Prediction of Bioconcentration Factors (BCF) using Graph Neural Networks

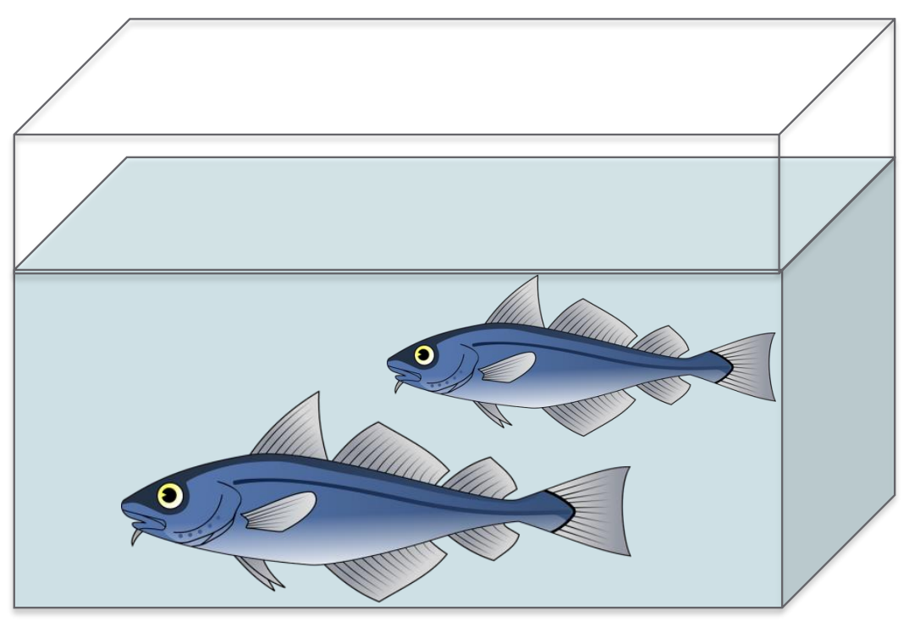
Edgar Ivan Sanchez Medina[1], Steffen Linke[1], Kai Sundmacher[1,2]

1 Chair for Process Systems Engineering, Otto von Guericke University, Universitätspl. 2, Magdeburg, 39106, Germany
2 Process Systems Engineering, Max Planck Institute for Dynamics of Complex Technical Systems, Sandtorstraße 1, Magdeburg, 39106, Germany

INTRODUCTION

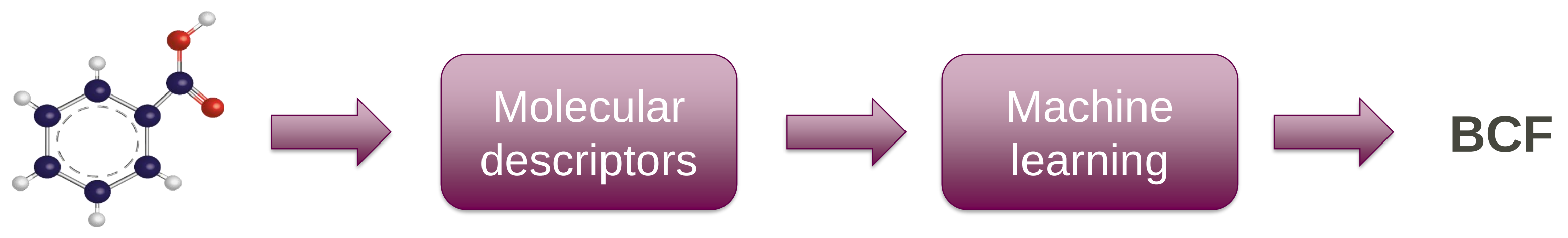
Bioconcentration factors (BCF)

When moving towards a sustainable chemical industry, chemical substances involved need to be assessed according to their environmental, health and safety properties. One of these properties refers to the capacity of the chemical substance to accumulate in body tissues. This specific property is usually expressed as a bioconcentration factor (BCF), which measures the ratio between the concentration of the substance in the organism's tissue to that in the environment [1].

