Can deep learning algorithms enhance the prediction of solute descriptors for linear solvation energy relationship approaches?

Nadin Ulrich[1,2,*], Andrea Ebert[1]

[1]Department of Analytical Environmental Chemistry, Helmholtz Centre for Environmental Research - UFZ, Permoserstrasse 15, D-04318 Leipzig, Germany

[2]Department of Ecological Chemistry, Helmholtz Centre for Environmental Research - UFZ, Permoserstrasse 15, D-04318 Leipzig, Germany

*Corresponding Author: Phone + 49 341 235 1818 E-mail: nadin.ulrich@ufz.de

Highlights

- Deep learning models predict solute descriptors of LSERs
- Singletask models are better compared to multitask models due to the small dataset
- Data augmentation strategies based on tautomers improve the training of DNNs

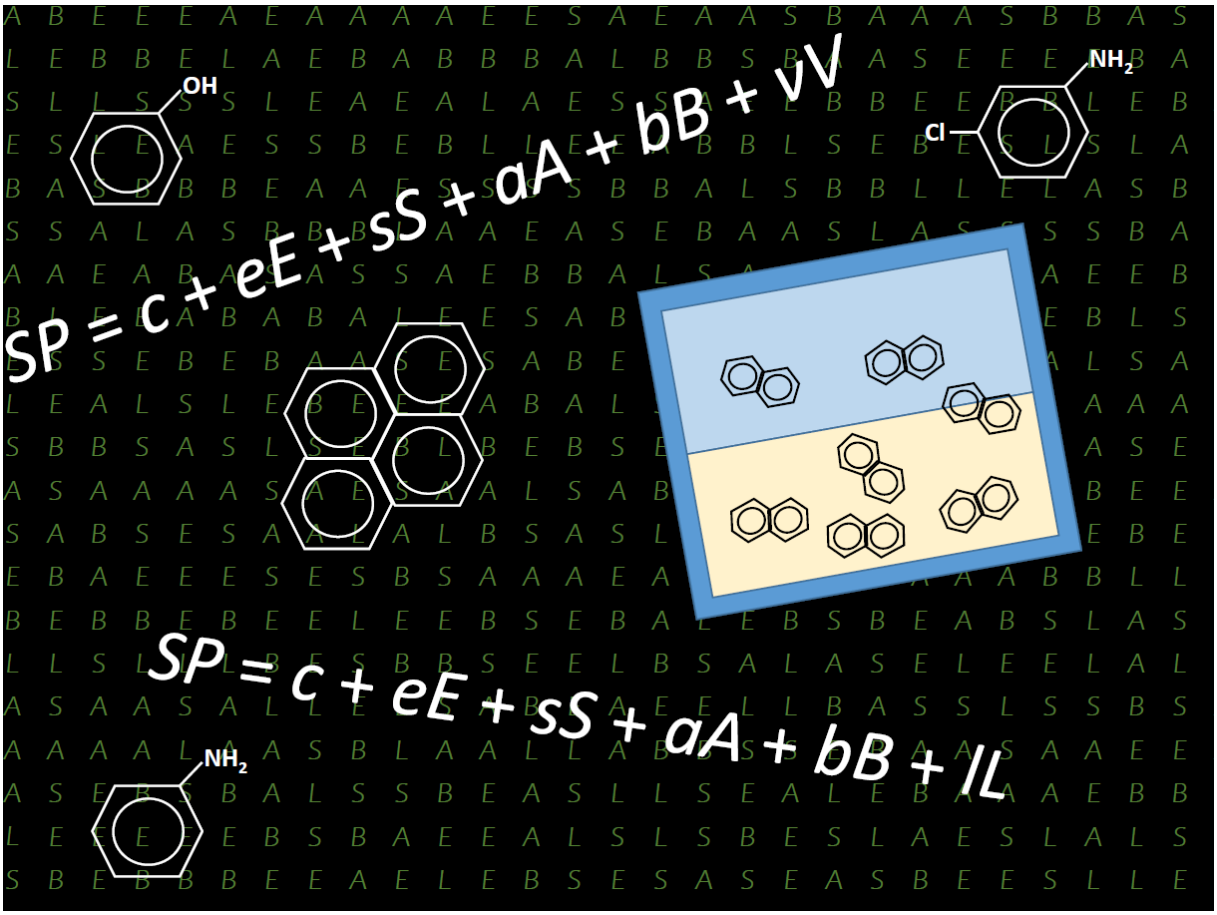
Abstract

Experimental solute descriptors for about 8,000 chemicals are currently available to apply physicochemical property predictions based on linear solvation energy relationship (LSER) models. The solute descriptors can be predicted by fragmental-based quantitative structure-property relationship (QSPR) models. However, the predictions are problematic for larger chemical structures, including multiple functional groups. We developed deep neural networks (DNNs) as alternative prediction models based on graph representations of the chemicals. The root mean square errors *rmse*s range between 0.11 and 0.46 for the different solute descriptors. The predictions of the solute descriptors were compared to predictions from the QSPR of LSERD (an online database) and ACD/Absolv (a commercial software). We further investigated the predictive power of all tools based on three different datasets of experimentally determined partition coefficients, namely the octanol-water partition coefficient (K_{ow}), the octanol-air partition coefficient (K_{oa}), and the water-air partition coefficient (K_{wa}). Additionally, we used two different sets of retention data for GC and LC to evaluate the results of all prediction tools. All prediction tools perform comparably well with *rmse*s of ~ 1.0 log unit for the K_{ow} dataset (12010 chemicals) and ~ 1.3 log units for the K_{wa} dataset (696 chemicals), for example. Nevertheless, larger chemical structures are predicted poorly by each approach. We recommend to use the novel DNN model as a complementary prediction tool.

