Supplementary Information
Extending the RANGE of Graph Neural Networks: Relaying Attention Nodes for Global Encoding

Alessandro Caruso[†1], Jacopo Venturin[†1,2], Lorenzo Giambagli[§1], Edoardo Rolando[§1], Frank Noé[*3,2,1,5], and Cecilia Clementi[*1,4,5]

[1]Department of Physics, Freie Universität Berlin, *Arnimallee 12*, 14195, Berlin, Germany
[2]Department of Mathematics and Computer Science, Freie Universität Berlin, *Arnimallee 12*, 14195, Berlin, Germany
[3]Microsoft Research AI for Science, *Karl-Liebknecht Str. 32*, 10178, Berlin, Germany
[4]Center for Theoretical Biological Physics, Rice University, Bioscience Research Collaborative, *6500 Main Street*, Houston, 77005, TX, USA
[5]Department of Chemistry, Rice University, *6100 Main Street*, Houston, 77030, TX, USA

[*] Corresponding authors. E-mails: frank.noe@fu-berlin.de, cecilia.clementi@fu-berlin.de
[†] These authors contributed equally.
[§]These authors contributed equally.

Supplementary Note 1: The RANGE architecture

As illustrated in Fig. 1 of the main text, the RANGE architecture combines a local message-passing with an aggregation of all the network nodes into a master node M, followed by a broadcasting that redistributes the collected information back into the single nodes, effectively realizing long-range message-passing. The details on the aggregation and broadcast phases are provided below.

1.1 Aggregation

Since a multi-head attention system is implemented, master nodes funnel information into L d-dimensional spaces: the information stored in each subspace is concatenated into a h-dimensional vector so that $Ld = h$. The aggregated embedding is

$$\mathbf{H}^{(t)} = \sigma\left(\|_{l=1}^{L} \sum_i \hat{\alpha}_i^l A_V^l \tilde{\mathbf{h}}_i^{(t)}\right), \tag{1}$$

where $\mathbf{H}^{(t)} \in \mathbb{R}^h$ is the embedding of M, σ is an element-wise non-linear activation, $\|_{l=1}^{L}$ represents the concatenation operator, $A_V^l : \mathbb{R}^h \to \mathbb{R}^d$ is a learnable matrix, and $\tilde{\mathbf{h}}_i^{(t)}$ refers to the i-th node embedding after a local message-passing iteration. Based on the conventional implementation of additive self-attention[1,2], the weight $\hat{\alpha}_i^l$ of embedding i and head l is defined as:

$$\alpha_i^l = (\mathbf{a}^l)^\top \mathrm{LeakyReLU}(A_Q^l \mathbf{H}^{(t-1)} + A_K^l \tilde{\mathbf{h}}_i^{(t)} + A_E^l \mathbf{E}_i) \tag{2}$$

$$\hat{\alpha}_i^l = \mathrm{Softmax}(\alpha_i^l) = \frac{\exp \alpha_i^l}{\sum_j \exp \alpha_j^l}. \tag{3}$$

Here, $A_Q^l, A_K^l : \mathbb{R}^h \to \mathbb{R}^d$ and $A_E^l : \mathbb{R}^f \to \mathbb{R}^d$ are learnable matrices and $\mathbf{a}^l \in \mathbb{R}^d$ is a learnable vector. The query projection matrices A_Q^l always act on the previous virtual node embedding $\mathbf{H}^{(t-1)}$. The edge features between master node and the graph nodes, denoted as a function of their respective distances $\mathbf{E}_i = \mathrm{RBF}(r_i)$, are carefully designed to extend the standard radial basis expansion and accommodate non-bounded distances without introducing a cutoff. We achieve this by scaling the distances between M and the graph nodes by their maximum

$$r_i = \frac{||\mathbf{x}_i - \mathbf{X}_M||}{\max_j ||\mathbf{x}_j - \mathbf{X}_M||} \quad \in [0, 1], \tag{4}$$

where $\mathbf{x}_i$ denotes the position of node i and $\mathbf{X}_M$ is the position of the master node, $\frac{1}{N}\sum_i \mathbf{x}_i$. The new distances are then transformed into edge features via radial basis expansion. This allows for complete transferability of the trained network across different system sizes.

1.2 Broadcast

In order to update the embeddings of the base graph with the aggregated information while retaining learned short-range interactions, we opted to include self-loops in the attention mechanism as follows:

$$\mathbf{h}_i^{(t+1)} = \mathrm{MLP}\left(\|_{l=1}^{L}\left(\hat{\beta}_{i,\mathrm{self}}^l B_{V,\mathrm{self}}^l \tilde{\mathbf{h}}_i^{(t)} + \hat{\beta}_i^l B_V^l \mathbf{H}^{l(t)}\right)\right), \tag{5}$$

where $B^l_{V,\text{self}} : \mathbb{R}^h \to \mathbb{R}^d$ and $B^l_V : \mathbb{R}^d \to \mathbb{R}^d$; the latter operates on each l-th head representation $\mathbf{H}^{l(t)}$ separately, mantaining their independence. The attention weights are obtained with a slight modification of Eq. (2), by defining

$$\begin{aligned}\beta^l_{i,\text{self}} &= (\mathbf{b}^l)^\top \text{LeakyReLU}(B^l_Q \tilde{\mathbf{h}}^{(t)}_i + B^l_{K,\text{self}} \tilde{\mathbf{h}}^{(t)}_i) \\ \beta^l_i &= (\mathbf{b}^l)^\top \text{LeakyReLU}(B^l_Q \tilde{\mathbf{h}}^{(t)}_i + B^l_K \mathbf{H}^{l(t)} + B^l_E \mathbf{E}_i);\end{aligned} \tag{6}$$

