Electronic Supplementary Material (ESI) for Physical Chemistry Chemical Physics.

SUPPORTING INFORMATION
Graph Convolutional Neural Networks for (QM)ML/MM Molecular Dynamics Simulations

Albert Hofstetter, Lennard Böselt, Sereina Riniker*

[*] *Laboratory of Physical Chemistry, ETH Zurich, Vladimir-Prelog-Weg 2, 8093 Zurich, Switzerland*
Email: sriniker@ethz.ch

Contents

1 Model Selection

Table S1: Influence of the model depth on the mean absolute error (MAE) on the validation set for the basic GCNN model for the test systems benzene in water and uracil in water. For each property, the model with the lowest MAE is marked in bold.

Benzene			
	$\Delta-$learning		
Depth	E kJ mol^{-1}	F_{QM} kJ mol^{-1} nm^{-1}	F_{MM} kJ mol^{-1} nm^{-1}
2	1.7	**20.7**	4.0
4	1.2	21.0	3.6
6	**1.0**	22.9	3.5
8	**1.0**	24.6	**3.4**
Uracil			
	$\Delta-$learning		
Depth	E kJ mol^{-1}	F_{QM} kJ mol^{-1} nm^{-1}	F_{MM} kJ mol^{-1} nm^{-1}
2	**1.6**	**49.5**	**1.2**
4	2.0	55.5	**1.2**
6	2.2	60.2	1.3
8	32.6	62.7	1.3

Table S2: Influence of the number of features (n_f) on the mean absolute error (MAE) on the validation set for the basic GCNN model for the test systems benzene in water and uracil in water. For each property, the model with the lowest MAE is marked in bold.

Benzene			
	$\Delta-$learning		
n_f	E kJ mol^{-1}	F_{QM} kJ mol^{-1} nm^{-1}	F_{MM} kJ mol^{-1} nm^{-1}
32	1.7	22.1	4.2
64	1.3	**21.5**	3.9
128	1.0	22.9	3.5
256	**0.9**	22.9	**3.1**
Uracil			
	$\Delta-$learning		
n_f	E kJ mol^{-1}	F_{QM} kJ mol^{-1} nm^{-1}	F_{MM} kJ mol^{-1} nm^{-1}
32	2.5	54.3	**1.2**
64	**1.9**	**53.5**	1.3
128	2.2	60.2	1.3
256	2.5	60.8	1.3