Keywords

physicochemical property prediction partition coefficients data augmentation LSERD quantitative structure-property relationship QSPR Absolv

Graphical Abstract



1. Introduction

Researchers in environmental science and experts in chemical risk assessment need reliable predictions of physicochemical properties. Often they predict partition coefficients like the octanol-water partition coefficient (K_{ow}) and octanol-air partition coefficient (K_{oa}) and apply it to characterize the bioaccumulation potential of chemicals[1, 2]. Other frequently used partition coefficients are the water-air partition coefficient (K_{wa})[2, 3] and the organic carbon-water partition coefficient (K_{oc})[4], both describing the behavior of chemicals in the environment. Several partition coefficients for biocompartments explain enrichment of chemicals in specific tissues or organs in an organism[5, 6]. In each case, reliable results are only achieved by experimental values or precise predictions of the respective partition coefficients. An often applied mechanistic approach for the predictions of these partition coefficients are linear solvation energy relationship (LSER) models (Eq. (1), (2), and (3))[7, 8].

$$SP = c + eE + sS + aA + bB + vV \quad (1)$$

$$SP = c + eE + sS + aA + bB + lL \quad (2)$$

$$SP = c + sS + aA + bB + vV + lL \quad (3)$$

In eq. (1), (2), and (3) the solute property *SP* is defined by five different terms and a system constant *c*. Each term consists of an upper case letter defining the molecular interactions of the solute, and a lower case letter defining the molecular interactions of the surrounding phase system. The *eE* term includes the excess molar refraction *E* and van der Waals interactions of the two-phase system, *sS* includes polarizability and dipolarity in both cases. The *aA* term and *bB* term describe hydrogen bond interactions, namely the hydrogen bond acidity *A* and the hydrogen bond basicity *B*, and vice versa for the phase parameters. In the *vV* term the McGowan characteristic volume *V* and the cavity formation *v* are depicted. And the *lL* term includes the logarithmic hexadecane-air partition coefficient *L* and dispersion and cavity formation interactions *l*.

Eq. (1) describes condensed phase systems, whereas eq. (2) defines systems including air as one phase and a condensed phase as a second one. Eq. (3) is applicable for both systems[7]. However, the LSER approach is only valid for neutral chemicals. Some attempts have been made to adapt the equations to ionic chemicals by adding a j^+J^+ and j^-J^- term for cations and anions, respectively. The application domain is, nevertheless, rather small including only a few classes of chemicals[9].

LSER models are often applied to predict environmentally relevant partition coefficients for chemicals such as K_{wa}[10, 11] or K_{ow}[10, 12]. Further, they describe the sorption of chemicals to carbonaceous sorbents such as humic acids[13, 14]. The sorption to passive sampling materials[15, 16], which are commonly used for environmental analysis of aqueous samples e.g., is also predicted by the use of the LSER models. LSERs also depict the biopartitioning of chemicals[6], which includes the description of the sorption to proteins[17, 18] and lipids[19, 20] in an organism. There are many other partitioning processes of chemicals which are described by the equations, for example retention behavior in chromatography[21, 22] or adsorption processes to surfaces[23, 24].

Currently, experimentally determined solute descriptors are available for about 8,000 chemicals. But these days, the Chemical Abstracts Service (CAS) includes more than 182 million registered chemicals[25]. This means that the solute descriptors cover only a small range of chemicals in the chemical universe. It is not feasible to determine these descriptors for each chemical individually when applying the respective property predictions. Thus, methods

for precise predictions of solute descriptors are needed. A QSPR (fragmental approach) for the prediction of solute descriptors is available at the free online platform LSERD[24]. And a similar prediction tool is included in the commercial software ACD Percepta (Absolv)[26]. Both prediction tools deliver valuable results for simple chemical structures including one functional group. But predictions for complex chemical structures with multiple functionalities are more erroneous.

Here, we develop a deep neural network (DNN) to predict solute descriptors. Compared to the QSPR predictions, DNN models may offer the opportunity to overcome the problems in the prediction of solute descriptors for chemicals with multiple functional groups. DNN models have already been successfully applied to predict physicochemical properties[27-29]. They further act as fast alternatives to classical quantum-chemistry based methods[30] with comparable predictive performance, for example on the QM9 dataset [27, 31, 32]. Thus, our aim is to develop DNN models for the precise prediction of solute descriptors, and to verify if their performance is better in comparison to existing QSPR approaches. Further, our aim is to check whether these models can overcome current problems with predicted solute descriptors. We test their applicability in the prediction of partition coefficients especially for chemicals with large and complex structures.

2. Methods

2.1 Curation of the initial dataset

The DNN models were developed based on the Abraham Absolv dataset[24]. The dataset consists of 7,881 chemicals. We only focused on chemicals with more than one descriptor being available and selected the *S* descriptor as marker (see SI 1). We skipped all chemicals without *S* descriptor, which reduced the dataset to 7241 chemicals. We then excluded metals and organometallics as well as gases like argon, nitrogen, and methane and started with the identification of potential errors in the dataset. The identifiers of the chemicals in the database were already curated. Thus, our focus was on the experimentally determined solute descriptors. We developed pre-DNNs to identify potential erroneous solute descriptors by the appearance of outliers. For each outlier, we checked the original solver file (if available). The solver file includes all partition coefficients and solubility data which were used initially by M. Abraham to determine the solute descriptors of the respective chemical. The solver files were provided

by M. Abraham (personal communication). We identified several problems in the solver files and excluded these chemicals from the dataset. Errors that lead to exclusion are: A) only two partition coefficients are used to determine the solute descriptors, B) the chemical was ionized, C) calculated partition coefficients were used to determine the solute descriptors. We did not exclude outliers where the corresponding solver files were not available. We used JChem for Excel v. 20.6.0.618[33] for the generation of tautomeric forms based on the SMILES representation and for the calculation of p*K*a and p*K*b values. Details are given in the Table Dataset_and_predictions.xlsx in the GIT repository.

2.2 Technical details and development of DNN models

The models were trained using a Tuxedo book (Intel core i7, 64 GB RAM) with an NVIDIA RTX2080 Max Q (8 GB). Feed-forward graph convolutional networks were developed using the DeepChem library v. 2.2[34, 35] and Tensorflow v. 1.14.0 in Python v. 3.5.6[36]. The full code for the DNN development is given at the GIT repository https://github.com/nadinulrich/solute_descriptor_prediction. For model development, we used the DeepChem library, which is available at https://github.com/deepchem/deepchem. We adapted and modified the code given in the library for the prediction of the solute descriptors. The graph generated from the SMILES is an initial feature vector with a corresponding neighbor list for each atom in the molecule. The information included are atom-type, hybridization type, and valence structure[27]. The singletask DNNs consist of an input layer followed by two hidden layers, including 16 and 32 neurons (each hidden layer included graph convolution, batch normalization and graph pool layer), and a dense layer with ReLU activation function. After the dense layer, a batch normalization was applied with a dropout of 0.1. In the output layer a tanh activation function was implemented. The training was performed for each solute descriptor over 80 epochs for the singletask DNN without applying data augmentation (DNN_{mono}) and for the singletask DNN_{taut} where data augmentation was used. A batch size of 50 was used in both cases, and the learning rate was set to 0.0001. The parameters stated here resulted from an optimization procedure where different DNNs (two-layer and three-layer structure with different number of neurons) and different learning rates were tested. The performance of the different DNN model structures is strongly dependent on the size of the training set itself. Our dataset is relatively small. Thus, DNN models with more layers and neurons tended to overfitting and performed worse compared to the two-layer net with a smaller number of neurons, see SI 2. The same structure of the DNN and learning rate were used for

the multitask models. Only the number of epochs for training differed, with 150 epochs for the multitask DNN_{mono} and 100 epochs for multitask DNN_{taut}. The *rmse*s for the training set and validation set over epochs for the respectively best models are shown in SI 2. We randomly split 10% of our curated dataset as test set and further divided the remaining set into 80% training set and 20% validation set. The split was done for the *S* descriptor (which was the largest dataset). We kept the assignment of the chemicals for all descriptors, resulting in different numbers of the chemicals in each dataset (see Figure 1 and SI 1).