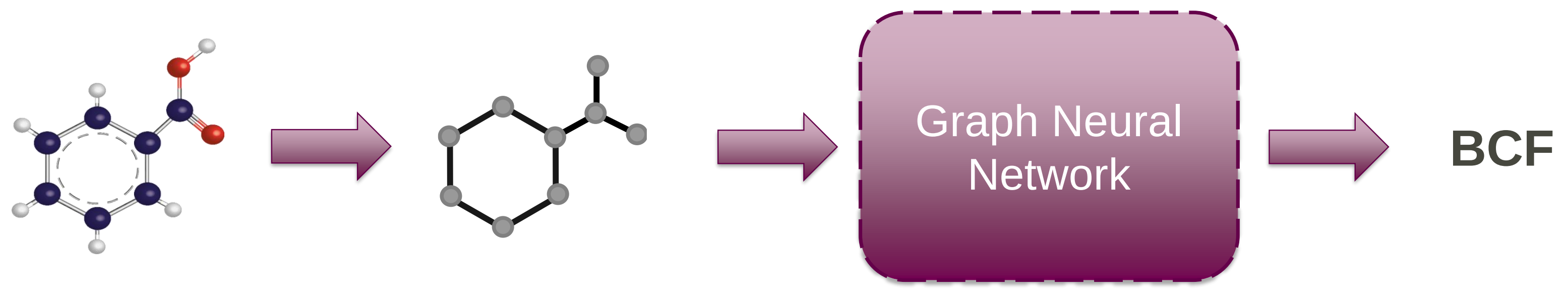


$$BCF\left[{}^{L}/_{kg}\right] = \frac{Concentration\ in\ tissue\ \left[{}^{\mu g}/_{kg}\right]}{Concentration\ in\ solution\ \left[{}^{\mu g}/_{L}\right]}$$

