Boltz-1 as a force field – why co-folding models struggle with learning physics and how to fix it

Anonymous Authors[1]

Abstract

Modern co-folding models have achieved remarkable success in biomolecular structure prediction. However, their ability to generalise to out-of-distribution examples and to capture physical laws remains limited. These limitations may stem from either data or architecture; here, we focus on the latter by examining whether the training objectives and architectural choices of co-folding models hinder the learning of physical laws. Drawing on insights from physics-constrained Machine Learning Interatomic Potentials (MLIPs), we investigate the expressiveness of attention-based modules as implemented in the co-folding models. We evaluate the exemplary co-folding model Boltz-1 as an MLIP and find that it underperforms on energy surface learning. Our analysis shows that accurate energy learning requires inter-atomic distances to be encoded appropriately in the attention pair bias, whereas Boltz-1 constructs these features in a way that fails to support this task. We further identify Boltz-1's strong reliance on pairformer features as an additional limitation in this context, even though this reliance might be beneficial in the structure prediction task. Based on these insights from MLIPs, we introduce simple architectural modifications, including a revised pair bias encoding, and show that they significantly improve energy landscape learning.

1. Introduction

Modern day biomolecular structure prediction models (also known as co-folding models) such as AlphaFold 3 (Abramson et al., 2024), Boltz-1 and 2 (Wohlwend et al., 2025; Passaro et al., 2025), or Chai (Discovery et al., 2024), have achieved great success in recent years and are becoming a basis for commercial drug design pipelines. Though very useful in many cases of interest, it has been shown that their performance drops as the similarity of the query biomolecule to the training set decreases (Škrinjar et al., 2025), suggesting that patterns learnt from the training data do not suffice for full generalisation. The most general patterns – the universal laws of physics that govern biomolecular formation – remain to be captured. For example, an investigation by (Masters et al., 2024) shows a ligand-binding case in which AlphaFold 3 (AF3) fails to account for a single mutation in the sequence, and its predictions violate physical laws. Similarly, Boltz-1 predicts structures that violate the realistic bond lengths and angles, and requires inference-time correction that guides the design towards a minimum-energy state (Wohlwend et al., 2025).

As for any deep learning model, these failure modes may stem from either data or model architecture. In this work, we focus on the latter, and investigate whether the specific architectural choices in co-folding models hinder learning of physical laws. Interestingly, the problem of optimal architecture for learning the biomolecular energy surface has been extensively studied in the field of Machine Learning Interatomic Potentials (MLIP). However, progress in MLIP has diverged from that in structure prediction models, and conclusions from MLIP cast doubt on whether the current design principles in structure prediction models allow for efficient learning of physical laws. Co-folding models such as AF3 and its replicas are designed to operate on coarse-grained tokens for the majority of the layers, and a large portion of their parameters reside in the trunk processing evolutionary information. While geometric information is used in multiple parts of the models, the atom attention encoder plays a central role in transforming atom coordinates into representations that can inform whether a generated structure lies in a high or low-energy region. Therefore, it remains unclear how well the representations learned by this module transfer to modelling physical energy and forces.

In this work, we address the following question: *are the attention based modules used in modern co-folding models expressive enough to learn the energy surface with high accuracy, and which contemporary architecture design choices make this task easier or more difficult?* To answer this,

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Submitted to the 2026 Workshop on Generative and Agentic AI for Biology (ICML 2026). Do not distribute.

we study transformer modules operating on atomic coordinates that incorporate elements from both structure prediction models and MLIPs, and analyse how individual design choices affect the learning dynamics and final performance. In particular, we design an attention-based MLIP with an architecture similar to the AF3 atom attention encoder, which allows us to identify the encoder components that are essential for strong MLIP performance. We further show how insights from specialised, physics-constrained MLIPs can benefit contemporary co-folding model design in terms of energy accuracy. In particular, we demonstrate that a simple yet effective Radial Basis Function featurisation of inter-atomic distances improves the performance of a physics-unconstrained atom encoder.

We then evaluate one of the state-of-the-art co-folding models, Boltz-1, when trained explicitly as an MLIP, in order to understand the limitations of its architecture for energy learning. We show that the original Boltz-1 architecture underperforms at learning the potential energy surface, and that this is linked to the way pair bias features are initialised and atomic coordinates are processed. Guided by our preceding analysis, we show that simple architectural modifications significantly improve the performance of Boltz-1 as an MLIP. We hypothesise that these modifications may help current co-folding models learn more faithful physics-based representations.

Our contributions are as follows:

- An MLIP-based analysis of attention module design in co-folding models operating on atomic coordinates.
- Evidence that simple MLIP-inspired distance featurisation improves energy learning in physics-unconstrained atom encoders.
- An analysis of Boltz-1 as an MLIP. We show how Boltz-1 performance on potential energy surface learning task can be significantly improved without fundamental changes to its overall design.

2. Related work

Machine-learned interatomic potentials (MLIPs) have been proposed to emulate expensive quantum mechanical calculations at a fraction of the cost by learning the potential energy surface of a molecule or material from data. Recent progress has been driven by $\mathrm{SE}(3)$ equivariant architectures that encode local molecular structure via intermediate features that transform according to a group representation of $\mathrm{SE}(3)$, such as the PaiNN(Schütt et al., 2021) architecture, NequIP (Batzner et al., 2022), Allegro (Musaelian et al., 2023), MACE (Batatia et al., 2022), TensorNet (Simeon & De Fabritiis, 2023), or ViSNet (Wang et al., 2024b). Modelling biomolecules like proteins or RNA has been particularly challenging, since their dynamics is largely governed by entropic effects, solvent interaction and long-range interactions, which are difficult to capture with local message-passing architectures, and because the corresponding system sizes and timescales prohibit the use of expensive large models. While equivariant MLIPs based on fragmentation or hybrid approaches such as AI^{2}BMD (Wang et al., 2024a) and GEMS (Unke et al., 2024) have been proposed to address these challenges, classical force fields – with tabulated or machine learned parameters as in Grappa (Seute et al., 2025) – are still widely used.