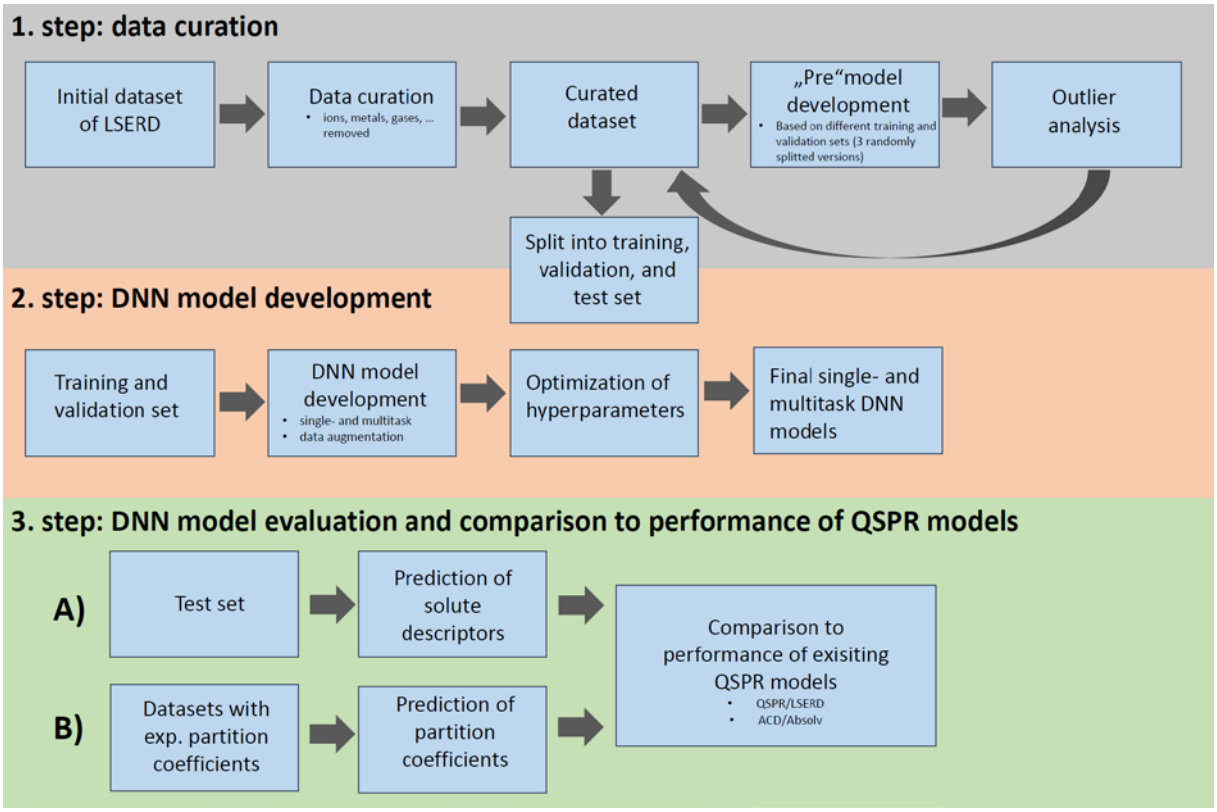


Figure 1 Scheme of DNN model development. After data curation (first step), the dataset was split into training, validation, and test set. Training and validation set were used for DNN model development in the second step and for the optimization of the hyperparameters of the DNN models. Model evaluation was performed afterwards (third step) using the test set and additional datasets with experimental partition coefficients.

3. Results and discussion

3.1 DNN model development

The curated dataset contained 6364 chemicals. Not all descriptor sets of these chemicals were complete. Thus, the number of chemicals included in the training of a particular solute descriptor differs (see SI 1). We developed two different types of deep learning models based

on the respective datasets: singletask and multitask models. Singletask models are trained only for one variable (in our case one solute descriptor) and multitask models include various variables (in our case the set of all solute descriptors) in the training. The number of chemicals included in the development of multitask models is lower compared to all other models, because only complete datasets of solute descriptors could be used for multitask model development.

We further applied data augmentation based on the generation of all tautomeric forms of a chemical as input to generate the graphs for the DNNs to improve the prediction outcomes of the models based on the relatively small chemical dataset. Thus, we developed four different models: singletask models with (DNN_{taut}) and without (DNN_{mono}) data augmentation and multitask models with and without data augmentation. We optimized the hyperparameters for the training of the four different models, evaluating our optimization results based on the validation set. The root mean square errors *rmse*s for the training of the models over epochs for the finally selected single- and multitask models are given in SI 2. Data augmentation, for example by the use of multiple SMILES strings for one molecule, increases dataset size and can thus improve model performance and robustness[37, 38].

3.2 Performance of the developed DNN models

We used the test set as an independent set to compare the different models. The results for the predictions of the DNN models are given in Table 1 and SI 3. The multitask models performed worse than the singletask models for the *E*, *S*, *A*, and *B* descriptors. The *rmse* values for the *L* descriptor are lower for both multitask models. Since the solute descriptors are not entirely independent from each other and are determined as a complete descriptor set, we expected that multitask models would perform better compared to singletask models. But, only complete descriptor sets can be used for the training of multitask models. Thus, the number of chemicals in the training set was reduced to 3864 for the multitask DNN models. The reduction of the number of chemicals included in the training set for the multitask models is mainly caused by the *L* descriptor, which is only available for 4011 chemicals of the training set (compared to 4582 chemicals where the *S* descriptor is available, e.g.). Thus, we would expect a positive impact on the *rmse* of the *L* descriptor using the multitask models compared to the singletask models, which is the case. At the same time, we see a negative impact for the other descriptors due to the substantial reduction of the number of chemicals in the training set. To confirm this, we also trained singletask models with the training set used for the multitask models (SI 3, Supplementary Table 1). The *rmse* values are in the same range or slightly better for the models trained on the reduced dataset of chemicals compared to the multitask models but still worse

compared to the initial singletask models. Only for the *L* descriptor, the performance is worse which confirms our findings.