Modeling approach 1



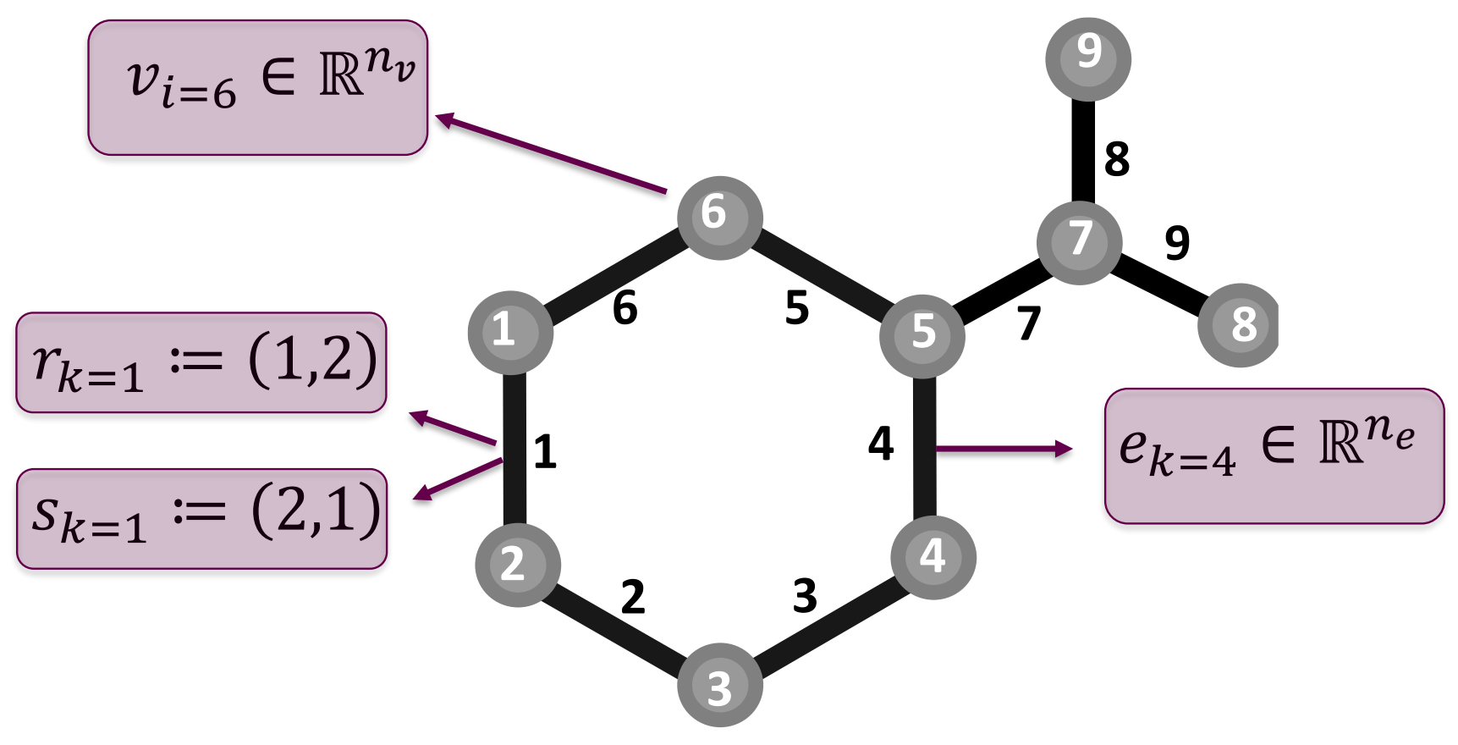
Modeling approach 2



BACKGROUND

Graphs

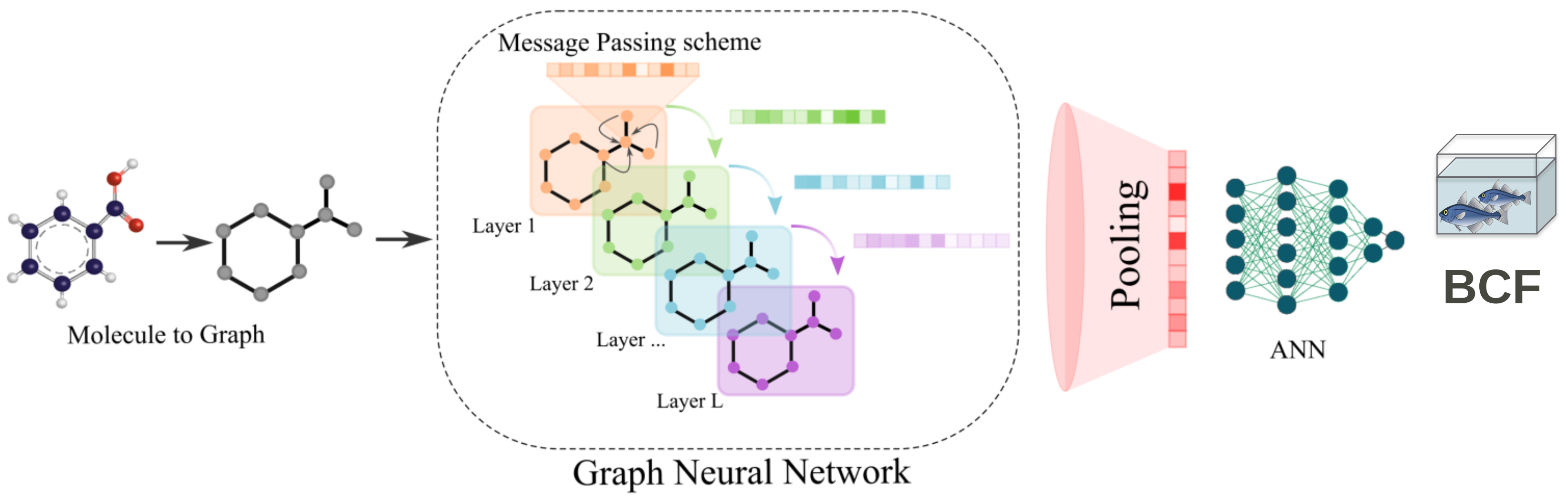
- Graph: $G = (V, E)$
- Node : $V \coloneqq \{v_i\}_{i=1:N^v}$
- Edge : $\mathrm{E} \coloneqq \{(e_k, r_k, s_k)\}_{k=1:N^e}$


$v_{i=6} \in \mathbb{R}^{n_v}$
$r_{k=1} \coloneqq (1,2)$
$s_{k=1} \coloneqq (2,1)$
$e_{k=4} \in \mathbb{R}^{n_e}$
1
2
3
4
5
6
7
8
9

Graph Neural Networks [2]

For each convolutional layer l:

- Message passing : $m_{v_i,v_j}^{(l)} = \phi_M\left(v_i^{(l-1)}, v_j^{(l-1)}, e_k\right)$, with k involving nodes i and j
- Message aggregation: $a_{v_i}^{(l)} = \phi_A\left(\left\{m_{v_i,v_j}^{(l)} : v_j \in \mathcal{N}(v_i)\right\}\right)$
- Features updating : $v_i^{(l)} = \phi_U\left(v_i^{(l-1)}, a_{v_i}^{(l)}\right)$



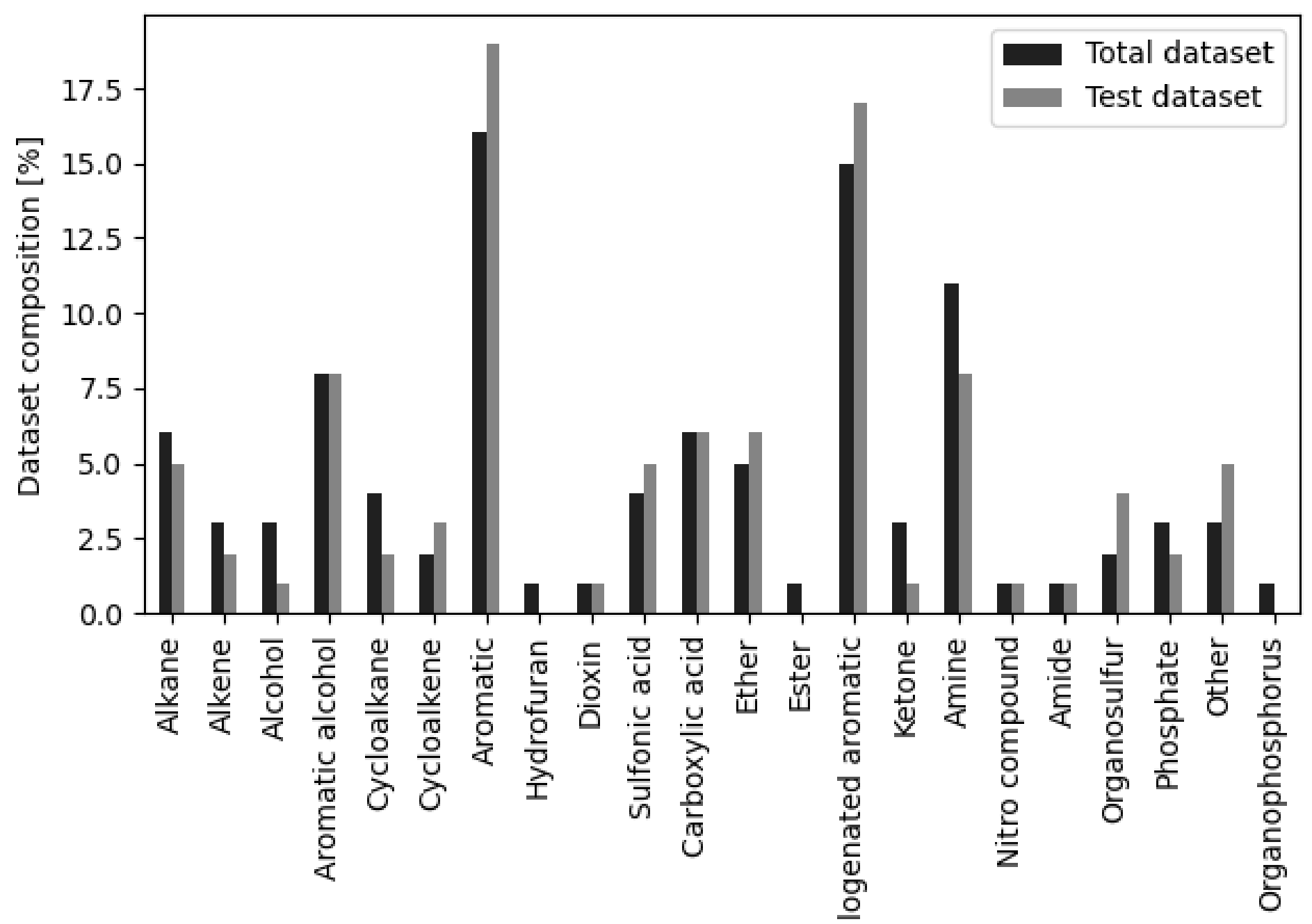
Pooling for vectorial molecular representation

The final-updated graph is pass through a permutation invariant pooling operation such as sum, max, mean or Set2Set [3].