Boltz-1 is more similar to unconstrained MLIPs that have been proposed recently. Brehmer et al. (2025) compare equivariant models vs non equivariant transformer-based models, and conclude equivariant architectures outperform non-equivariant, however, their transformers do not use pair bias features informative of coordinates, and it is unclear if their conclusions generalise to other datasets. Point Edge Transformer (Pozdnyakov & Ceriotti, 2023) is an unconstrained transformer that performs multi-head attention over individual neighbourhoods centred on each atom, effectively creating a unique set of keys per query atom. TransIP (Elhag et al., 2025) shows how enforcing equivariance in latent space of a non-equivariant transformer MLIP outperforms augmentation based approach. (Bigi et al., 2026) shows how scaling up of the Point Edge Transformer makes it match the predictive accuracy of equivariant models.

In this work, we aim to study the potential of co-folding models for predicting the potential energy surface of proteins. We will suggest several architectural modifications of Boltz-1 – a non-equivariant de-noising architecture – in order to increase its performance for energy prediction, which, as we argue, can guide folding model architecture development towards more physically relevant representations.

3. SPICE dataset

For both Section 4 and Section 5 we use the SPICE dataset (Eastman et al., 2024) – a collection of quantum mechanical data for ensembles of molecules – that was designed with learning potential functions for small molecule protein interactions in mind. We choose this dataset because it resembles the task of protein-ligand binding that co-folding models must solve: at least a basic understanding of the energy surface is required to find appropriate ligand poses. We choose dipeptides and amino-acid–ligand pairs from the SPICE dataset: dipeptides represent interatomic forces between protein residues that a co-folding model should respect, and amino-acid–ligand pairs represent ligand-protein interactions that are crucial for pose prediction. For each molecule in the dataset, ensembles of multiple high and low energy conformers per molecule were generated. Each conformer is annotated with its energy and per-atom forces.

3.1. Train, validation, test

The dipeptides group of SPICE contains 676 dipeptides with ACE and NME caps. We randomly sample 50 dipeptides to form the validation set. The SPICE amino-acid–ligand pair group contains 79,967 pairs. From these, we first remove pairs with only a single conformation, and then remove pairs for which the amino acid could not be matched to the reference conformer. This leaves 43,930 pairs. We further exclude all pairs containing a ligand that appears in the Boltz test split PDBs, resulting in 43,844 pairs, which we use to construct the training and validation splits. The validation split contains all pairs whose ligands appear in the PDB entry from the Boltz validation split, together with all pairs involving 500 additional randomly sampled ligands. This yields 41,602 pairs in the training split and 2,242 in the validation split. Our test split consists of 86 amino-acid–ligand pairs that share a ligand with at least one PDB entry in the Boltz test split.

3.2. Batching for SPICE data

Because of our formulation of the energy regression (models are trained to predict the energy difference to the mean energy of a molecule – details in Section 4.1), batches with SPICE molecules need to contain many conformers of the same molecule. We construct amino-acid–ligand batches by sequentially loading entire ensemble of a molecule until the max batch size is reached. For dipeptides, we always take 32 conformers per molecule, and then batch molecules together until max batch size is reached.

3.3. Reference conformer matching

Boltz-1 was originally trained with reference conformers for amino acids and ligands, encoding local structural information and connectivity. We hence match the SPICE structure to one of the reference conformer templates used for training. Since the corresponding CCD codes are not deposited for the SPICE datasets, we match the residues and ligands to their template using a search algorithm that compares the molecular graphs and identifies candidate molecule and template if their graphs are isomorphic. Due to modifications such as protonation states in amino-acids or ligands that were missing from the set of templates, it was not always possible to match the SPICE molecule to one of the templates. For those molecules, we simply mask template information.

4. Atom encoder as MLIP study

In this Section we analyse how specific inductive biases or specific components of transformer MLIP affect its performance and data efficiency. Our non-equivariant transformer is based on the self-attention mechanisms and blends in elements from physically constrained MLIP. Ultimately, we are interested in understanding whether the particular kind of atom attention present in structure prediction models is expressive enough to be an efficient MLIP, and which components of the architecture make the training most efficient. For this reason, our transformer module (Algorithm 1) incorporates feature projections and other components present in the atom attention encoder from AF3, which are also found in multitude of AF3 inspired architectures. Atom attention encoder is a module which embeds chemical features and atomic coordinates in the noisy sample as per-atom features, and it uses reference conformer geometry as inductive bias for the pair features. If implemented as a part of the structure module, it also broadcasts single and pair embedding from the pairformer to update per atom single and pair embeddings. Then it consequently updates atom features in a series of pair bias attention+conditional transition blocks. For more details we refer to the AF3 Supplementary Material (Abramson et al., 2024). In this work, we investigate the following components:

- The dimension of pair bias z. We compare the commonly used 16 to a larger 128.
- Encoding of pairwise distances in pair bias: inverse distance versus radial basis function (RBF, which is a simple and effective encoding common in constrained MLIP (Simeon & Fabritiis, 2023; Schütt et al., 2017)).
- Projection of pairwise displacements into pair bias features, which breaks pair bias equivariance. We call it 'edge coord project' in Algorithm 1.
- The presence of cutoff radius for attention mechanism. Constrained MLIPs introduce cutoff radius to avoid going OOM, but its introduction has extra long-range interactions filtering effect.
- The presence of conditional transition block after attention layer. As in AF3 encoder, it scales the atom embeddings by initial chemical features and pairformer embeddings.

We compare the attention based module to TensorNet (Simeon & Fabritiis, 2023): an exemplary physics constrained model equivariant by design. During message passing, messages are decomposed into irreducible representations that transform under the scalar, vector, or tensor representation of the Euclidean group $\mathrm{SE}(3)$. Relying on equivariance, TensorNet achieved state-of-the-art results with less parameters than baselines.

We use the implementation of TensorNet from PhysicsML (Exscientia, 2023) and apply the default training parameters fine-tuned for SPICE dataset in the original publication (cutoff 10Å, 64 RBF, 3.7M parameters). We also re-use TensorNet RBF embeddings module for all our experiments that require RBF encoding. Our Boltz-1 inspired atom encoder has 3 layers, hidden dimension 256, 8 heads and is trained with diffusion multiplicity 8. Models with $z = 16$ had 3.6M

Algorithm 1 Attention module as MLIP. Components switched on/off are coloured in blue.