DNN_{taut} and DNN_{mono} perform equally well despite the increased dataset of DNN_{taut} due to data augmentation. The reason might be the well-curated dataset, which presumably already states the most likely tautomeric forms. One might suspect a bias in performance towards chemicals with several tautomeric forms for DNN_{taut} because these chemicals are overrepresented in its training. However, we detected no bias when analyzing the performance of chemicals with one and chemicals with multiple tautomers in the test set (SI 3, Supplementary Table 2).

We further compared our results to two QSPR models for the prediction of solute descriptors: the QSPR of ACD/Absolv[26] and the QSPR of LSERD[24]. Both QSPR models are based on molecular fragment contributions, whereas ACD/Absolv is an optimized version of the group contribution approach by Platts et al.[39]. The group contribution approach of LSERD is an adapted version of the group contribution approach by T. N. Brown[40]. As can be seen from Table 1, the *rmse*s of the singletask models for the solute descriptors are in the same range as the predictions based on the QSPR models from LSERD and ACD/Absolv. We performed a detailed analysis of the prediction error of the four different DNN models and the QSPR models (Figure 2). In general, the *rmse*s of the solute descriptors depend on the range of values of each solute descriptor. For *A*, *B*, and *E*, the range is smaller than for the *S* and *L* descriptor. For chemicals with a larger number of non-hydrogen atoms NHA (>20), the corresponding *rmse* of the descriptor predictions is noticeably higher for all models (Figure 2A). The QSPR models predict the *A* descriptors more precisely since *A* is set to zero if no hydrogen bond acceptor is available in the structure, which is not the case for the DNN models. Thus, DNN models perform slightly worse. Nitrogen containing chemicals seem to be problematic for predicting *A* for all models (Figure 2B), but still, the errors are in a small range. Predictions of the *B* descriptor by the singletask DNN models and the QSPR/LSERD model are in a close range, and the QSPR model is slightly better for chemicals with NHA>25. Interestingly, DNN_{mono} performs slightly better than DNN_{taut}. The QSPR ACD/Absolv performs slightly worse for oxygen containing chemicals. But, ACD/Absolv performs best in the prediction of the *E* descriptor. In fact, the *E* descriptor of the experimental datasets was often calculated using the ACD software. Thus, the performance of the QSPR ACD/Absolv should be better in this case. The QSPR LSERD is slightly worse in predicting the *E* descriptor of halogen containing chemicals. QSPR LSERD predicts negative *E* descriptors for polyfluorinated chemicals, resulting in larger prediction errors. The performance in predicting the *L* descriptor for chemicals with NHA >5 is also worse for QSPR LSERD. *Rmse* values of the predictions from

QSPR LSERD are especially larger for oxygen and nitrogen containing chemicals and chemicals with various heteroatoms. While both QSPR models perform exceptionally well in predicting *L* for C, H containing, and halogen containing chemicals and worse for all other classes, the DNN models deliver stable prediction results for all classes of chemicals. For the *S* descriptor, QSPR LSERD performs worse than the other tools for oxygen and nitrogen containing chemicals and chemicals with various heteroatoms, while the singletask DNNs and QSPR ACD/Absolv are in a comparable range. Overall, there is no clear indication that one prediction tool delivers systematically better results on all descriptors than the other tools.

Table 1 The root mean square error (*rmse*) and corresponding variance (*sdev*) are given for the predictions of the test set chemicals of the two different singletask DNNs (DNN_{mono} and DNN_{taut}) and multitask DNNs compared to the QSPR model of LSERD and the QSPR model of ACD/Absolv. Note that the mean value and variance were estimated using bootstrapping according to Vorberg and Tetko[41].

model	number of chemicals in the test set	singletask				multitask				LSERD		ACD/Absolv	
		DNN_{mono}		DNN_{taut}		DNN_{mono}		DNN_{taut}		QSPR			
solute descriptor		*rmse*	*sdev*	*rmse*	*sdev*	*rmse*	*sdev*	*rmse*	*sdev*	*rmse*	*sdev*	*rmse*	*sdev*
E	635	0.13	0.01	0.12	0.01	0.16	0.01	0.18	0.01	0.14	0.04	0.10	0.01
S	636	0.21	0.01	0.22	0.01	0.24	0.01	0.26	0.01	0.28	0.02	0.23	0.01
A	635	0.12	0.01	0.11	0.01	0.14	0.01	0.15	0.01	0.09	0.01	0.09	0.01
B	607	0.13	0.01	0.14	0.01	0.18	0.01	0.19	0.01	0.13	0.01	0.16	0.01
L	549	0.46	0.03	0.42	0.02	0.36	0.01	0.38	0.02	0.52	0.04	0.44	0.03