METHODOLOGY

Database

- Dataset collected according to the REACH legislation [4].
- 473 molecules covering $log(BCF)\left[{}^{L}/_{kg}\right] \in [-1.0, 4.85]$ and molecular weights in range $[68, 943]\ {}^{g}/_{mol}$.



Molecule to graph

Table 1: Node features used to define molecular graphs.

Node feature	Description	Dimensions
Type *	Type of atom (C, O, Cl, N, F, Br, S, Other)	8
Ring	Whether the atom is part of a ring	1
Aromaticity	Whether the atom is part of an aromatic ring	1
Hybridization *	Hybridization of the atom (sp, sp^2, sp^3, sp^{3}d, sp^{3}d^2)	5
Bonds *	Number of bonds to the atom	6

* Implemented using one-hot-encoding (vector of binary values for each unique integer value).

Table 2: Edge features used to define molecular graphs.

Edge feature	Description	Dimensions
Type *	Type of bond (single, double, triple, aromatic)	4
Conjugated	Whether the bond is conjugated	1
Ring	Whether the bond is in a ring	1

* Implemented using one-hot-encoding (vector of binary values for each unique integer value).

Graph Neural Network architecture

- $m_{v_i,v_j}^{(l)} = v_j^{(l-1)} \cdot MLP_e(e_k)$ with k involving nodes i and j.
- $a_{v_i}^{(l)} = \sum_{v_j \in \mathcal{N}(v_i)} m_{v_i,v_j}^{(l)}$
- $v_i^{(l)} = GRU\left(\Theta^{(l)} v_i^{(l-1)} + a_{v_i}^{(l)}\right)$
- MLP_e with single hidden layer with 128 neurons and ReLU activation.
- 2 Graph-convolutional layers with size 21 with Set2Set pooling layer with 3 processing steps.
- 3-hidden layers in final MLP with 64, 32 and 16 neurons.

Training, validation and testing

- Trained with Adam (300 epochs, batch size of 30) using MSE as loss function and MAE as decision metric.
- 80% training (out of which 20% was used for validation) and 20% test.
- Ensemble learning using 10 models constructed with differend random seeds.
- Dropout of 50% used in convolutional layers and final MLP.