Require: Molecular features $\mathcal{F} = \{\mathbf{e}, \mathbf{c}, \mathbf{x}, \mathbf{m}\}$ where $\mathbf{e} \in \mathbb{R}^{bM\times n\times d_e}$ element features, $\mathbf{c} \in \mathbb{R}^{bM\times n}$ charges, $\mathbf{x} \in \mathbb{R}^{bM\times n\times 3}$ coordinates, $\mathbf{m} \in \{0,1\}^{bM\times n}$ atom mask. M - augmentation multiplicity.

$\mathbf{h}_{\text{chem}} \leftarrow \text{Linear}([\mathbf{e}, \mathbf{c}]) \in \mathbb{R}^{bM\times n\times d_s}$
$\mathbf{M}_{\text{pair}} \leftarrow \mathbf{m} \otimes \mathbf{m}^T \in \{0,1\}^{bM\times n\times n}$
$\mathbf{r}_{bij} \leftarrow \mathbf{x}_{b,i} - \mathbf{x}_{b,j} \in \mathbb{R}^{bM\times n\times n\times 3}$
$d_{ij} \leftarrow \|\mathbf{r}_{ij}\|_2 \in \mathbb{R}^{bM\times n\times n\times 1}$

if RBF **then**
  $\mathbf{z}_+ = \text{LinearNoBias}(\text{RBF}(d_{ij}))$
else
  $\mathbf{z}_+ = \text{LinearNoBias}(1/(1+d_{ij}))$
end if
if edge coord project **then**
  $\mathbf{z} \leftarrow \text{LinearNoBias}(\mathbf{r}_{ij}) \odot \mathbf{M}_{\text{pair}} + \mathbf{z}_+$
else
  $\mathbf{z} \leftarrow \mathbf{z}_+$
end if

$\mathbf{z} \leftarrow \mathbf{z} \odot \mathbf{M}_{\text{pair}}$
$\mathbf{z} \leftarrow \text{LinearNoBias}(\mathbf{z})$

 # Chemical features into attention bias
$\mathbf{z_{ij}} \leftarrow \mathbf{z_{ij}} + \text{LinearNoBias}(\text{ReLU}(\mathbf{h}_{\text{chem}}))_i + \text{LinearNoBias}(\text{ReLU}(\mathbf{h}_{\text{chem}}))_j$
$\mathbf{z} \leftarrow \text{MLP}(\mathbf{z})$

 # Init atom features
$\mathbf{q} \leftarrow \mathbf{h}_{\text{chem}} + \text{LinearNoBias}(\mathbf{x})$

if cutoff attention enabled **then**
  $\mathbf{M}_{\text{dist}} \leftarrow (d_{ij} < r_c) \wedge \neg \mathbf{I}_n$
else
  $\mathbf{M}_{\text{dist}} \leftarrow \text{None}$
end if

for $\ell = 1, \ldots, L$ **do**
  $\Delta\mathbf{q} \leftarrow \text{Attention}_\ell(\mathbf{q}, \mathbf{z}, \mathbf{m}, \mathbf{M}_{\text{dist}})$
  if conditional transition **then**
    $\mathbf{q} \leftarrow \text{CondTransition}_\ell(\mathbf{q}, \mathbf{h}_{\text{chem}})$
  end if
  $\mathbf{q} \leftarrow \mathbf{q} + \Delta\mathbf{q}$
end for

 # Energy prediction
$\hat{E} \leftarrow \text{LinearNoBias}(\mathbf{q})$
return $\hat{E}$

parameters, and with $z = 128$ 3.7M parameters. While this might seem small in comparison to other MLIPs (Bigi et al., 2026), it is still larger than the atom attention encoder in AlphaFold3, which reasons over atomic coordinates and which casts atom embeddings into coarse grained tokens (AF3 atom attention encoder + decoder have 2.3M parameters). Since we are ultimately interested in the architecture maximally comparable to the current co-folding attention modules, we deliberately keep it this size.

4.1. Training methodology

We train two sets of models on dipeptides and amino-acid–ligand pairs separately. We investigate the data efficiency and convergence speed on two datasets, and finally compare the performance of models trained on amino-acid–ligand pairs on the test set.

Models were trained with starting learning rate 1e-4 achieved during warmup of 500 steps, and learning rate started to reduce after 20K steps by a factor of 0.9 every 4K steps up to minimum 1e-6. Training augmentation multiplicity (AF3-like center random augmentation) was $M = 8$. We used Adam optimiser with $\beta_1 = 0.9$ and $\beta_2 = 0.95$. To retrain TensorNet, we used the same training setup as in the original publication.

We trained with an MSE loss on energies and forces, with weights 0.5 for each. Following previous works on the SPICE dataset (Seute et al., 2025; Wang et al., 2022), we trained the model to predict energy differences between individual states of a given molecule, which can be realised effectively by subtracting the mean energy of the molecule in both reference and prediction. We hence discarded energies of formation since we are not attempting to model bond-breaking chemical reactions. This training objective teaches the model to discriminate between high and low energy states, which bears similarity to ligand pose ranking, for which state-of-the-art computational and experimental methods have an accuracy of around 1kcal/mol (Ross et al., 2023). Force predictions are obtained from the gradient of the predicted energy w.r.t coordinates, $F = -\nabla_x E$, hence the force loss can be interpreted as gradient matching.

4.2. Results

Figure 1 compares the learning dynamics of different model variants in terms of validation energy MAE on amino-acid–ligand pairs and dipeptides, when trained up to the same number of gradient updates. A clear model ranking arises: the most important modifications are RBF embedding and direct projection of coordinates into pair features. Models without direct projection of coordinates into pair features benefit greatly from increasing $z = 16$ to $z = 128$. Cutoff radius of attention does not have significant effect, and surprisingly the condition transition block is essential. Those findings are reflected in the performance on the SPICE amino-acid–ligand test set (Table 1), which evaluates models trained on amino-acid–ligand pairs. The best performing are RBF variants. Still, none of them approaches the performance of the TensorNet which achieves energy MAE 0.333 kcal/mol and force MAE 1.121 kcal/mol/Å. Since our