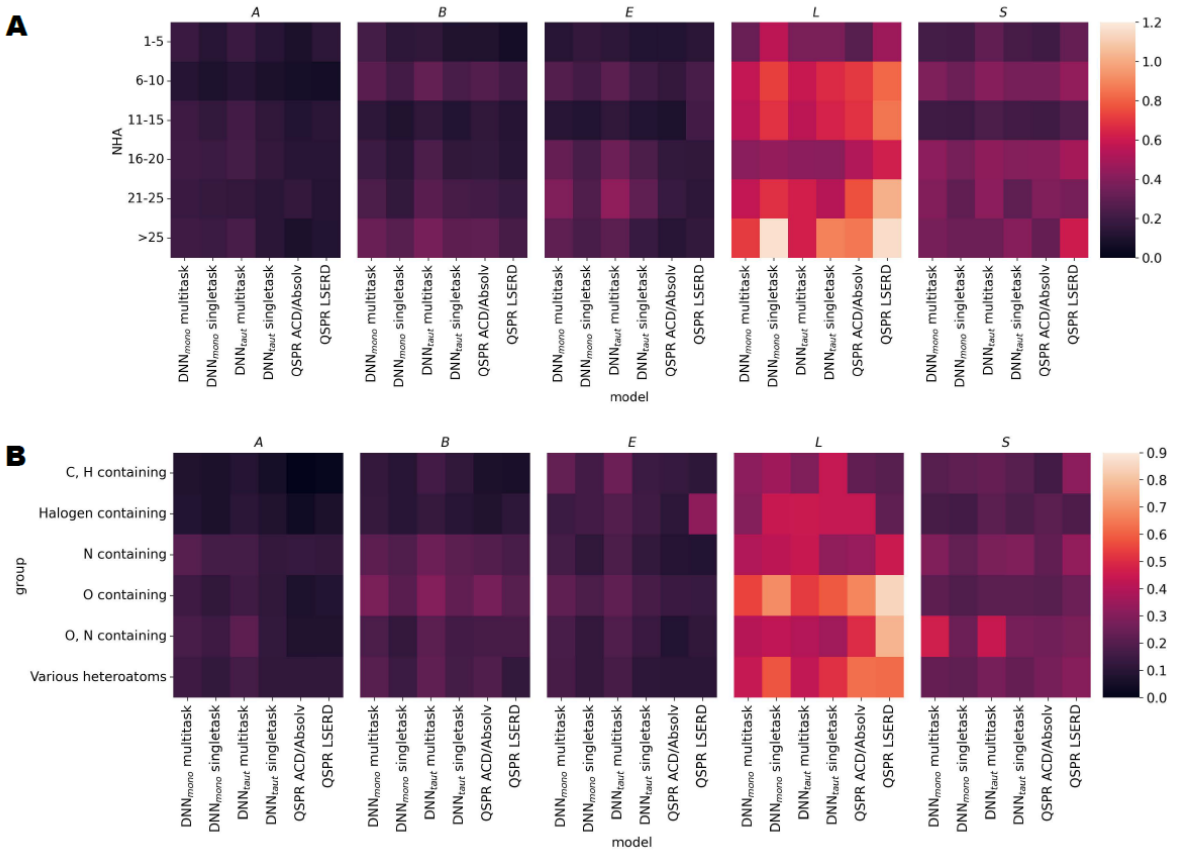


Figure 2 Prediction errors for different groups of chemicals of the test set of the four DNN models and the two QSPR models. R*mse* values are calculated for different groups of chemicals. Groups are generated based on the composition of the chemicals: A) chemicals grouped according to the number of non-hydrogen atoms NHA B) chemicals grouped according to different heteroatoms in the molecule.

Since both QSPR models are trained on the Abraham Absolv dataset, we cannot be sure that the respective chemicals in our test set were not partially or fully included in the training of the two QSPR models and the partial better performance in descriptor prediction is evoked by this. We therefore used different datasets of experimental partition coefficients and retention parameters for a further evaluation of our models and comparison to the two QSPR models (Table 2, SI 4). We selected five independent datasets, three of these are datasets of partition coefficients, namely the K_{ow}, K_{wa}, and K_{oa}. The datasets are taken from Mansouri et al.[42], who curated the datasets with respect to the chemical identifier. The K_{ow} dataset was further curated in a prior study concerning errors in the experimentally determined K_{ow} values[38]. Additionally, we used retention data for liquid chromatography (LC) and gas chromatography (GC), both measured in solvent and temperature gradient mode, respectively. LC data are given

as *CHI* values (chromatographic hydrophobicity index)[22], GC data are provided as Kováts retention indices *KRI*[43].

It can be seen that the DNN_{taut} performs equally well as the DNN_{mono}. Again, the singletask DNN models perform better than the multitask DNN models. The predictions made by the QSPR models are in a similar range when comparing the *rmse* values. In some cases, like the log K_{wa} or the *CHI*, for example, the QSPR model of LSERD performs better, and in other cases, the QSPR model of ACD/Absolv is better (e.g., log K_{oa}). However, the results do not indicate which model is the best choice for descriptor predictions.

Table 2 The root mean square error (*rmse*) and corresponding variance (*sdev*) are given for the predictions of partition coefficients and retention data based on LSER equations of the different DNNs compared to the QSPR model of LSERD and the QSPR model of ACD/Absolv. Note that the units of K_{ow}, K_{oa}, and K_{wa} are [$L_{water}/L_{octanol}$], [$L_{air}/L_{octanol}$], and [L_{air}/L_{water}], respectively. K_{wa} values were originally given as HL values (air-water, [m^3 atm/mole]) and converted according to Sander[44]. Note that the mean value and variance were estimated using bootstrapping according to Vorberg and Tetko[41].

dataset	model	DNN_{mono}		DNN_{taut}		QSPR LSERD		QSPR ACD	
	no. chemicals	*rmse*	*sdev*	*rmse*	*sdev*	*rmse*	*sdev*	*rmse*	*sdev*
log K_{ow} Eq. (1)	12010	1.01	0.02	1.04	0.02	0.91	0.01	0.87	0.01
log K_{ow} Eq. (3)	12010	0.86	0.01	0.89	0.01	1.11	0.01	0.90	0.01
log K_{oa} Eq. (1)	270	0.60	0.04	0.63	0.05	0.49	0.07	0.59	0.07
log K_{oa} Eq. (3)	270	0.59	0.04	0.61	0.05	0.50	0.08	0.59	0.07
log K_{wa} Eq. (1)	696	1.46	0.07	1.36	0.08	1.28	0.07	1.39	0.09
log K_{wa} Eq. (3)	696	1.36	0.05	1.24	0.07	1.20	0.07	1.34	0.12
KRI	454	127	7	129	7	104	10	120	7
CHI	204	4.58	0.29	4.49	0.29	5.52	0.55	3.19	0.14

3.3 Restrictions in the application of the models

We decided to have a closer look at the details of the predictions. First, we determined the NHA of the chemicals represented in the different datasets for descriptor prediction and the prediction of partition coefficients and chromatographic indices. We grouped the chemicals in the different datasets according to their NHA and determined the respective *rmse* values for specific NHA ranges. The results can be seen in Figure 3 for the predictions of the log K_{ow} and in SI 5 and 6 for the other datasets. Again, the *rmse* values increase for higher NHA. But still, there is no clear indication for one model performing better compared to the other tools. QSPR LSERD, e.g., performs best in predicting the *KRI* of chemicals with NHA >25 but performs worst in the prediction of *CHI*. For the log K_{ow} dataset, the DNN and QSPR models do not differ in their performance for chemicals with NHA lower than 25.