2 Force Loss Contribution

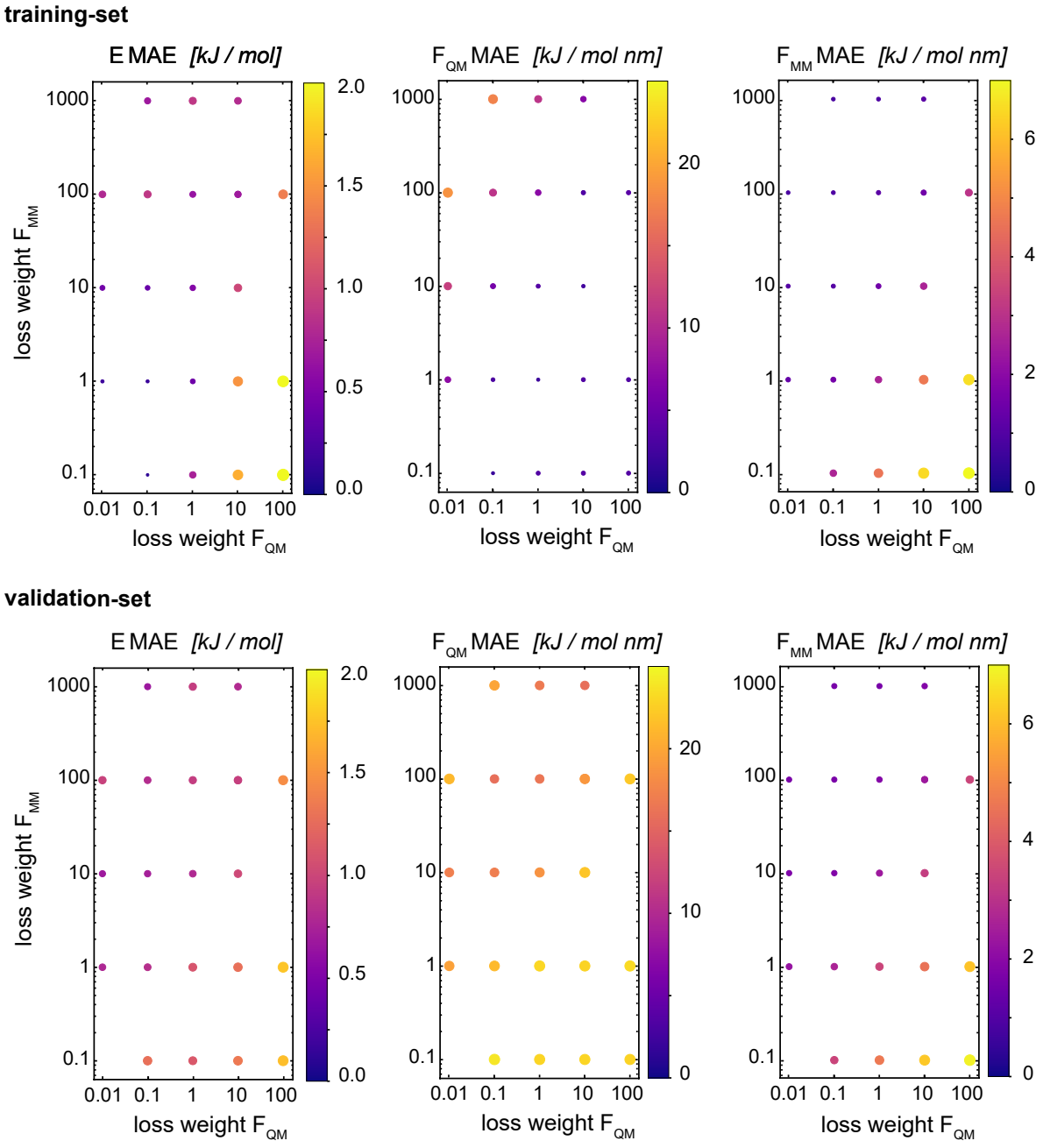

Figure S1: Influence of relative weights (wF_{QM} and wF_{MM}) for the different loss terms (QM forces, MM forces) on the mean absolute error (MA of the energy (**left**), forces on QM particles (**middle**), and forces on the MM particles from the QM zone (**right**). The Δ-learning GCNN was used with the test system benzene in water. The weight of the energy loss (wE) was kept constant at 1.0. (**Top**): MAE for the training set. (**Bottom**): MAE for the validation set. The color map and the relative size of the points indicate the MAE.