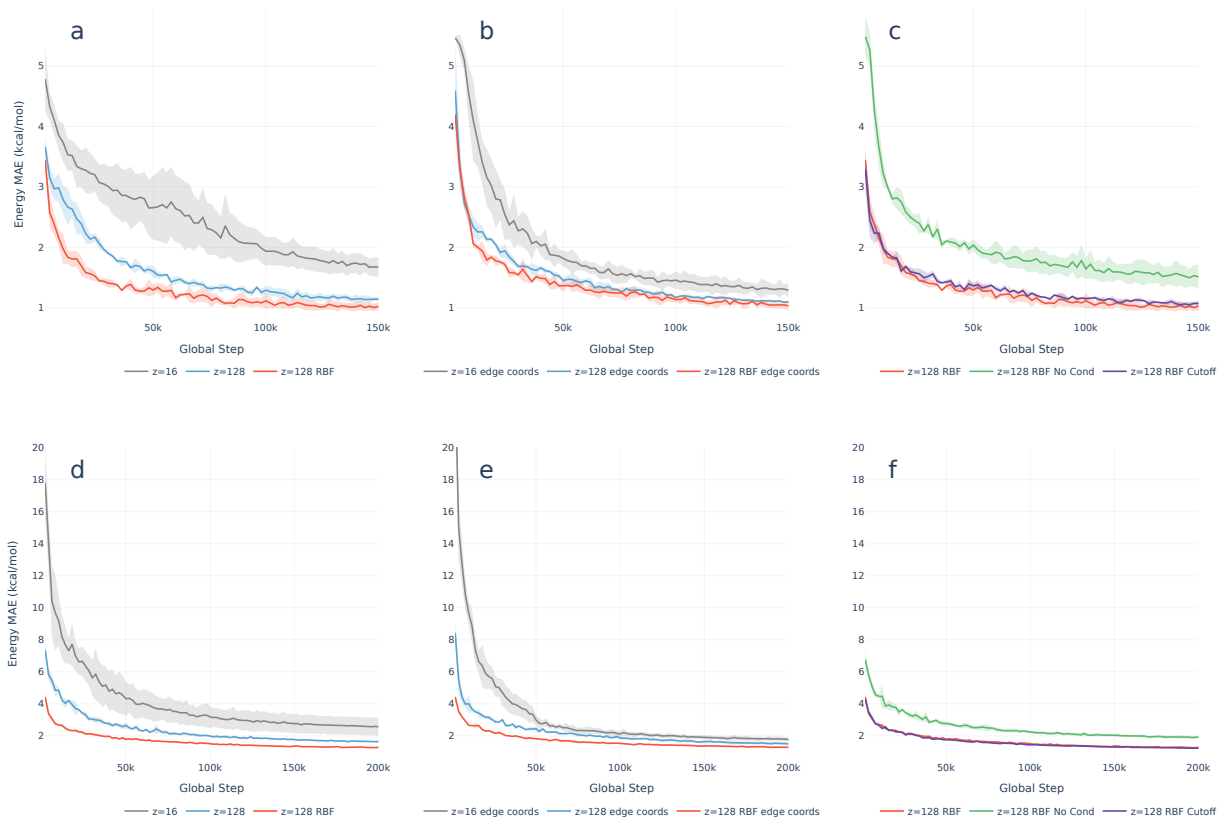

Figure 1. Training dynamics of different models variants on amino-acid–ligands (panels *a,b,c*) and dipeptides (panels *d,e,f*). All models were trained to the same number of gradient updates (150K and 200K for amino-acid–ligand pairs and dipeptides). RBF variants have the quickest and the most stable convergence. It is beneficial to increase $z = 16$ to $z = 128$, even though the direct projection of coordinates into pair features greatly benefits variants $z = 16$.

attention module was designed to test AF3 atom encoder components, this lagging behind the baseline urges to reconsider the current architecture trends in co-folding atom encoder modules. A way to address this can be for example improved distance embedding in pair features, or different allocation of parameters in AF3-like structure module (atom encoder is much smaller than token transformer operating on coarse-grained representations).

5. Boltz as MLIP

In this Section we investigate whether the contemporary design of large structure prediction models allows them to be used as efficient MLIPs. To this end we use Boltz-1, an open-source AF3 replica. It is a suitable model for our case-study, since it was widely adopted by the community (Wohlwend et al., 2024) with its training data being publicly available. Like AF3, it consists of the pairformer for processing evolutionary information and a structure module for coordinates generation, using reference conformers of ligands as initial geometric priors within the pairformer.

Our study in Section 4 showed that attention based MLIPs in most cases can reach the energy prediction accuracy of 1kcal/mol, however, Boltz-1 and other AF3 replicas contain other sizeable modules that might affect the performance as MLIP. For example, in Boltz-1 most of the parameters in the structure module are found in the token transformer (283M out of 294M), which operates on coarse-grained tokens rather than individual atoms. Moreover, the structure module inherits pair features from pairformer, and they encode separations between atoms in the reference conformers instead of those in the sample being generated. The structure module projects sample coordinates r directly via linear layers into per-atom embeddings q. Given the importance of pair bias features revealed in Section 4, it is not clear whether this stark difference in coordinate embedding allows for a reproduction of the results of our carefully designed attention-based MLIP from Section 4. To answer this question and to find potential failure modes, we train Boltz on the energy regression task, starting from its published checkpoint, as described in Section 5.1. Also, drawing on the lesson from Section 4, we check the influence of adding

Table 1. Results on the test set of all model variants. Model variants with RBF embedding show superior performance compared to other variants, and direct projection of coordinates into pair features is particularly important for the variant $z = 16$. Energy is given in kcal/mol, and force is given in kcal/mol/Å. Results are averaged over 3 runs.

z dimension	z16		z128					
model	edge inv dist	edge inv dist + coord project	edge inv dist	edge inv dist + coord project	edge RBF	edge RBF + coord project	edge RBF + cutoff	edge RBF + no cond transition
Energy MAE	1.82 ± 0.19	1.43 ± 0.04	1.28 ± 0.14	1.34 ± 0.05	1.17 ± 0.05	1.12 ± 0.07	1.22 ± 0.03	1.56 ± 0.15
Force MAE	6.32 ± 0.56	4.65 ± 0.35	4.13 ± 0.04	4.14 ± 0.24	3.21 ± 0.14	3.24 ± 0.07	3.45 ± 0.17	4.71 ± 0.20

the RBF encoding of pairwise distances coming from the structure being generated, not from the reference conformers.