these are then normalized using Softmax, as defined in Eq. (3), to obtain the final attention weights $\hat{\beta}^l_{i,\text{self}}$ and $\hat{\beta}^l_i$. A Multi-layered Perceptron (MLP) mixes the contributions from different heads at the end of the broadcast phase, effectively integrating different classes of non-local interactions. Remarkably, this method enables transfer of information across the system with a computational complexity that scales linearly with the number of nodes in the input graph. This is particularly advantageous when considering predictions on large systems, as it represents an improvement over standard FFT-based methods used for the treatment of long range interactions (e.g. Particle Mesh Ewald in the context of molecular dynamics), whose $N \log N$ scaling might represent a bottleneck during simulations of large molecules. While we considered a single master node in the description above, this design limits the amount of relevant global information that can be aggregated without loss, thereby constraining the scalability of the model. In the following section, we will address this limitation by introducing multiple master nodes, adapting the model to tasks where the number of nodes varies significantly across the dataset.

1.3 Spatial scalability

When several master nodes N_M with indices $I \in \{1 \dots N_M\}$ are employed, each one is initialized with a different embedding $\mathbf{H}^{(0)}_I$, and Eqs. (1) and (2) become, respectively,

$$\mathbf{H}^{(t)}_I = \sigma\left(\|_{l=1}^L \sum_i \alpha^l_{iI} A^l_V \mathbf{h}_i \right) \tag{7}$$

and

$$\alpha^l_{iI} = (\mathbf{a}^l)^\top \text{LeakyReLU}(A^l_Q \mathbf{H}^{(t-1)}_I + A^l_K \mathbf{h}_i + A^l_E \mathbf{E}_{iI}). \tag{8}$$

In this context, the edge features $\mathbf{E}_{iI}$ can be master node-dependent but, in order to maximize parameter sharing without sacrificing performances, the same edge features are allocated for all master nodes. Similarly, the broadcast phase can be generalized to the case of multiple master nodes. Each d-dimensional portion of the output vector $\mathbf{h}^{(t+1)}_i$ can select from multiple global representations, and Eq. (5) and the second of Eq. (6) become, respectively,

$$\mathbf{h}^{(t+1)}_i = \text{MLP}\left(\|_{l=1}^L \left(\hat{\beta}^l_{i,\text{self}} B^l_{V,\text{self}} \tilde{\mathbf{h}}^{(t)}_i + \sum_I \hat{\beta}^l_{iI} B^l_V \mathbf{H}^{l(t)}_I \right) \right) \tag{9}$$

and

$$\beta^l_{iI} = (\mathbf{b}^l)^\top \text{LeakyReLU}(B^l_Q \tilde{\mathbf{h}}^{(t)}_i + B^l_K \mathbf{H}^{l(t)}_I + B^l_E \mathbf{E}_{iI}). \tag{10}$$

After normalizing, a regularization parameter

$$\lambda_I \in \begin{cases} \{1\} & \text{if } I = 1 \\ [0, 1) & \text{if } I > 1, \end{cases} \tag{11}$$

biased on the system size, rescales the contribution from each master node during broadcast by:

$$\Lambda_I(n) = \lambda_I^{\gamma(n)} \tag{12}$$

$$\gamma(n) = (1 + a_I)|\max[0, (1 - n)] + \tanh(b_I)\min[1, n]|. \tag{13}$$

Here, a_I and b_I are positive trainable parameters, and $n = (N - N_{\text{min}})/(N_{\text{max}} - N_{\text{min}})$ is the normalized number of nodes in the graph, with N_{min} and N_{max} being the minimum and maximum number of nodes present in the dataset during training, respectively. While the scalar λ_1 is designed always to ensure at least one fully activated master node, the intensity of all the $\lambda_{I\neq 1}$ is controlled by the factor $\gamma(n)$ as a function of the system size n. Intuitively, $\gamma(n)$ should a) decrease with n, following the intuition that larger molecules need larger capacity per head, and b) always be greater than zero. Given these requirements, we opted for the parametric function in Eq. (13), enforcing $\gamma(n) > 1$ for small molecules and $\gamma(n) < 1$ for large molecules, with the values a_I and b_I controlling this behavior. Finally, the broadcast attention weights are rescaled as follows:

$$\hat{\beta}_{iI}^l \leftarrow \Lambda_I(n)\hat{\beta}_{iI}^l \qquad \text{for } I \in \{1 \dots N_M\}. \tag{14}$$

Approaches as the one delineated in Eq. (14), which aim at regularizing the overall usage of a given node in the trained model, are theoretically motivated[3] and have been proven effective in real word scenarios[4].

1.4 Application to equivariant models

Typically, SE(3)-equivariant MLFFs are designed considering 1) an invariant features representation, and 2) a set of high-order equivariant features; a mixing step is often implemented to exchange information between the two representations[5–8]. The RANGE aggregation and broadcast procedures, as defined in Eq. (1) and Eq. (5), cannot be directly applied to SE(3)-equivariant features due to the presence of nonlinear transformations. In agreement to other designs[9,10], we transfer long-range information via the invariant features and possibly propagate it to the equivariant embeddings via the mixing step in the baseline model. While it is possible to explicitly incorporate higher-order equivariant features in the aggregation-broadcast scheme, this design choice maximizes computational efficiency and enables modularity in RANGE.

Supplementary Note 2: Datasets

All the models reported in the main manuscript have been trained on energies and forces of configurations extracted from the QM7-X[11], AQM[12], and MD22[13] atomic datasets. The labels are calculated at the DFT level of theory, with either the PBE or PBE0 exchange-correlation functional. All datasets include explicit treatment of van der Waals interactions, that are predominantly long-range, via many-body dispersion (MBD)[14–17].