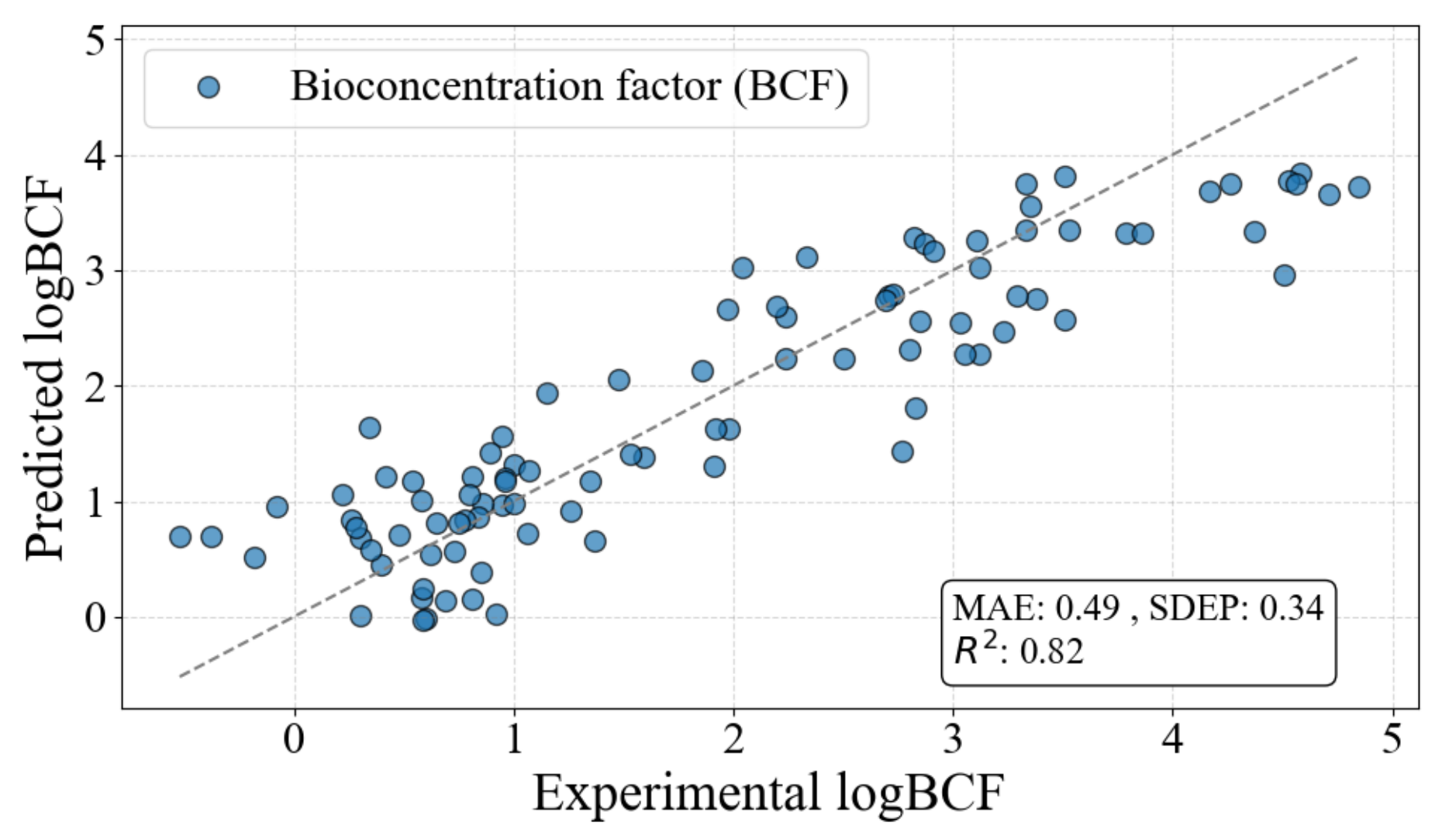
RESULTS AND DISCUSSION

- Comparable performance to the best QSAR model reported by Zhao et al. [4].
- 1022 descriptors [4] vs 8 structural parameters.
- Differences in physical insight.
- Future research: modeling of other EHS properties and determination of GNN applicability domains.

Table 3: Comparison of models for predicting BCF.

Model	Test set			Validation set			Training set		
	R^2	MAE	SDEP	R^2	MAE	SDEP	R^2	MAE	SDEP
GNN	**0.82**	0.49	**0.34**	**0.83**	**0.44**	**0.32**	0.82	**0.44**	**0.35**
Zhao	0.79	**0.45**	0.59	0.79	-	0.66	**0.83**	-	0.56

Bold numbers indicate preferred value.



ACKNOWLEDGEMENTS

This work is partly funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - TRR 63 "Integrated Chemical Processes in Liquid Multiphase Systems" (Gefördert durch die Deutsche Forschungsgemeinschaft (DFG) - TRR 63 "Integrierte chemische Prozesse in flüssigen Mehrphasensystemen" (Teilprojekt B9) - 56091768). E. I. S. M. and S. L. are also affiliated with the International Max Planck Research School for Advanced Methods in Process and Systems Engineering - IMPRS ProEng at the Max Planck Institute for Dynamics of Complex Technical Systems Magdeburg.

REFERENCES

[1] Wang, W.X., 2016. Bioaccumulation and biomonitoring. In Marine Ecotoxicology (pp. 99-119). Academic Press.
[2] Battaglia, P.W., Hamrick, J.B., Bapst, V., Sanchez-Gonzalez, A., Zambaldi, V., Malinowski, M., Tacchetti, A., Raposo, D., Santoro, A., Faulkner, R. and Gulcehre, C., 2018. Relational inductive biases, deep learning, and graph networks. arXiv preprint arXiv:1806.01261.
[3] Vinyals, O., Bengio, S. and Kudlur, M., 2015. Order matters: Sequence to sequence for sets. arXiv preprint arXiv:1511.06391.
[4] Zhao, C., Boriani, E., Chana, A., Roncaglioni, A. and Benfenati, E., 2008. A new hybrid system of QSAR models for predicting bioconcentration factors (BCF). Chemosphere, 73(11), pp.1701-1707.