5.1. Modifications to predict energies

Boltz was originally trained on a denoising diffusion task, enabling the generation of a biomolecule by performing a number of reversed diffusion steps. The structure module receives intermediate noisy coordinates of the sample r and updates them iteratively until it arrives at fully denoised coordinates r_0. For completeness we provide the algorithm for the Boltz-1 diffusion module in the Appendix C. In the energy prediction task, however, we do not noise the samples before passing them as input: the structure module receives ground truth coordinates r_0, and its output is passed to the newly introduced MLIP head, which returns an energy prediction. The energy prediction head is a two-layer MLP (Algorithm 4 in the Appendix C). It takes as input the atom decoder embeddings returned by the structure module – the very same from which the output denoised coordinates would be predicted in the denoising task. Since backpropagation through the structure module is computationally expensive, models in this Section were trained without force prediction (since forces are derived from the energy gradient, training with forces requires calculating higher order derivates and increases memory usage). Also, during training we updated only the parameters within the structure module, keeping the rest frozen.

We aim to investigate the following aspects of Boltz as MLIP:

- **Reliance on the pairformer embeddings**. Pairformer precedes structure module and it is meant to extract evolutionary information form protein sequences. In the case of SPICE dipeptides and amino-acid–ligand pairs we do not have evolutionary information, however the pairformer still encodes the atomic coordinates in the reference positions. We ablate how much the structure module depends on pairformer embeddings, even when provided r_0 sufficient for the successful energy prediction. We compare two variants of the model, 'ignore pairformer' vs 'with pairformer'. 'With pairformer' does the forward pass through the pairformer, and passes the embeddings to structure module, as would be in the original unmodified architecture. 'Ignore pairformer' initialises pairformer embeddings to zeros and does not update them. For the 'ignore pairformer' variant we are interested in decoupling the performance of atom encoder and atom decoder from the token transformer that emphasizes evolutionary information stored in the coarse-grained tokens, therefore for this model variant we decrease depth of the token transformer from 24 layers to 1.

- **The impact of pair bias features** within the Boltz atom attention encoder for the energy prediction. Atom attention encoder pair bias feature encode information about atom displacements in the reference conformers, and the coordinates r_0 are projected into atom features q. Seeing the importance of pair bias feature in Section 4, we optionally add RBF encoding of pairwise distance in r_0 to pair bias features, and expand their dimension to 128 from 16.

Algorithm 2 shows the atom attention encoder with our modifications. Before training the models for the energy prediction, we pretrain Boltz with the diffusion task to ensure that we keep the pairformer embeddings adjusted to represent both original molecules from Boltz-1's RCSB training dataset, and also dipeptides and amino-acid–ligand data. Pretraining started from the published checkpoint. No modules were frozen at this stage. We sampled batches from both the RCSB Boltz-1 training dataset and the SPICE dataset. Samples from SPICE were noised only partially to cover the last 26 out of total 200 inference steps noise levels. After the pretraining, the RMSD on the validation set for dipeptides and amino-acid–ligand pairs was 0.10 and 0.09 Å (where partial diffusion for dipepetides and amino-acid–ligand pairs started from the noise level of last 26 inference step), indicating successful diffusion pretraining. Details of the diffusion pretraining are in the Appendix B.

For the energy prediction experiments, we load parameters from the newly pretrained checkpoint. For the model variants with increased dimension of pair features p in the structure module, we newly initialise layers that project p.

Algorithm 2 Atom attention encoder. Ablated components in blue.

Require: Reference conformer features $\{\mathbf{f}^*\}$, coordinates $\{\mathbf{r}_l\}$, pairformer token $\{\mathbf{s}_l^{\text{trunk}}\}$ and pair $\{\mathbf{z}_{ij}\}$ features.

 # Embed per-atom data
$\tilde{\mathbf{f}}_l = \text{concat}(\tilde{\mathbf{f}}_l^{\text{pos}}, \tilde{\mathbf{f}}_l^{\text{charge}}, \tilde{\mathbf{f}}_l^{\text{mask}}, \tilde{\mathbf{f}}_l^{\text{ele}}, \tilde{\mathbf{f}}_l^{\text{atom\_name}})$
$\mathbf{c}_l = \text{LinearNoBias}(\tilde{\mathbf{f}}_l)$ $\quad l \in \{1, \ldots, N_{\text{atoms}}\}$

 # Embed offsets between atom reference positions
$\tilde{\mathbf{d}}_{lm} = \tilde{\mathbf{f}}_l^{\text{ref\_pos}} - \tilde{\mathbf{f}}_m^{\text{ref\_pos}}$ $\quad \tilde{\mathbf{d}}_{lm} \in \mathbb{R}^3$

 # Consider distances within the same conformer only
$v_{lm} = (\tilde{f}_l^{\text{ref\_space\_uid}} == \tilde{f}_m^{\text{ref\_space\_uid}})$ $\quad v_{lm} \in \mathbb{R}$
$\mathbf{p}_{lm} = \text{LinearNoBias}(\tilde{\mathbf{d}}_{lm}) \cdot v_{lm}$ $\quad \mathbf{p}_{lm} \in \mathbb{R}^{16}$

 # Embed pairwise inverse squared distances, and the valid mask
$\mathbf{p}_{lm}$ += $\text{LinearNoBias}\left(1/\left(1 + \|\tilde{\mathbf{d}}_{lm}\|^2\right)\right) \cdot v_{lm}$
$\mathbf{p}_{lm}$ += $\text{LinearNoBias}(v_{lm}) \cdot v_{lm}$

 # Initialise the atom single representation as the single conditioning
$\mathbf{q}_l = \mathbf{c}_l$ $\quad \mathbf{q}_l \in \mathbb{R}^{128}$

 # If provided, add trunk embeddings and noisy positions
if $\{\mathbf{r}_l\} \neq \emptyset$ **then**
  if not ignore pairformer **then**
    $\mathbf{c}_l$ += $\text{LinearNoBias}(\text{LayerNorm}(\mathbf{s}^{\text{trunk}}_{\text{tok\_idx}(l)}))$
    $\mathbf{p}_{lm}$ += $\text{LinearNoBias}(\text{LayerNorm}(\mathbf{z}_{\text{tok\_idx}(l)\text{tok\_idx}(m)}))$
  end if
  # Add the noisy positions.
  $\mathbf{q}_l$ += $\text{LinearNoBias}(\mathbf{r}_l)$
end if

 # Add the combined single conditioning to the pair representation
$\mathbf{p}_{lm}$ += $\text{LinearNoBias}(\text{relu}(\mathbf{c}_l)) + \text{LinearNoBias}(\text{relu}(\mathbf{c}_m))$