2.1 QM7-X

The QM7-X dataset comprises 42 physicochemical properties calculated for $\sim$ 4.2 millions equilibrium and non-equilibrium structures of organic molecules with up to 23 atoms. These cover the set of elements that is the most predominant in biomolecules, that is H, C, N, O, S, Cl. In order to

better represent the effect of long-range interactions, a subset of QM7-X encompassing structures with more than 20 atoms was selected to train and validate the different models. The reduced dataset contains approximately 200 000 different structures, with 99% of all pairwise distances below 7 Å and an average of 3.4 ± 1.3 Å.

2.2 AQM

The Aquamarine dataset contains over 40 global and local physicochemical properties of $\sim 60\,000$ low- and high-energy conformers of 1 653 molecules with up to 92 atoms, both in gas phase and implicit water[12]. In our tests, we only considered the gas phase version of the dataset and we further filtered out all structures with less than 30 atoms. This selection led to $\sim 52\,000$ structures with mean pairwise distance of 6 ± 3 Å. Approximately 65% of all pairwise distances are below 7 Å, 83% are below 9 Å and 95% are below 12 Å.

2.3 DHA

We selected the portion of the MD22 dataset associated to the Docosahexaenoic Acid (*DHA*), a lipid of biological interest composed of 56 atoms. Atomic and molecular properties are reported for $\sim 70\,000$ structures. The mean pairwise distance between the atoms of each molecule in the dataset is 6 ± 3 Å with 63% of them below 7 Å, 81% below 9 Å and 94% below 12 Å.

Supplementary Note 3: Model training

All the models where trained using the combined force and energy loss:

$$\mathcal{L} = \alpha \sum_{i=1}^{N} |E_i - E(\mathbf{X_i}; \theta)|^2 + \sum_{i=1}^{N} |\mathbf{F}_i + \nabla E(\mathbf{X_i}; \theta)|^2. \tag{15}$$

Here, N is the number of molecules, E_i and $\mathbf{F}_i$ are the potential energy and forces acting on the i-th molecule. $E(\mathbf{X_i}; \theta)$ and $\nabla E(\mathbf{X_i}; \theta)$ are energy and forces predicted by the model, that depend on the network parameters θ. Finally, α is a scalar value controlling the relative numerical weight between force and energy contribution. A term that acts specifically on the parameters that regulate the activation of multiple master nodes is introduced in the loss function as

$$\mathcal{L}_{\text{reg}} = \sum_{I} \delta |\lambda_I + a_I + b_I|, \tag{16}$$

where the scalar δ was set to 2.0 during all the trainings. All models were trained on the QM7-X and AQM datasets for 200 epochs, while the training on the DHA dataset was extended to 500 epochs. The AdamW[18] optimizer was used in all training, with initial learning rate of 0.0001 and a weight decay of 0.01. For the first 125 epochs, α was set to 0.01, and subsequently increased to 0.1. A linear scheduler was used with a gamma factor of 0.8 and learning rate step size of 19 for optimizing the model parameters, 6 for the regularization parameter λ_I, and 8 for the parameters a_I and b_I. In order to scale different parameter groups with different step sizes, we employed a custom implementation of the standard *LinearLR* class in the PyTorch library[19]. Model hyperparameters are reported in Supplementary Table 1 and Supplementary Table 2.

Supplementary Table 1: Training hyperparameters. Neural network hyperparameters used for all baseline models and their RANGE counterparts.

	Training setup
Hidden channels (H)	512
Number of Filters (L)	512
Interaction Blocks (T)	3
Activation	tanh
Cutoff function	CosineCutoff
Distance Expansion Basis	Gaussian RBF[20]
Master node RBF dimension	7
Output Network	MLP, 2 layers, [128,64] features
Output Prediction	energy, forces
Attention heads	16

Supplementary Table 2: Radial basis expansion. Dimension of the radial basis expansion used in all baseline models and their RANGE counterparts for different cutoff radii.

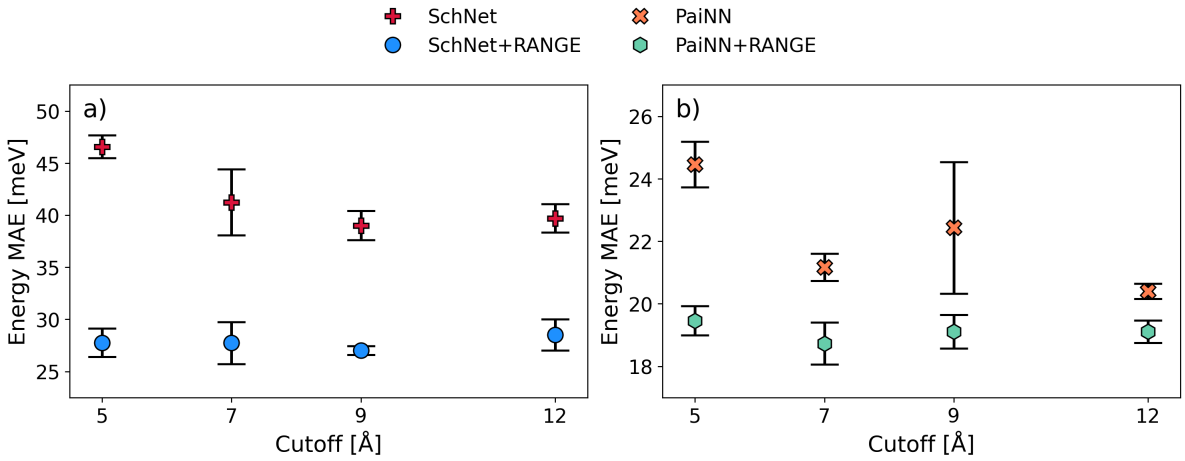
Radius (Å)	Number of RBF SchNet	Number of RBF PaiNN
4.0	27	-
5.0	33	20
7.0	47	28
9.0	60	36
12.0	80	48

