ΔG errors in the electrostatic regime and positive ΔG errors in the VDW regime.

One possible explanation is the structure of the alchemical schedule. Electrostatic interactions are annihilated over only four windows, whereas VDW interactions are annihilated over fifteen windows. Consequently, adjacent Boltzmann distributions differ more strongly in the electrostatic regime. The observed pattern suggests that the model learns a smoother, averaged evolution across λ, effectively underestimating the rate of change in the electrostatic region and overestimating it in the VDW region.

The AHFE absolute error can be written as

$$\text{AHFE AE} = \left|\Delta\hat{G}^{\text{model}}_{1,K} - \Delta\hat{G}^{\text{MD}}_{1,K}\right| \tag{22}$$

$$= \left|\sum_{i=1}^{K-1}\left(\Delta\hat{G}^{\text{model}}_{i,i+1} - \Delta\hat{G}^{\text{MD}}_{i,i+1}\right)\right|. \tag{23}$$

Under this interpretation, the signed area under the error curve in Figure 4c corresponds to the final AHFE error. Positive and negative deviations can therefore cancel, making it possible for a model with large local free-energy errors to nevertheless achieve a small AHFE AE.

To quantify these local deviations directly, we define the cumulative absolute error (CAE):

$$\text{CAE} = \sum_{i=1}^{K-1}\left|\Delta\hat{G}^{\text{model}}_{i,i+1} - \Delta\hat{G}^{\text{MD}}_{i,i+1}\right|. \tag{24}$$

By the triangle inequality,

$$\text{AHFE AE} \leq \text{CAE}, \tag{25}$$

with equality only when all pairwise error terms have the same sign. CAE therefore provides an upper bound on AE that is insensitive to error cancellation and more directly measures inaccuracies in the modeled overlap between adjacent alchemical distributions.

C.5 Regression baselines.

We provide additional details on the regression baselines to which we compare AquaGen in Table 1:

1. **Random Forest (ECFP Fingerprint):** As a simple, structure-agnostic baseline, we train a random forest model to regress to the ground truth AHFE value given the Extended Connectivity Fingerprint (ECFP) descriptor (Rogers & Hahn, 2010) of the ligand.
2. **GNN:** We train a GNN (with the same architecture as AquaGen) to map from the minimal energy configuration of each ligand to the ground truth AHFE value. We consider two variants: 1) providing only the ligand structure (**GNN (Vacuum Structure)**) and 2) providing both the ligand and water structure (**GNN (Solvated Structure)**). This baseline is in the spirit of recent works which integrate binding affinity prediction as an auxiliary regression task along with structure prediction (Passaro et al., 2025).

Baselines are trained on all compounds from the AquaGen training set, excluding compounds with true AHFE values in the bottom or top 5% of the original training data (for fair comparison, the AquaGen models reported in Table 1 are also re-trained on the same compounds). All methods are then evaluated on four test splits:

1. **Target split of internal dataset - interpolation (100 compounds):** Compounds in the test set whose true AHFE lies within the middle 90% of values in the original AquaGen training set .
2. **Target split of internal dataset- extrapolation (224 compounds):** Compounds in the test set whose true AHFE lies in the bottom or top 5% of values in the original AquaGen training set .
3. **FreeSolv (Mobley & Guthrie, 2014) (589 compounds**): A database of neutral fragment-like compounds, spanning a wide range of molecular weights (around 16 - 499 Daltons) and polarities.
4. **CombiSolv (Vermeire & Green, 2021) (575 compounds)**: A multi-solvent solvation energy mega-database that was curated from multiple sources, including FreeSolv (Mobley & Guthrie, 2014),

MNSol (Marenich et al., 2020), CompSol (Moine et al., 2017), and the Abraham dataset (Grubbs et al., 2010). Here we used only compounds solvated with water, and compounds originating from FreeSolv were excluded .

Unlike our internal dataset, which is composed of drug-like molecules (median molecular weight of >300 g/mol, median heavy atom count >20), FreeSolv and CombiSolv are mostly comprised of fragment-sized, low-molecular-weight organic solutes (median MW of 120.6 g/mol and 184.3 g/mol, respectively) (relevant statistics are summarized in Table 4).

Table 4: Comparison of molecular size statistics across the FreeSolv, CombiSolv, and internal dataset.

Metric	**FreeSolv (FS)**	**CombiSolv (CS)**	**Internal Dataset**
Total Compounds	589	575	thousands
Median MW (g/mol)	120.6	184.3	>300
Median Heavy Atom Count	8.0	13.0	>20
95th Percentile MW	295.0	362.5	>500

Figure 8 plots the correlation between AHFE Mean AE and Tanimoto similarity for the regression baselines. The baselines exhibit a stronger negative correlation than AquaGen (Figure 4b), suggesting they may be less reliable in out-of-distribution settings.

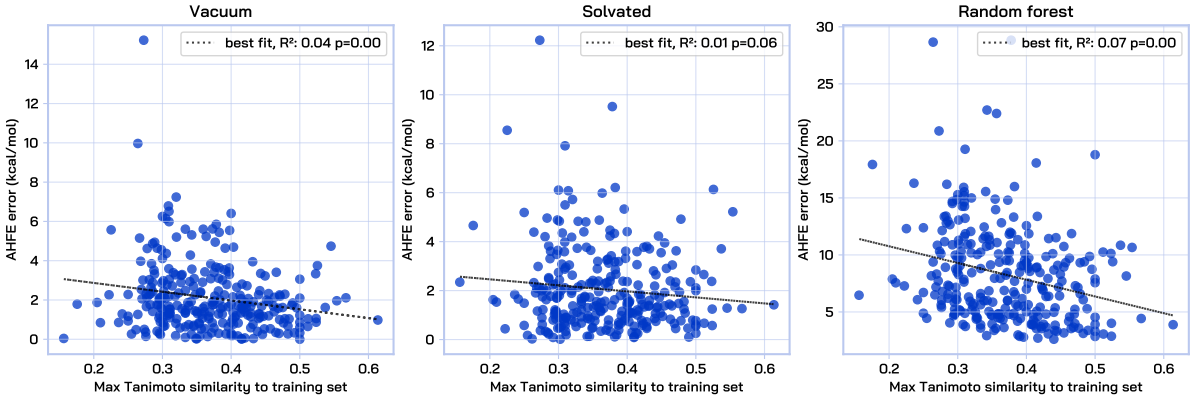

Figure 8: **Baseline AHFE Mean AE vs similarity to the training set for baselines**. The weak, negative trend suggests that baselines succeed closer to their training set.