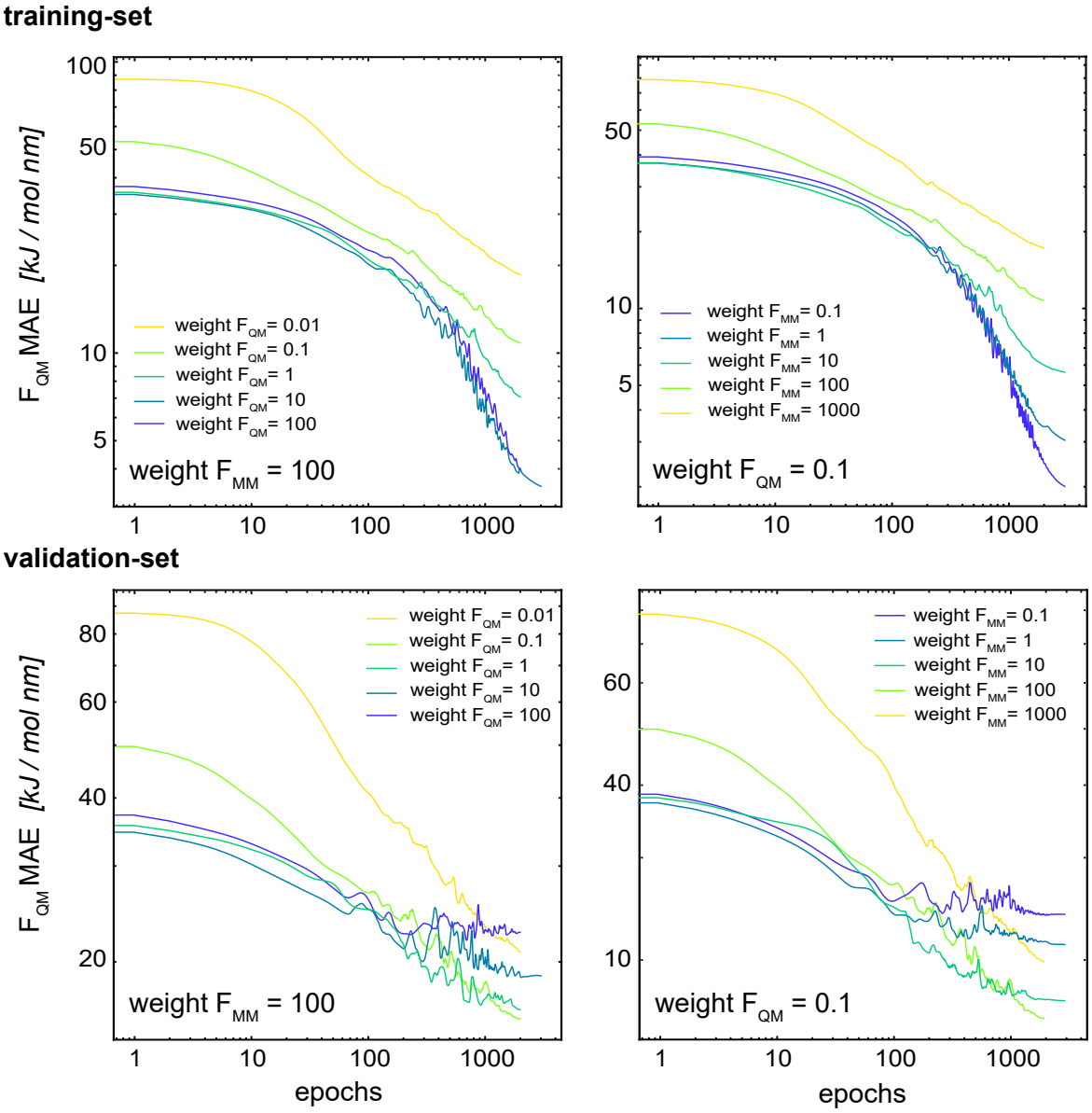

Figure S2: Learning curves for F_{QM} in the training set (**top**) and validation set (**bottom**) when varying the relative loss weightings wF_{QM} (**left**) and wF_{MM} (**right**). The Δ-learning GCNN was used with the test system benzene in water. The weight of the energy loss (wE) was kept constant at 1.0.

3 Neighborhood Reduction

Table S3: Influence of the neighborhood reduction schemes on the mean absolute error (MAE) on the validation set for the Δ-learning GCNN model with the test systems benzene in water and uracil in water. For each property, the model with the lowest MAE is marked in bold.

Benzene			
	Δ−learning		
Scheme	E	F_{QM}	F_{MM}
	kJ mol^{-1}	kJ mol^{-1} nm^{-1}	kJ mol^{-1} nm^{-1}
Complete	**0.7**	**17.0**	**1.8**
KNN-8	2.2	28.4	5.1
KNN-12	1.5	26.3	4.3
Voronoi	2.1	34.8	3.1
Uracil			
	Δ−learning		
Scheme	E	F_{QM}	F_{MM}
	kJ mol^{-1}	kJ mol^{-1} nm^{-1}	kJ mol^{-1} nm^{-1}
Complete	**1.6**	**52.8**	**0.8**
KNN-8	2.6	60.8	1.0
KNN-12	2.0	57.6	1.0
Voronoi	3.1	73.9	0.9

3.1 Data Set Ordering

For the farthest-point-sampling (fps), we used a simple fingerprint together with a hash function. For each frame from the training set, we calculated all atom-atom distances (r_{ij}) rounded to 0.001 nm and stored these together with the corresponding atom types as $[Z_i, Z_j, r_{ij}]$. Next, we hashed these objects to 64 bit and counted the occurrence of each bit within the hash of each frame. Finally, we compared the resulting bit counts of different frames using the Jaccard/Tanimoto index [1, 2],

$$T = \frac{\sum_k C_k}{\sum k(U_k - C_k)} \tag{1}$$

where C_k is the intersection of the counts at bit k between two frames and U_k is the union of the counts at bit k between two frames.