 # Run a small MLP on the pair activations.
$\mathbf{p}_{lm}$ += $\text{LinearNoBias}(\text{relu}(\text{LinearNoBias}(\text{relu}(\text{LinearNoBias}(\text{relu}(\mathbf{p}_{lm}))))))$

if use RBF **then**
  $\mathbf{r}_{\mathbf{dist,lm}} = r_l - r_m$
  $\mathbf{p}_{lm} = \text{LinearNoBias}(\text{concat}(\mathbf{p}_{\mathbf{lm}}, RBF(\mathbf{r}_{\mathbf{dist,lm}})))$ $\quad \mathbf{p}_{lm} \in \mathbb{R}^{128}$
end if

 # Cross attention transformer
$\{\mathbf{q}_l\} = \text{AtomTransformer}(\{\mathbf{q}_l\}, \{\mathbf{c}_l\}, \{\mathbf{p}_{lm}\}, N_{\text{block}} = 3, N_{\text{head}} = 4)$

 # Aggregate per-atom representation to per-token representation
$\mathbf{a}_i = \underset{\substack{l \in \{1,\ldots,N_{\text{atoms}}\} \\ \text{tok\_idx}(l) = i}}{\text{mean}} (\text{relu}(\text{LinearNoBias}(\mathbf{q}_l)))$ $\quad \mathbf{a}_i \in \mathbb{R}^{c_{\text{token}}}$

$\mathbf{q}_l^{\text{skip}}, \mathbf{c}_l^{\text{skip}}, \mathbf{p}_{lm}^{\text{skip}} = \mathbf{q}_l, \mathbf{c}_l, \mathbf{p}_{lm}$
return $\{\mathbf{a}_i\}, \{\mathbf{q}_l^{\text{skip}}\}, \{\mathbf{c}_l^{\text{skip}}\}, \{\mathbf{p}_{lm}^{\text{skip}}\}$

5.2. Results

Table 2 shows the energy MAE on the amino-acid–ligand test set of all model variants. The correct embedding of distances in r_0 is crucial for successful energy landscape learning: RBF variants outperform no RBF variants by a large margin. Moreover, 'ignore pairformer' + RBF outperforms 'with pairformer' + RBF, which shows that reference conformers encoded in pairformer output have confounding effect for the task of energy prediction. While their inclusion might be beneficial for the structure prediction task, the best energy prediction results are achieved with the

Table 2. Comparison of 'ignore pairformer' variants vs 'with pairformer' in terms of energy MAE in kcal/mol on the amino-acid–ligand pairs test set. Results are averaged over 3 runs. Consistently with the findings from Section 4, RBF variants outperform non-RBF. Moreover, 'ignore pairformer' variant with token transformer reduced achieves lower error than original structure module with pairformer embeddings.

	Ignore PF + RBF	Ignore PF	With PF + RBF	With PF
Energy MAE	0.86 ± 0.10	1.70 ± 0.07	1.68 ± 0.10	2.55 ± 0.16

improved embedding of distances in r. This finding calls for re-thinking the current design trends in AF3 replicas and disentangling how information about reference conformers and r_0 are merged together, for example by keeping separate pair bias features based on reference conformers and r.

6. Discussion

Modern co-folding models have achieved remarkable success in biomolecular structure prediction, but our results show that this success does not automatically translate into accurate learning of the potential energy surface. Through the lens of MLIPs, we examined whether the attention-based modules used in contemporary co-folding models are expressive enough for energy learning, and which architectural choices help or hinder this objective.

Our findings point to a clear conclusion: accurate energy learning depends critically on how geometric information is encoded in the attention mechanism. In particular, interatomic distances must be represented appropriately in the attention pair bias for the model to learn the energy surface effectively. We show that the original Boltz-1 architecture underperforms on this task, and that this underperformance is linked to the way pair bias features are constructed, together with its strong reliance on pairformer features. More broadly, our results suggest that architectural choices that work well for structure prediction are not necessarily well suited to learning physically faithful energy landscapes.

At the same time, our results are encouraging. The limitation is not inherent to attention-based architectures themselves. Rather, simple architectural modifications, motivated by MLIP design principles, substantially improve the ability of Boltz-1 to learn the potential energy surface. In particular, improving the encoding of inter-atomic distances in the pair bias leads to markedly better performance. This suggests that modern co-folding architectures can be made more physics-aware without fundamental changes to their overall design.

More broadly, this work highlights a gap between success in structure prediction and success in capturing the physical laws underlying biomolecular structure. Bridging this gap may be important for improving robustness and out-of-distribution generalisation in future co-folding models, and may matter directly in applications where physical faithfulness is essential, such as mutation sensitivity, conformational ranking, and molecular design.

In summary, we show that contemporary co-folding architectures, exemplified by Boltz-1, underperform as MLIPs not because attention-based models are fundamentally inadequate, but because key geometric information is not encoded in the most effective way. By incorporating simple MLIP-inspired modifications, we significantly improve energy surface learning and provide a concrete route towards more physically grounded co-folding models.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning for biology. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Abramson, J., Adler, J., Dunger, J., et al. Accurate structure prediction of biomolecular interactions with alphafold 3. *Nature*, 630:493–500, 2024. doi: 10.1038/s41586-024-07487-w.

Batatia, I., Kovács, D. P., Simm, G. N. C., Ortner, C., and Csányi, G. Mace: Higher order equivariant message passing neural networks for fast and accurate force fields. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Batzner, S., Musaelian, A., Sun, L., Geiger, M., Mailoa, J. P., Kornbluth, M., Molinari, N., Smidt, T. E., and Kozinsky, B. E(3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials. *Nature Communications*, 13(1):2453, 2022. doi: 10.1038/s41467-022-29939-5.

Bigi, F., Pegolo, P., Mazitov, A., and Ceriotti, M. Pushing the limits of unconstrained machine-learned interatomic potentials, 2026. URL https://arxiv.org/abs/2601.16195.

Brehmer, J., Behrends, S., de Haan, P., and Cohen, T. Does equivariance matter at scale?, 2025. URL https://arxiv.org/abs/2410.23179.

Discovery, C., Boitreaud, J., Dent, J., McPartlon, M., Meier, J., Reis, V., Rogozhnikov, A., and Wu, K. Chai-1: Decoding the molecular interactions of life. *bioRxiv*, 2024. doi: 10.1101/2024.10.10.615955. URL https://www.biorxiv.org/content/early/2024/10/15/2024.10.10.615955.