3.1 Timing

All time measurements were performed considering the mean training time averaged over 200 epoch. To ensure accurate and reliable evaluation of this metric, all time measurements were performed in a controlled environment: a compute node with 4 *NVIDIA RTX A6000-ADA* GPUs isolated from the main compute cluster and a refrigerating system were reserved for this work in order to avoid slow downs due to over-warming. Temperature and power were constantly measured for every GPU during training as indicators of the experiments' stability. The goodness of the experimental setting is confirmed further by the low relative errors reported in Supplementary Tables 3 and 4.

Supplementary Note 4: Simulation details

All-atom simulations of DHA were conducted using a SchNet+RANGE model with a baseline cutoff of 5.0 Å, 3 master nodes, and 16 attention heads for stability analysis. Each simulation was run for 16 ns using a Langevin integrator at 300 K, with a timestep of 2 fs. To gather robust statistics on the conformational space exploration by each model, 20 parallel simulations were performed. Supplementary Fig. 4 presents the time series of the radius of gyration during the

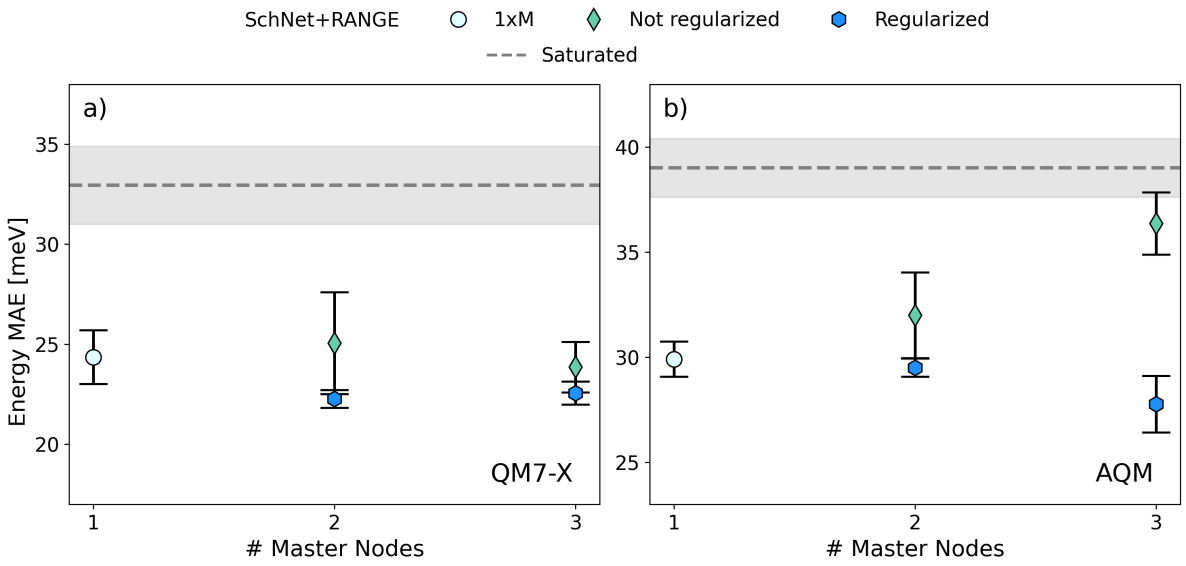


Supplementary Figure 1: Accuracy dependence on message-passing cutoff. The MAE on the predicted energy of the AQM dataset is reported for a) SchNet and b) PaiNN, and the same models with the RANGE extension, as a function of the message-passing cutoff. All the reported values are averaged on 4 models independently trained with different dataset seeds.

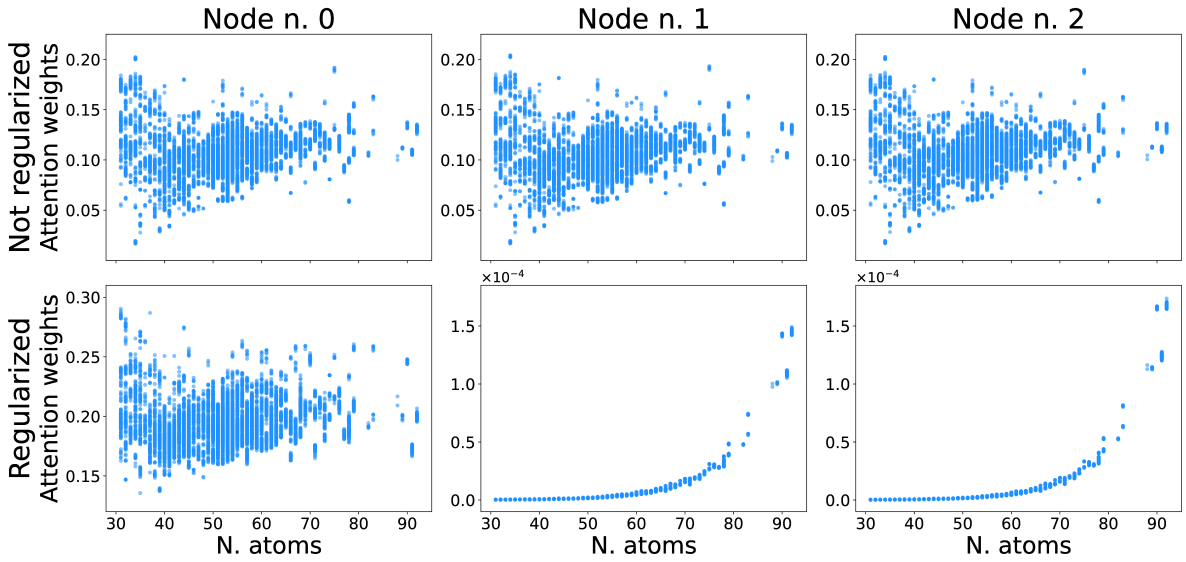
simulations. Notably, the model successfully explored a diverse range of DHA conformations, spanning compact and extended states.