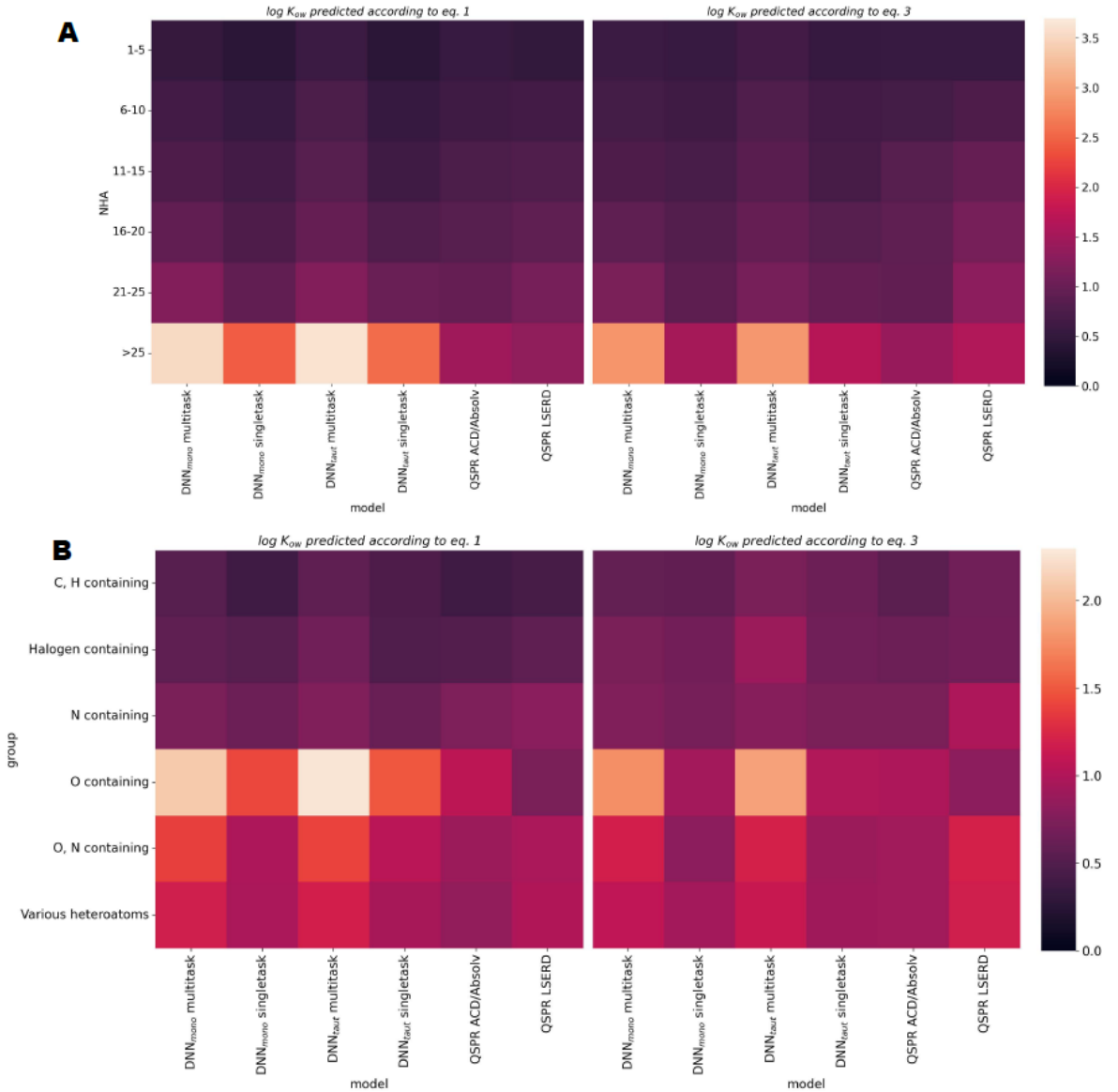


Figure 3 *Rmse*s of the log K_{ow} predictions based on the different DNN approaches and QSPR depending on (A) the number of non-hydrogen atoms NHA and (B) the composition of the

molecule. The log K_{ow} is predicted based on two different LSER equations (1 – left heat map, 2 – right heat map). The respective ranges in the heat maps are given independently for (A) and (B). The heat maps of the other datasets are shown in SI 5 and 6.

The reason for the larger *rmse*s for chemicals with larger NHA is the Absolv dataset of experimental solute descriptors itself. Most of the chemicals included in this dataset are small chemicals with no or only one functional group and smaller NHAs (SI 7). As a result, DNN and QSPR models, both trained on the Absolv dataset, have a limited application domain. As can be seen in SI 7 (Figure S30), many chemicals of the external datasets lie well beyond this application range.

The fragmental-based approaches of the QSPR models partially cover chemicals with NHA greater than 25. They perform, therefore, slightly better than the DNN models. Nevertheless, all the models should be used with care when applying them for chemicals with NHA greater than 25.

Again, there is no clear trend in the *rmse* values when the chemicals are grouped according to different atom types in the molecules. For the log K_{oa} and log K_{wa}, e.g., predicting partition coefficients for chemicals with various heteroatoms (which means multiple functional groups) is problematic. The DNN models show problems for the prediction of log K_{ow} for oxygen containing chemicals. And QSPR LSERD shows larger *rmse* values on the predictions of the *CHI* dataset. There is no clear indication of which model should be used for predictions in general.

Thus, we recommend applying different prediction tools (if available), like a direct prediction of the log K_{ow} by a DNN approach[38], to reduce the prediction error. An extension of the Absolv dataset for chemicals with NHA greater than 25 is problematic from our perspective. Some of these chemicals are extremely hydrophobic chemicals, which are problematic in experimental setups for descriptor determination. Another problem is that these larger molecules often are chemicals with multifunctional groups, for which the measurement and the prediction of the pK_as can be extremely erroneous. As a consequence the chemical might not be present in its neutral form during the measurement, but as an ion. The classical LSER approach cannot cover ionic chemicals.

4. Conclusions

In general, the error based on the LSER prediction models is relatively high, with an overall *rmse* of ~ 1 log unit (according to our selected datasets). The error results from errors in the

prediction of each solute descriptor itself and the error of the LSER equations. DNN models are complementary prediction methods that overcome problems of the group contribution approaches such as predicting negative *E* descriptors for fluorinated chemicals. They, therefore, offer an independent prediction of the descriptors and partition coefficients. For chemicals with large NHA and multiple functionalities, users should be aware that the prediction error might be significant (independent of the tool used). The novel prediction methods do not solve these problems since the limitation is given by the available experimental dataset of the descriptors. If within their application domain, LSER models are nevertheless helpful for the prediction of partition coefficients especially where no other tools are available.