Eastman, P., Pritchard, B. P., Chodera, J. D., and Markland, T. E. Nutmeg and spice: Models and data for biomolecular machine learning. *Journal of Chemical Theory and Computation*, 20(19):8583–8593, 2024. doi: 10.1021/acs.jctc.4c00794. URL https://doi.org/10.1021/acs.jctc.4c00794. PMID: 39318326.

Elhag, A. A., Raja, A., Morehead, A., Blau, S. M., Morris, G. M., and Bronstein, M. M. Learning inter-atomic potentials without explicit equivariance, 2025. URL https://arxiv.org/abs/2510.00027.

Exscientia. Physicsml, 2023. URL https://exscientia.github.io/physicsml/. Accessed: 2026-04-14.

Masters, M. R., Mahmoud, A. H., and Lill, M. A. Do deep learning models for co-folding learn the physics of protein-ligand interactions? *bioRxiv*, 2024. doi: 10.1101/2024.06.03.597219. URL https://www.biorxiv.org/content/early/2024/06/04/2024.06.03.597219.

Musaelian, A., Batzner, S., Johansson, A., Sun, L., Owen, C. J., Kornbluth, M., and Kozinsky, B. Learning local equivariant representations for large-scale atomistic dynamics. *Nature Communications*, 14(1):579, 2023. doi: 10.1038/s41467-023-36329-y.

Passaro, S., Corso, G., Wohlwend, J., Reveiz, M., Thaler, S., Somnath, V. R., Getz, N., Portnoi, T., Roy, J., Stark, H., Kwabi-Addo, D., Beaini, D., Jaakkola, T., and Barzilay, R. Boltz-2: Towards accurate and efficient binding affinity prediction. *bioRxiv*, 2025. doi: 10.1101/2025.06.14.659707.

Pozdnyakov, S. and Ceriotti, M. Smooth, exact rotational symmetrization for deep learning on point clouds. In Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., and Levine, S. (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 79469–79501. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/fb4a7e3522363907b26a86cc5be627ac-Paper-Conference.pdf.

Ross, G. A., Lu, C., Scarabelli, G., Albanese, S. K., Houang, E., Abel, R., Harder, E. D., and Wang, L. The maximal and current accuracy of rigorous protein-ligand binding free energy calculations. *Communications Chemistry*, 6(1):222, 2023. doi: 10.1038/s42004-023-01019-9. URL https://doi.org/10.1038/s42004-023-01019-9.

Schütt, K., Kindermans, P.-J., Sauceda Felix, H. E., Chmiela, S., Tkatchenko, A., and Müller, K.-R. Schnet: A continuous-filter convolutional neural network for modeling quantum interactions. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/303ed4c69846ab36c2904d3ba8573050-Paper.pdf.

Schütt, K., Unke, O., and Gastegger, M. Equivariant message passing for the prediction of tensorial properties and molecular spectra. In *International conference on machine learning*, pp. 9377–9388. PMLR, 2021.

Seute, L., Hartmann, E., Stühmer, J., and Gräter, F. Grappa – a machine learned molecular mechanics force field. *Chemical Science*, 16(6):2907–2930, 2025. doi: 10.1039/D4SC05465B.

Simeon, G. and De Fabritiis, G. Tensornet: Cartesian tensor representations for efficient learning of molecular potentials. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Simeon, G. and Fabritiis, G. D. Tensornet: Cartesian tensor representations for efficient learning of molecular potentials. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=BEHlPdBZ2e.

Škrinjar, P., Eberhardt, J., Durairaj, J., and Schwede, T. Have protein-ligand co-folding methods moved beyond memorisation? *bioRxiv*, 2025. doi: 10.1101/2025.02.03.636309.

Unke, O. T., Stöhr, M., Ganscha, S., Unterthiner, T., Maennel, H., Kashubin, S., Ahlin, D., Gastegger, M., Medrano Sandonas, L., Berryman, J. T., Tkatchenko, A., and Müller, K.-R. Biomolecular dynamics with machine-learned quantum-mechanical force fields trained on diverse chemical fragments. *Science Advances*, 10(14):eadn4397, 2024. doi: 10.1126/sciadv.adn4397.

Wang, T., He, X., Li, M., Li, Y., Bi, R., Wang, Y., Cheng, C., Shen, X., Meng, J., Zhang, H., Liu, H., Wang, Z., Li, S., Shao, B., and Liu, T.-Y. Ab initio characterization of protein molecular dynamics with ai^{2}bmd. *Nature*, 635:1019–1027, 2024a. doi: 10.1038/s41586-024-08127-z.

Wang, Y., Fass, J., Kaminow, B., Herr, J. E., Rufa, D., Zhang, I., Pulido, I., Henry, M., Bruce Macdonald, H. E., Takaba, K., and Chodera, J. D. End-to-end differentiable construction of molecular mechanics force fields. *Chemical Science*, 13:12016–12033, 2022. doi: 10.1039/D2SC02739A.

Wang, Y., Wang, T., Li, S., He, X., Li, M., Wang, Z., Zheng, N., Shao, B., and Liu, T.-Y. Enhancing geometric representations for molecules with equivariant vector-scalar interactive message passing. *Nature Communications*, 15 (1):313, 2024b. doi: 10.1038/s41467-023-43720-2.

Wohlwend, J., Corso, G., and Passaro, S. Boltz: Official repository for the boltz biomolecular interaction models, 2024. URL https://github.com/jwohlwend/boltz. Accessed: 2026-04-14.

Wohlwend, J., Corso, G., Passaro, S., Getz, N., Reveiz, M., Leidal, K., Swiderski, W., Atkinson, L., Portnoi, T., Chinn, I., Silterra, J., Jaakkola, T., and Barzilay, R. Boltz-1 democratizing biomolecular interaction modeling. *bioRxiv*, 2025. doi: 10.1101/2024.11.19.624167. URL https://www.biorxiv.org/content/early/2025/05/06/2024.11.19.624167.