Supplementary Note 5: Interpretation and singular value decomposition analysis

For each configuration in the validation set of DHA, V_{DHA}, two N dimensional vector, containing aggregation and broadcast weights of the master node with $\lambda_1 = 1$ during the last interaction block, are stored as matrix rows to analyze the attention patterns of the RANGE model. The two matrices of size $|V_{\mathrm{DHA}}| \times N$ are decomposed in singular values for every attention head separately. Supplementary Fig. 5 shows the results for aggregation and broadcast. Singular values within each matrix are normalized with respect to their maximum, highlighted in red. A single, dominant pattern associated to an N-dimensional principal component emerges, and its coefficients can be mapped onto the molecular graph with a color index (Supplementary Fig. 6).



Supplementary Figure 2: MAE of the regularized and non-regularized RANGE model. Energy MAE of the regularized and non-regularized RANGE models with different number of master nodes are reported for the a) QM7-X and b) AQM datasets. The gray line represents the lowest MAE achieved by the baseline model upon increasing the message-passing cutoff. All the reported values are averaged on 4 models independently trained with different dataset seeds.



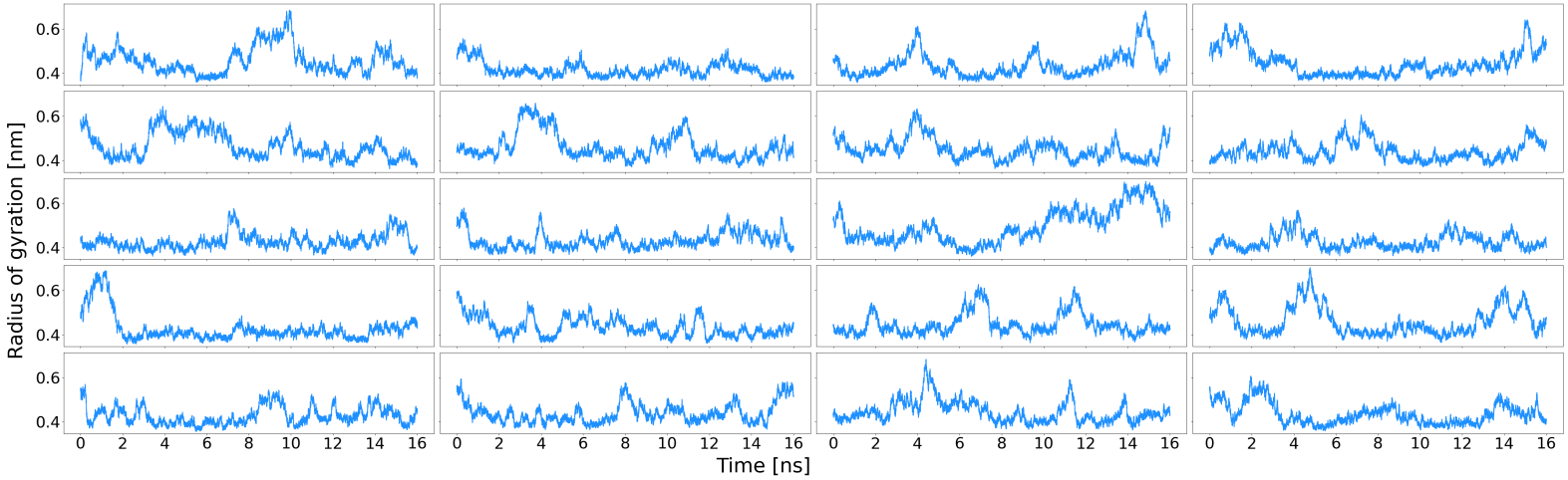
Supplementary Figure 3: Magnitude of the regularization. Comparison of mean molecular broadcast attention weights between the non-regularized and regularized SchNet+RANGE model with 3 master nodes on the AQM dataset. The regularized model effectively reduces the relevance of nodes 1 and 2, mitigating the redundancy observed in the non-regularized model, for the smallest samples in the validation set.

Supplementary Table 3: Accuracy and training time on QM7-X, AQM, and DHA datasets. Accuracy and training time are reported for different SchNet models, and the RANGE model with varying number of master nodes M (1, 2, and 3). Non regularized RANGE models are indicated as RANGE-NR. All the reported values are averaged on 4 models independently trained with different dataset seeds. The best results are in bold lettering.