Appendix A

Supporting Information

The supporting information is available free of charge at https://

Additional information on the dataset, DNN model development, performance of DNNs and QSPR models. (PDF)

Author Information

Corresponding Author

Nadin Ulrich – Department of Ecological Chemistry, Helmholtz Centre for Environmental Research– UFZ, 04318 Leipzig, Germany; https://orcid.org/0000-0002-1267-0429; Email: nadin.ulrich@ufz.de

Authors

Nadin Ulrich – Department of Analytical Environmental Chemistry, Helmholtz Centre for Environmental Research – UFZ, 04318 Leipzig, Germany

Department of Ecological Chemistry, Helmholtz Centre for Environmental Research – UFZ, 04318 Leipzig, Germany

Andrea Ebert – Department of Analytical Environmental Chemistry, Helmholtz Centre for Environmental Research – UFZ, 04318 Leipzig, Germany; https://orcid.org/0000-0002-1095-0259

CRediT authorship contribution statement

N.U. and A.E. conceived and designed the study, corrected the dataset, performed the calculations, interpreted the data and wrote the manuscript.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We thank Karsten Voigt and Sebastian Klaus for support in Python. We thank Kai-Uwe Goss for helpful comments on our manuscript.

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

References

[1] F.A.P.C. Gobas, B.C. Kelly, J.A. Arnot, Quantitative structure activity relationships for predicting the bioaccumulation of POPs in terrestrial food-webs, Qsar Comb Sci, 22 (2003) 329-336.

[2] D.C.G. Muir, P.H. Howard, Are There Other Persistent Organic Pollutants? A Challenge for Environmental Chemists, Environ Sci Technol, 40 (2006) 7157-7166.

[3] S.H. Hilal, S.N. Ayyampalayam, L.A. Carreira, Air−Liquid Partition Coefficient for a Diverse Set of Organic Compounds: Henry's Law Constant in Water and Hexadecane, Environ Sci Technol, 42 (2008) 9231-9236.

[4] S. Endo, P. Grathwohl, S.B. Haderlein, T.C. Schmidt, LFERs for Soil Organic Carbon−Water Distribution Coefficients (KOC) at Environmentally Relevant Sorbate Concentrations, Environ Sci Technol, 43 (2009) 3094-3100.

[5] K. Halbach, N. Ulrich, K.-U. Goss, B. Seiwert, S. Wagner, S. Scholz, T. Luckenbach, C. Bauer, N. Schweiger, T. Reemtsma, Yolk Sac of Zebrafish Embryos as Backpack for Chemicals?, Environ Sci Technol, 54 (2020) 10159-10169.

[6] S. Endo, T.N. Brown, K.U. Goss, General Model for Estimating Partition Coefficients to Organisms and Their Tissues Using the Biological Compositions and Polyparameter Linear Free Energy Relationships, Environ Sci Technol, 47 (2013) 6630-6639.
[7] S. Endo, K.U. Goss, Applications of Polyparameter Linear Free Energy Relationships in Environmental Chemistry, Environ Sci Technol, 48 (2014) 12477-12491.
[8] M. Vitha, P.W. Carr, The chemical interpretation and practice of linear solvation energy relationships in chromatography, J Chromatogr A, 1126 (2006) 143-194.
[9] M.H. Abraham, W.E. Acree, Equations for the Transfer of Neutral Molecules and Ionic Species from Water to Organic phases, J Org Chem, 75 (2010) 1006-1015.
[10] K.-U. Goss, Predicting the equilibrium partitioning of organic compounds using just one linear solvation energy relationship (LSER), Fluid Phase Equilibr, 233 (2005) 19-22.
[11] M.H. Abraham, J. Andonian-Haftvan, G.S. Whiting, A. Leo, R.S. Taft, Hydrogen bonding. Part 34. The factors that influence the solubility of gases and vapours in water at 298 K, and a new method for its determination, Journal of the Chemical Society, Perkin Transactions 2, (1994) 1777-1791.
[12] M.H. Abraham, R.E. Smith, R. Luchtefeld, A.J. Boorem, R. Luo, W.E. Acree, Prediction of Solubility of Drugs and Other Compounds in Organic Solvents, Journal of Pharmaceutical Sciences, 99 (2010) 1500-1515.
[13] G. Bronner, K.U. Goss, Predicting Sorption of Pesticides and Other Multifunctional Organic Chemicals to Soil Organic Carbon, Environ Sci Technol, 45 (2011) 1313-1319.
[14] C. Niederer, R.P. Schwarzenbach, K.-U. Goss, Elucidating Differences in the Sorption Properties of 10 Humic and Fulvic Acids for Polar and Nonpolar Organic Chemicals, Environ Sci Technol, 41 (2007) 6711-6717.
[15] L. Sprunger, A. Proctor, W.E. Acree, M.H. Abraham, Characterization of the sorption of gaseous and organic solutes onto polydimethyl siloxane solid-phase microextraction surfaces using the Abraham model, J Chromatogr A, 1175 (2007) 162-173.
[16] S. Endo, S.T.J. Droge, K.U. Goss, Polyparameter Linear Free Energy Models for Polyacrylate Fiber-Water Partition Coefficients to Evaluate the Efficiency of Solid-Phase Microextraction, Anal Chem, 83 (2011) 1394-1400.
[17] S. Endo, K.U. Goss, Serum Albumin Binding of Structurally Diverse Neutral Organic Compounds: Data and Models, Chemical Research in Toxicology, 24 (2011) 2293-2301.
[18] S. Endo, J. Bauerfeind, K.U. Goss, Partitioning of Neutral Organic Compounds to Structural Proteins, Environ Sci Technol, 46 (2012) 12697-12703.

[19] S. Endo, B.I. Escher, K.U. Goss, Capacities of membrane lipids to accumulate neutral organic chemicals, Environ Sci Technol, 45 (2011) 5912-5921.
[20] A. Geisler, S. Endo, K.U. Goss, Partitioning of Organic Chemicals to Storage Lipids: Elucidating the Dependence on Fatty Acid Composition and Temperature, Environ Sci Technol, 46 (2012) 9519-9524.
[21] N. Ulrich, J. Mühlenberg, H. Retzbach, G. Schüürmann, W. Brack, Linear solvation energy relationships as classifiers in non-target analysis – A gas chromatographic approach, J Chromatogr A, 1264 (2012) 95-103.
[22] N. Ulrich, G. Schüürmann, W. Brack, Linear Solvation Energy Relationships as classifiers in non-target analysis—A capillary liquid chromatography approach, J Chromatogr A, 1218 (2011) 8192-8196.
[23] K.U. Goss, R.P. Schwarzenbach, Adsorption of a diverse set of organic vapors on quartz, CaCO3, and alpha-Al2O3 at different relative humidities, J Colloid Interf Sci, 252 (2002) 31-41.
[24] N. Ulrich, S. Endo, T.N. Brown, N. Watanabe, G. Bronner, M.H. Abraham, K.U. Goss, UFZ-LSER database v 3.2 [Internet], (2017).
[25] https://www.cas.org/about/cas-content [accessed on 09.06.2021], in.
[26] ACD/Percepta, (2015 Release www.acdlabs.com).
[27] Z. Wu, B. Ramsundar, Evan N. Feinberg, J. Gomes, C. Geniesse, A.S. Pappu, K. Leswing, V. Pande, MoleculeNet: a benchmark for molecular machine learning, Chem Sci, 9 (2018) 513-530.
[28] C.W. Coley, R. Barzilay, W.H. Green, T.S. Jaakkola, K.F. Jensen, Convolutional Embedding of Attributed Molecular Graphs for Physical Property Prediction, J Chem Inf Model, 57 (2017) 1757-1772.
[29] A. Lusci, G. Pollastri, P. Baldi, Deep Architectures and Deep Learning in Chemoinformatics: The Prediction of Aqueous Solubility for Drug-Like Molecules, Journal of Chemical Information and Modeling, 53 (2013) 1563-1575.
[30] M. Rupp, R. Ramakrishnan, O.A. von Lilienfeld, Machine Learning for Quantum Mechanical Properties of Atoms in Molecules, The Journal of Physical Chemistry Letters, 6 (2015) 3309-3313.
[31] S.S.S. Justin Gilmer, Patrick F. Riley, Oriol Vinyals , George E. Dahl, Neural Message Passing for Quantum Chemistry, (2017).

