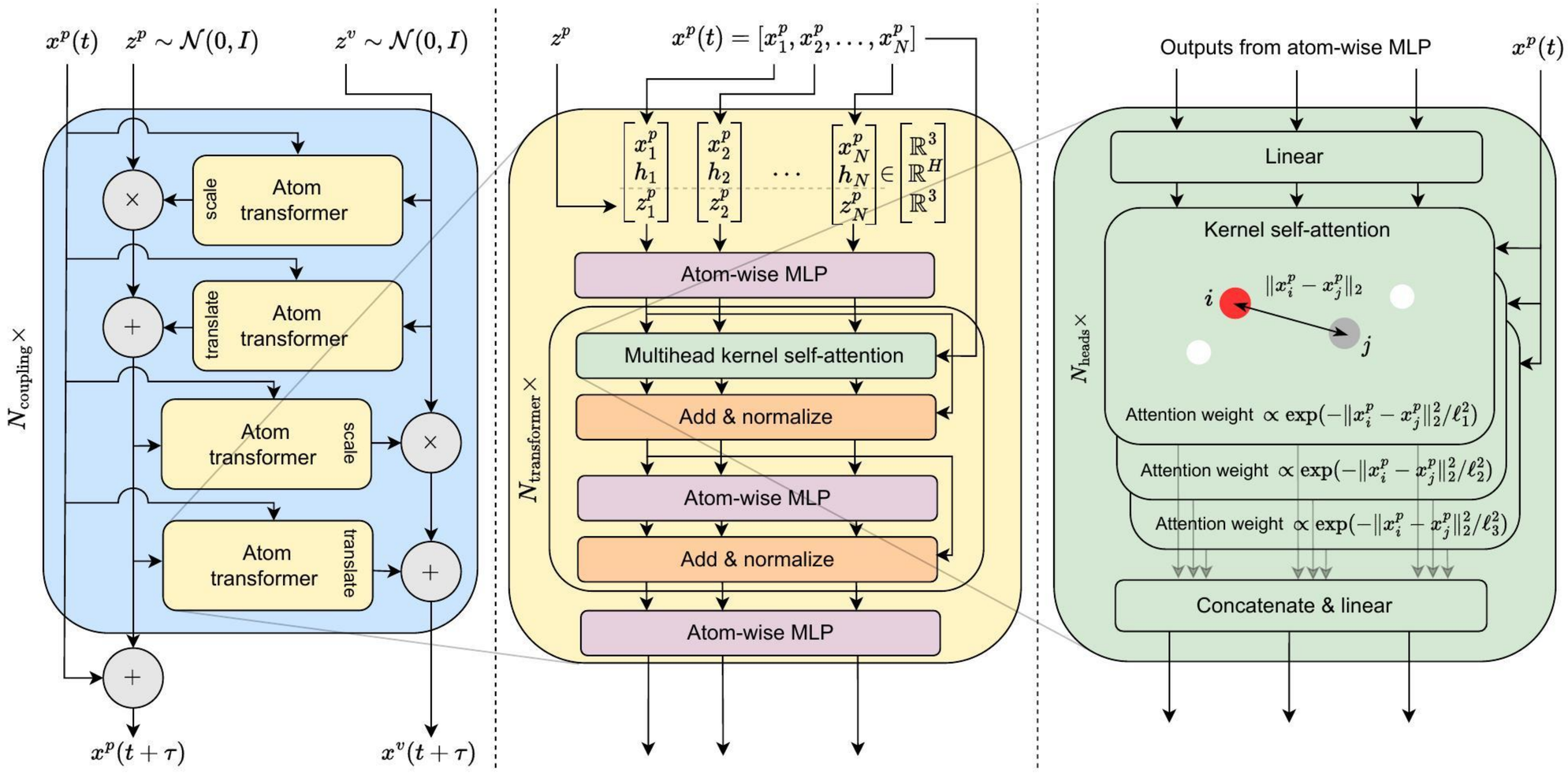

Figure 2. Schematic illustration of the Timewarp conditional flow architecture, described in Section 4. *Left*: A single conditional RealNVP coupling layer. *Middle*: A single atom transformer module. *Right*: the multihead kernel self-attention module.

not use a positional encoding. Second, instead of dot product attention, we use a simple *kernel self-attention* module, which we describe next.

Kernel self-attention We motivate the kernel self-attention module with the observation that physical forces acting on the atoms in a molecule are *local*: *i.e.*, they act more strongly on nearby atoms. Intuitively, for values of τ that are not too large, the positions at time $t+\tau$ will be more influenced by atoms that are nearby at time t, compared to atoms that are far away. Thus, we define the attention weight w_{ij} for atom i attending to atom j as follows:

$$w_{ij} = \frac{\exp(-\|x_i^p - x_j^p\|_2^2/\ell^2)}{\sum_{j'=1}^{N} \exp(-\|x_i^p - x_{j'}^p\|_2^2/\ell^2)}, \tag{11}$$

where ℓ is a lengthscale hyperparameter. The output vectors $\{r_{\text{out},i}\}_{i=1}^N$, given the input vectors $\{r_{\text{in},i}\}_{i=1}^N$, are then:

$$r_{\text{out},i} = \sum_{j=1}^{N} w_{ij} V \cdot r_{\text{in},j}, \tag{12}$$

where $V \in \mathbb{R}^{d_{\text{out}} \times d_{\text{in}}}$ is a learnable matrix. This kernel self-attention is an instance of the RBF kernel attention investigated in Tsai et al. (2019). Similarly to Vaswani et al. (2017), we introduce a *multihead* version of kernel self-attention, where each head has a different lengthscale. This is illustrated in Figure 2, Right. We found that kernel self-attention was significantly faster to compute than dot product attention, and produced similar or improved performance.

4.1. Symmetries

The MD dynamics respects certain physical *symmetries* that would be advantageous to incorporate. We now describe how each of these symmetries is incorporated in Timewarp.

Permutation equivariance Let σ be a permutation of the N atoms. Since the atoms have no intrinsic ordering, the only effect of a permutation of $x(t)$ on the future state $x(t+\tau)$ is to permute the atoms similarly, *i.e.*,

$$\mu(\sigma x(t+\tau)|\sigma x(t)) = \mu(x(t+\tau)|x(t)). \tag{13}$$

Our conditional flow satisfies permutation equivariance exactly. To show this, we use the following proposition proved in Appendix A.1, which is an extension of Köhler et al. (2020); Rezende et al. (2019) for conditional flows:

Proposition 4.1. *Let σ be a symmetry action, and let $f(\cdot;\cdot)$ be an equivariant map such that $f(\sigma z; \sigma x) = \sigma f(z; x)$ for all σ, z, x. Further, let the base distribution $p(z)$ satisfy $p(\sigma z) = p(z)$ for all σ, z. Then the conditional flow defined by $z \sim p(z)$, $x(t+\tau) := f(z; x(t))$ satisfies $p(\sigma x(t+\tau)|\sigma x(t)) = p(x(t+\tau)|x(t))$.*

Our flow satisfies $f_\theta(\sigma z; \sigma x(t)) = \sigma f_\theta(z; x(t))$, since the transformer is permutation equivariant, and permuting z and $x(t)$ together permutes the inputs. Furthermore, the base distribution $p(z) = \mathcal{N}(0, I)$ is permutation invariant. Note that the auxiliary variables allow us to easily construct a