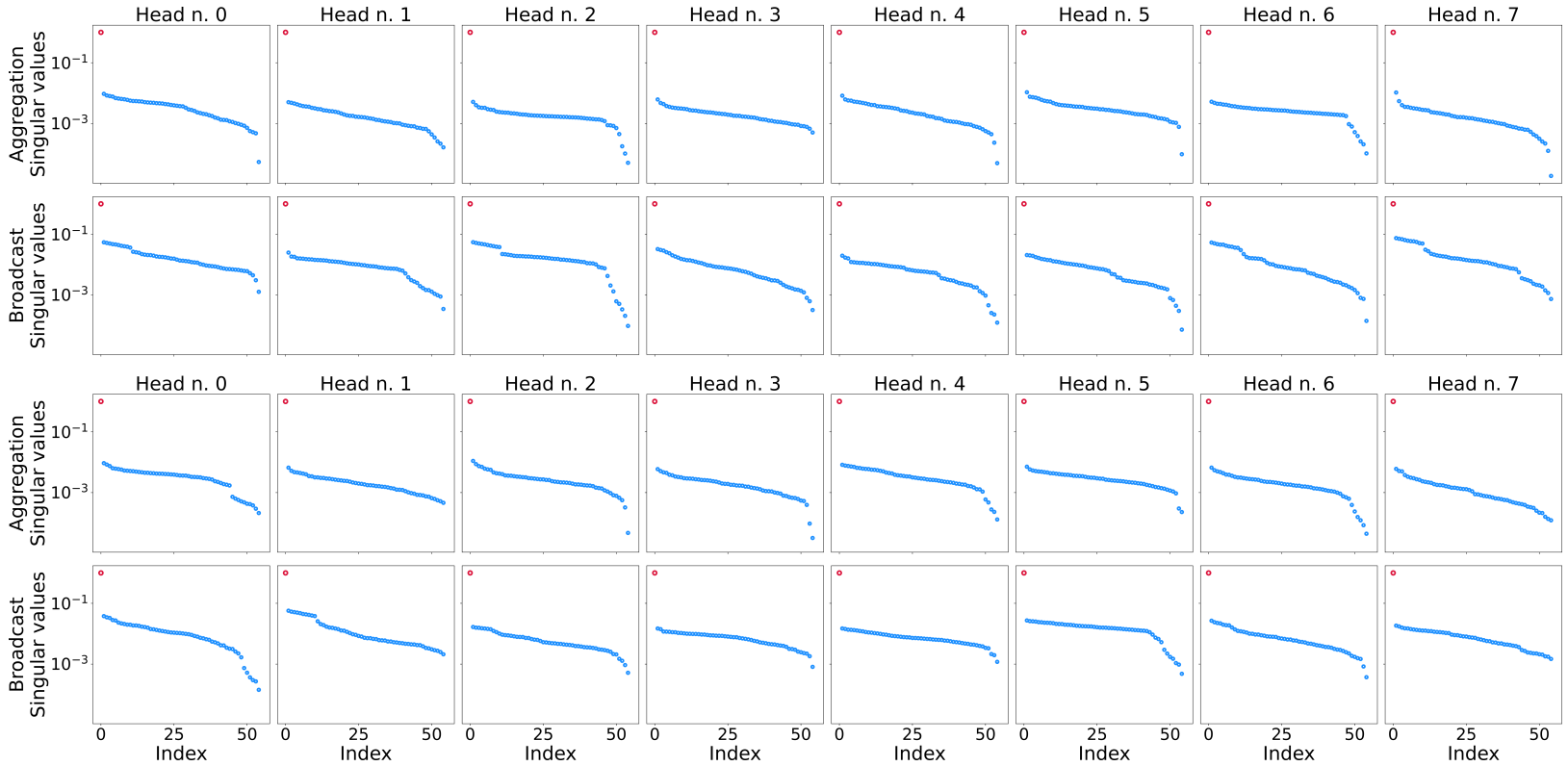
Model	MAE energy [meV]	MAE forces [meV/Å]	Training time [min/epoch]
	QM7-X		
Baseline 4 Å	39.2 ± 1.8	51.4 ± 0.1	0.809 ± 0.002
Baseline 5 Å	34.6 ± 1.3	47.3 ± 0.1	0.993 ± 0.003
Baseline 7 Å	33 ± 2	45.0 ± 0.2	1.123 ± 0.002
Baseline 9 Å	30.5 ± 0.8	43.9 ± 0.2	1.131 ± 0.002
RANGE 4 Å (1xM)	24.4 ± 1.4	33.3 ± 0.4	1.243 ± 0.005
RANGE 4 Å (2xM)	$\mathbf{22.3} \pm 0.4$	32.73 ± 0.12	1.316 ± 0.005
RANGE 4 Å (3xM)	22.6 ± 0.6	$\mathbf{32.57} \pm 0.14$	1.391 ± 0.004
RANGE-NR 4 Å (2xM)	25 ± 3	33.4 ± 0.3	1.32 ± 0.01
RANGE-NR 4 Å (3xM)	23.9 ± 1.3	33.6 ± 0.4	1.40 ± 0.02
	AQM		
Baseline 5 Å	46.6 ± 1.1	20.3 ± 0.2	0.831 ± 0.002
Baseline 7 Å	41 ± 3	18.6 ± 0.2	1.257 ± 0.002
Baseline 9 Å	39.0 ± 1.4	18.6 ± 0.3	1.550 ± 0.002
Baseline 12 Å	39.7 ± 1.4	18.7 ± 0.3	1.791 ± 0.003
RANGE 5 Å (1xM)	29.9 ± 0.8	13.6 ± 0.3	1.212 ± 0.005
RANGE 5 Å (2xM)	29.5 ± 0.4	13.4 ± 0.4	1.250 ± 0.006
RANGE 5 Å (3xM)	$\mathbf{27.8} \pm 1.4$	$\mathbf{12.9} \pm 0.4$	1.284 ± 0.006
RANGE-NR 5 Å (2xM)	32 ± 2	14.4 ± 0.7	1.241 ± 0.002
RANGE-NR 5 Å (3xM)	36.4 ± 1.5	15.1 ± 0.3	1.267 ± 0.005
	DHA		
Baseline 5 Å	34.9 ± 0.3	40.9 ± 0.3	-
Baseline 7 Å	28.2 ± 0.3	37.2 ± 0.2	-
Baseline 9 Å	25.1 ± 0.1	36.2 ± 0.1	-
Baseline 12 Å	23.1 ± 0.4	36.0 ± 0.2	-
RANGE 5 Å (1xM)	16.6 ± 0.3	26.6 ± 0.1	-
RANGE 5 Å (2xM)	16.00 ± 0.08	26.0 ± 0.1	-
RANGE 5 Å (3xM)	$\mathbf{15.7} \pm 0.4$	$\mathbf{25.7} \pm 0.2$	-

Supplementary Table 4: Accuracy and training time of SchNet+RANGE and PaiNN+RANGE on the AQM dataset. Accuracy and training time are reported for different SchNet and PaiNN models, and their RANGE-corrected variants. All the reported values are averaged on 4 models independently trained with different dataset seeds. The best results are in bold lettering.