A. SPICE data input

Dipeptides that consisted of two main residues + ACE and NME caps were processed as a protein type with 4 unknown residues. Hydrogens were explicitly passed, but were not matched to any template atom. We passed the reference molecule formal charge, but we did not pass the SPICE partial charge as a feature. Amino-acids in amino-acid ligand pairs were also capped with ACE and NME and were processed as a protein data type. For ligands in amino-acid pairs, hydrogens were again passed but not matched to any template atom. For both types of data, we used the published Boltz processing script to extract set of features we extract and pass to the model.

B. Training details

B.1. Diffusion pretraining

During the diffusion pretraining, we start from Boltz-1 published checkpoint and continue training on RCSB and SPICE datasets. In order to keep the pass through the model simple, even batches come from SPICE datasets, odd from RCSB. We limit the max number of tokens to 384, which applies only to samples from RCSB since samples from SPICE were not padded to max tokens. Learning rate was 1e-3 with warmup of 100 steps and decay by a factor of 0.8 every 2 epochs, which was started after 8 epochs, down to minimum learning rate 1e-4. Each epoch had 125 optimizer steps. Training was done on 8 NVIDIA A100 80G until convergence. We compare the pretrained model to another model trained in the identical way but on RCSB dataset only. This serves as a sanity check to confirm that inclusions of SPICE dataset does not degrade pairformer embeddings for RCSB samples. We evaluate both models on the RCSB validation set cropped to 384 tokens. There is no significant difference in distogram LDDT between the two models (in Boltz-1, distogram is predicted from pairformer embeddings). except for small reduction in intra RNA distogram LDDT - 0.37 for RCSB only model and 0.28 for RCSB and SPICE datasets model.

B.2. Energy prediction training details

All models were trained on 4 GPUS with batch size per GPU 32 amino acid-ligand pairs, or 32 conformers per dipeptide with two dipeptides in one batch. We sampled batches from amino acid-ligand pairs and dipeptides sets with proportions 0.75 and 0.25 respectively. Models with 1 token transformer layer had warmup of 500 steps, starting learning rate 1e-3, reduced every 2 epochs by a factor of 0.8 which started after 12 epochs down to a minimum of 1e-6, diffusion multiplicity 16. Models with 24 token transformer layers had warmup of 500 steps, starting learning rate 1e-4, reduced every epoch by a factor of 0.8 after 12 epochs down to a minimum of 1e-6, diffusion multiplicity 8. Each epoch had 3000 gradient updates.

C. Algorithms

Algorithm 3 shows Boltz-1 diffusion module (structure module), a reproduction of AF3 diffusion module. x^{noisy} and $\hat{t}$ are the noised atom coordinates and the current noise level during the diffusion training. For our energy prediction experiments, coordinates are not noised, and noise level is 0.

Algorithm 3 Diffusion Module

Require: $\{\mathbf{x}_i^{\text{noisy}}\}, \hat{t}, \{\mathbf{f}^*\}, \{\mathbf{s}_i^{\text{input}}\}, \{\mathbf{s}_m^{\text{trunk}}\}, \{\mathbf{z}_{ij}^{\text{trunk}}\}, \sigma_{\text{data}}$

$\{\mathbf{s}_i\}, \{\mathbf{z}_{ij}\} = \text{DiffusionConditioning}(\hat{t}, \{\mathbf{f}^*\}, \{\mathbf{s}_m^{\text{input}}\}, \{\mathbf{s}_m^{\text{trunk}}\}, \{\mathbf{z}_{ij}^{\text{trunk}}\}, \sigma_{\text{data}})$

 # Scale positions to dimensionless vectors with approximately unit variance.

$\mathbf{r}_l^{\text{noisy}} = \mathbf{x}_l^{\text{noisy}} / \sqrt{\hat{t}^2 + \sigma_{\text{data}}^2}$ $\quad \mathbf{r}_l^{\text{noisy}} \in \mathbb{R}^3$

 # Atom attention and aggregation to coarse-grained tokens.

$\{\mathbf{a}_i\}, \{\mathbf{q}_l^{\text{skip}}\}, \{\mathbf{c}_l^{\text{skip}}\}, \{\mathbf{p}_{lm}^{\text{skip}}\} = \text{AtomAttentionEncoder}(\{\mathbf{f}^*\}, \{\mathbf{r}_l^{\text{noisy}}\}, \{\mathbf{s}_m^{\text{trunk}}\}, \{\mathbf{z}_{ij}\})$ $\quad \mathbf{a}_i \in \mathbb{R}^{c_{\text{token}}}$

$\mathbf{a}_i \mathrel{+}= \text{LinearNoBias}(\text{LayerNorm}(\mathbf{s}_i))$

 # Token transformer. Full self-attention on token level.

$\{\mathbf{a}_i\} \leftarrow \text{DiffusionTransformer}(\{\mathbf{a}_i\}, \{\mathbf{s}_i\}, \{\mathbf{z}_{ij}\}, \beta_{ij} = 0)$

$\mathbf{a}_i \leftarrow \text{LayerNorm}(\mathbf{a}_i)$

 # Broadcast token activations to atoms and run atom attention

$\{\mathbf{r}_l^{\text{update}}\} = \text{AtomAttentionDecoder}(\{\mathbf{a}_i\}, \{\mathbf{q}_l^{\text{skip}}\}, \{\mathbf{c}_l^{\text{skip}}\}, \{\mathbf{p}_{lm}^{\text{skip}}\})$

$\mathbf{x}_l^{\text{out}} = \sigma_{\text{data}}^2 / (\sigma_{\text{data}}^2 + \hat{t}^2) \cdot \mathbf{x}_l^{\text{noisy}} + \sigma_{\text{data}} \cdot \hat{t} / \sqrt{\sigma_{\text{data}}^2 + \hat{t}^2} \cdot \mathbf{r}_l^{\text{update}}$

return $\{\mathbf{x}_l^{\text{out}}\}$

Algorithm 4 shows the energy prediction head. $c_{atom} = 128$ is the Boltz default atom embedding dimension.

Algorithm 4 Energy Readout MLP

Require: $\mathbf{h} \in \mathbb{R}^{c_{\text{atom}}}, c_{atom} = 128$

$\mathbf{h} = \text{Linear}(\mathbf{h})$ $\quad \mathbf{h} \in \mathbb{R}^{2 \cdot c_{\text{hidden}}}$

$\mathbf{h} = \text{LeakyReLU}(\mathbf{h}, \alpha = 0.1)$

$E = \text{LinearNoBias}(\mathbf{h})$ $\quad E \in \mathbb{R}$

return E