The fps-ordering was initialized at the first MD frame. At each step, the frame with the least summed similarity to all already chosen frames was added to the fps-ordered set until all training frames were contained.

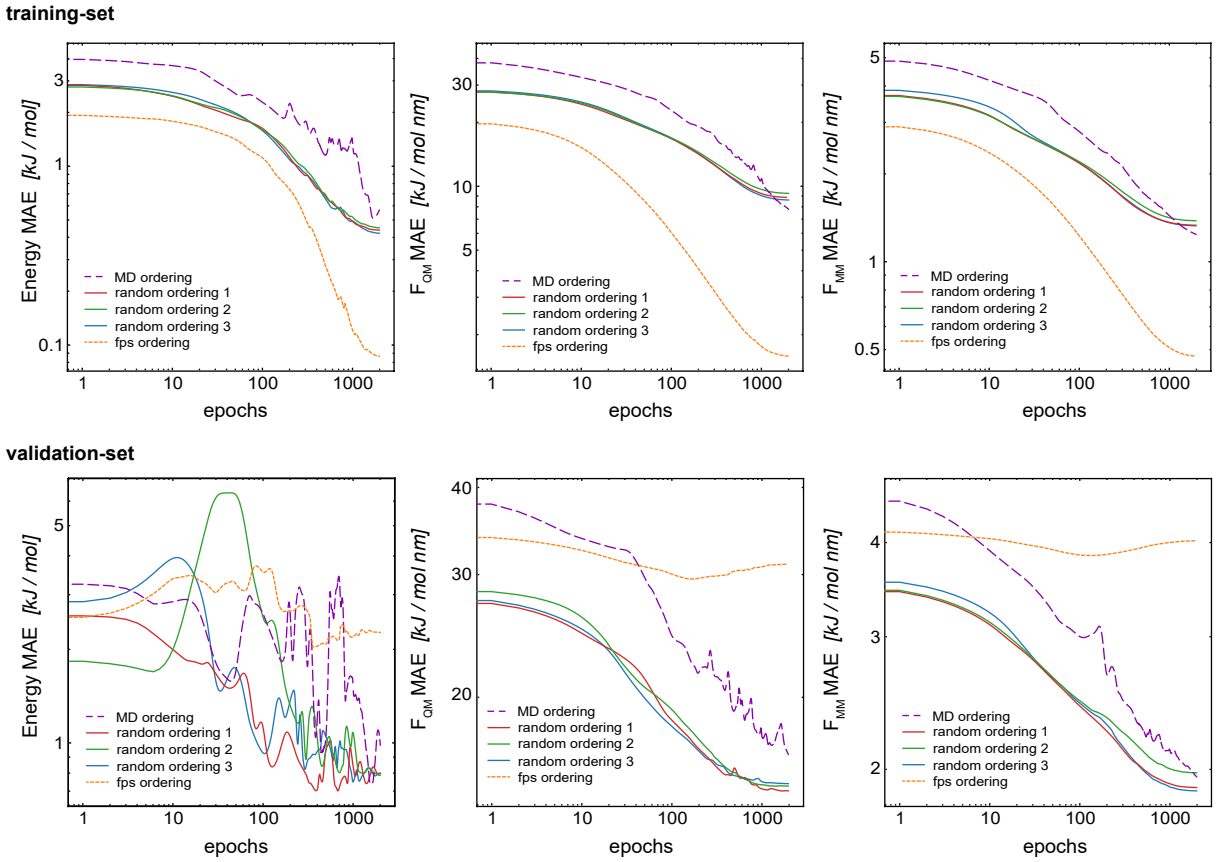

Figure S3: Learning curves for the training set (**top**) and validation set (**bottom**) when varying the order of the data points in the training set. The Δ-learning GCNN for benzene in water is shown.

References

[1] Paul Jaccard. The distribution of the flore in the alpine zone. 1. *New Phytologist*, 11:37–50, 1912.

[2] T T Tanimoto. *An Elementary Mathematical Theory of Classification and Prediction*. International Business Machines Corporation, 1958.