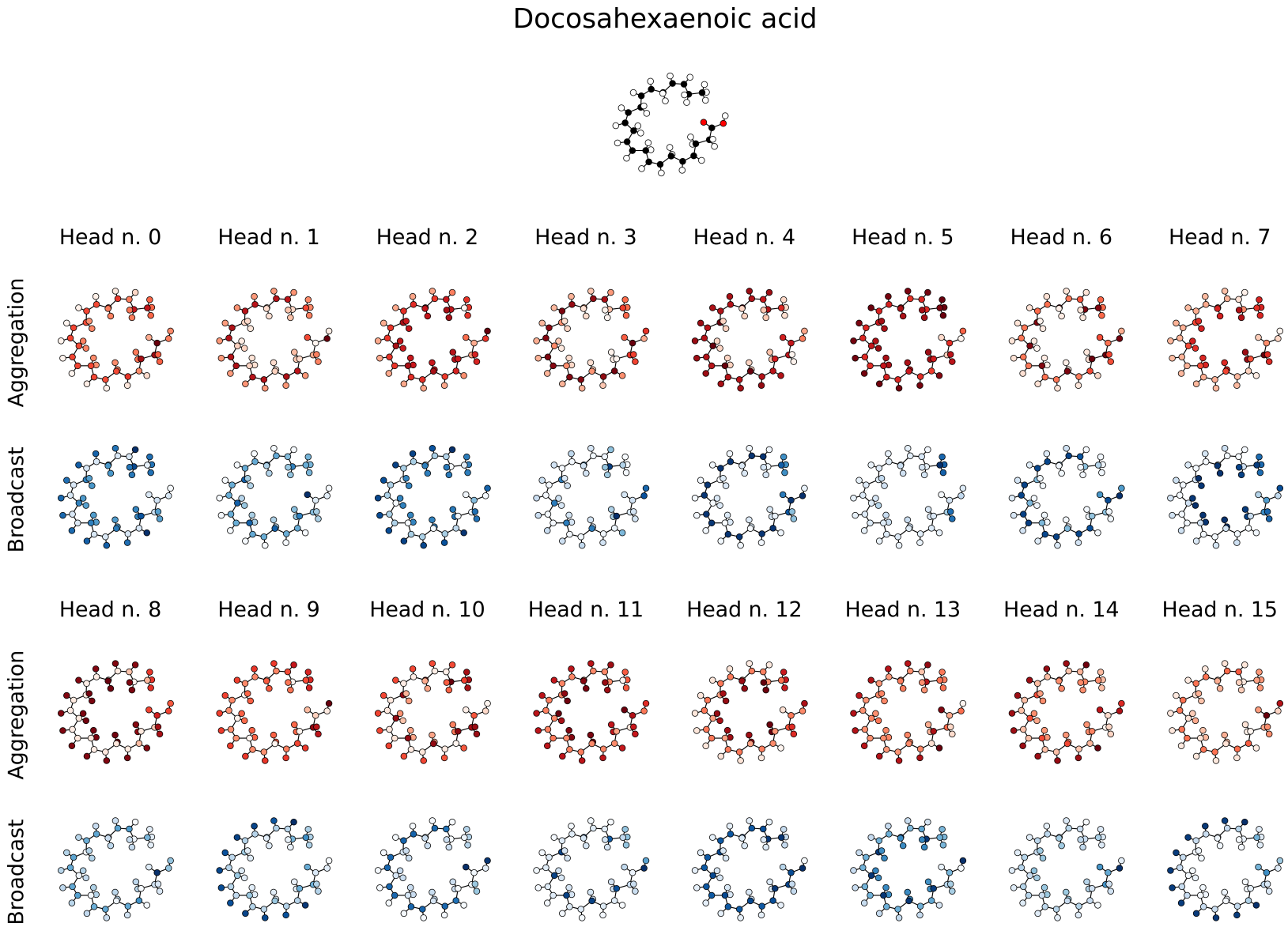
	Model	MAE energy [meV]	MAE forces [meV/Å]	Training time [min/epoch]
SchNet	Baseline 5 Å	46.6 ± 1.1	20.3 ± 0.2	0.831 ± 0.002
	Baseline 7 Å	41 ± 3	18.6 ± 0.2	1.257 ± 0.002
	Baseline 9 Å	39.0 ± 1.4	18.6 ± 0.3	1.550 ± 0.002
	Baseline 12 Å	39.7 ± 1.4	18.7 ± 0.3	1.791 ± 0.003
	RANGE 5 Å	27.8 ± 1.4	12.9 ± 0.4	1.284 ± 0.006
	RANGE 7 Å	28 ± 2	$\mathbf{12.7} \pm 0.3$	1.692 ± 0.017
	RANGE 9 Å	$\mathbf{27.0} \pm 0.4$	12.9 ± 0.3	1.971 ± 0.011
	RANGE 12 Å	28.5 ± 1.5	13.5 ± 0.3	1.971 ± 0.011
PaiNN	Baseline 5 Å	24.5 ± 0.7	8.92 ± 0.14	3.103 ± 0.005
	Baseline 7 Å	21.2 ± 0.4	8.59 ± 0.14	4.705 ± 0.006
	Baseline 9 Å	22 ± 2	8.7 ± 0.3	5.6 ± 0.4
	Baseline 12 Å	20.4 ± 0.2	8.62 ± 0.12	6.692 ± 0.003
	RANGE 5 Å	19.5 ± 0.5	7.68 ± 0.17	3.71 ± 0.01
	RANGE 7 Å	$\mathbf{18.7} \pm 0.7$	7.30 ± 0.06	5.28 ± 0.01
	RANGE 9 Å	19.1 ± 0.5	$\mathbf{7.26} \pm 0.18$	6.422 ± 0.008
	RANGE 12 Å	19.1 ± 0.4	7.47 ± 0.17	7.24 ± 0.02



Supplementary Figure 4: Radius of gyration of DHA as a function of simulation time. The radius of gyration is calculated along 16 ns of MD trajectory simulated with the RANGE architecture applied on SchNet with a 5 Å cutoff, across 20 independent trajectories.



Supplementary Figure 5: Singular value decomposition (SVD) of aggregation and broadcast weights. The SVD analysis is performed on the master node with $\lambda_1 = 1$. Its principal component, corresponding to the largest value, is marked in red.



Supplementary Figure 6: Principal component of attention weights. The colors in the top figure represent the atomic species (white: H, black: C, red: O). In the bottom figure, the principal component of the SVD on the attention weight distribution during aggregation and broadcast for all 16 attention heads is reported. Darker colors correspond to higher values.

References

1. Veličković, P., Cucurull, G., Casanova, A., *et al.* Graph attention networks. *arXiv preprint arXiv:1710.10903* (2017).
2. Brody, S., Alon, U. & Yahav, E. How attentive are graph attention networks? *arXiv preprint arXiv:2105.14491* (2021).
3. Giambagli, L., Buffoni, L., Chicchi, L., *et al.* How a student becomes a teacher: learning and forgetting through Spectral methods. *Adv. Neural Inf. Process.* **36,** 60291–60306 (2023).
4. Liu, Z., Li, J., Shen, Z., *et al. Learning efficient convolutional networks through network slimming* in *Proceedings of the IEEE international conference on computer vision* (2017), 2736–2744.
5. Satorras, V. G., Hoogeboom, E. & Welling, M. *E(n) equivariant graph neural networks* in *International conference on machine learning* (2021), 9323–9332.
6. Schütt, K. T., Unke, O. & Gastegger, M. *Equivariant message passing for the prediction of tensorial properties and molecular spectra* in *International Conference on Machine Learning* (2021), 9377–9388.
7. Batatia, I., Kovacs, D. P., Simm, G., *et al.* MACE: Higher order equivariant message passing neural networks for fast and accurate force fields. *Adv. Neural Inf. Process.* **35,** 11423–11436 (2022).
8. Fu, X., Wu, Z., Wang, W., *et al.* Forces are not enough: Benchmark and critical evaluation for machine learning force fields with molecular simulations. *arXiv preprint arXiv:2210.07237* (2022).
9. Kosmala, A., Gasteiger, J., Gao, N., *et al. Ewald-based long-range message passing for molecular graphs* in *International Conference on Machine Learning* (2023), 17544–17563.
10. Wang, Y., Cheng, C., Li, S., *et al.* Neural P^3M: A Long-Range Interaction Modeling Enhancer for Geometric GNNs. *arXiv preprint arXiv:2409.17622* (2024).
11. Hoja, J., Medrano Sandonas, L., Ernst, B. G., *et al.* QM7-X, a comprehensive dataset of quantum-mechanical properties spanning the chemical space of small organic molecules. *Sci. Data* **8,** 43 (2021).
12. Medrano Sandonas, L., Van Rompaey, D., Fallani, A., *et al.* Dataset for quantum-mechanical exploration of conformers and solvent effects in large drug-like molecules. *Sci. Data* **11,** 742 (2024).
13. Chmiela, S., Vassilev-Galindo, V., Unke, O. T., *et al.* Accurate global machine learning force fields for molecules with hundreds of atoms. *Sci. Adv.* **9,** eadf0873 (2023).
14. Tkatchenko, A., DiStasio Jr, R. A., Car, R., *et al.* Accurate and efficient method for many-body van der Waals interactions. *Phys. Rev. Lett.* **108,** 236402 (2012).
15. Ambrosetti, A., Reilly, A. M., DiStasio, R. A., *et al.* Long-range correlation energy calculated from coupled atomic response functions. *J. Chem. Phys.* **140** (2014).
16. Stöhr, M., Michelitsch, G. S., Tully, J. C., *et al.* Communication: Charge-population based dispersion interactions for molecules and materials. *J. Chem. Phys.* **144** (2016).

17. Mortazavi, M., Brandenburg, J. G., Maurer, R. J., *et al.* Structure and stability of molecular crystals with many-body dispersion-inclusive density functional tight binding. *J. Phys. Chem. Lett.* **9,** 399–405 (2018).

18. Loshchilov, I. & Hutter, F. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101* (2017).

19. Paszke, A., Gross, S., Massa, F., *et al.* Pytorch: An imperative style, high-performance deep learning library. *Adv. Neural Inf. Process.* **32** (2019).

20. Schütt, K. T., Kindermans, P.-J., Sauceda Felix, H. E., *et al.* Schnet: A continuous-filter convolutional neural network for modeling quantum interactions. *Adv. Neural Inf. Process.* **30** (2017).