[32] L.H. Felix A. Faber, Bing Huang, Justin Gilmer, Samuel S. Schoenholz, George E. Dahl, Oriol Vinyals, Steven Kearnes, Patrick F. Riley, O. Anatole von Lilienfeld, Machine learning prediction errors better than DFT accuracy, Chemical Physics, (2017).

[33] ChemAxon, JChem for Excel v. 20.6.0.618, (http://www.chemaxon.com), (2020).

[34] B. Ramsundar, Eastman P, Walters P, Pande V, Leswing K, Wu Z, Deep Learning for the Life Sciences, O'Reilly Media, 2019.

[35] B. Ramsundar, Democratizing Deep-Learning for Drug Discovery, Quantum Chemistry, Materials Science and Biology, GitHub repository - https://github.com/deepchem/deepchem, (2016).

[36] Python Software Foundation. Python Language Reference. Available at http://www.python.org.

[37] J.E. Bjerrum, SMILES Enumeration as Data Augmentation for Neural Network Modeling of Molecules, in, 2017, pp. arXiv:1703.07076.

[38] N. Ulrich, K.-U. Goss, A. Ebert, Exploring the octanol–water partition coefficient dataset using deep learning techniques and data augmentation, Communications Chemistry, 4 (2021) 90.

[39] J.A. Platts, D. Butina, M.H. Abraham, A. Hersey, Estimation of Molecular Linear Free Energy Relation Descriptors Using a Group Contribution Approach, Journal of Chemical Information and Computer Sciences, 39 (1999) 835-845.

[40] T.N. Brown, Predicting hexadecane–air equilibrium partition coefficients (L) using a group contribution approach constructed from high quality data, SAR and QSAR in Environmental Research, 25 (2014) 51-71.

[41] S. Vorberg, I.V. Tetko, Modeling the Biodegradability of Chemical Compounds Using the Online CHEmical Modeling Environment (OCHEM), Mol Inform, 33 (2014) 73-85.

[42] K. Mansouri, C.M. Grulke, R.S. Judson, A.J. Williams, OPERA models for predicting physicochemical properties and environmental fate endpoints, J Cheminform, 10 (2018) 10.

[43] N. Ulrich, G. Schüürmann, W. Brack, Prediction of gas chromatographic retention indices as classifier in non-target analysis of environmental samples, J Chromatogr A, 1285 (2013) 139-147.

[44] R. Sander, Compilation of Henry's law constants (version 4.0) for water as solvent, Atmos. Chem. Phys., 15 (2015) 4399-4981.

Can deep learning algorithms enhance the prediction of solute descriptors for linear solvation energy relationship approaches?

Nadin Ulrich[1,2*], Andrea Ebert[1]

[1]Department of Analytical Environmental Chemistry, Helmholtz Centre for Environmental Research - UFZ, Permoserstrasse 15, D-04318 Leipzig, Germany

[2]Department of Ecological Chemistry, Helmholtz Centre for Environmental Research - UFZ, Permoserstrasse 15, D-04318 Leipzig, Germany

*Corresponding Author: Phone + 49 341 235 1818 E-mail: nadin.ulrich@ufz.de

Table of Content

SI 1 Dataset for DNN model development

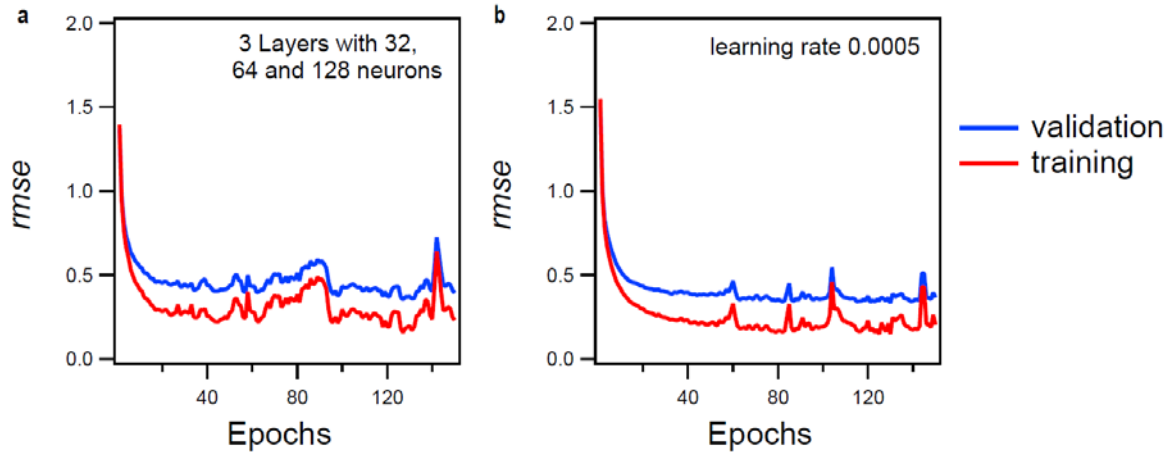
We did a preselection of the solute descriptors implemented in our study. We used the *S* descriptor for selection, because we believe it to be the most reliable descriptor. The solute descriptor *L* is often missing and cannot be used as criteria; the *B* descriptor cannot be determined by gas chromatography experiments, which are commonly used for the determination of solute descriptors. Thus, we cannot refer to this descriptor; the solute descriptor *E* is often calculated from molar refraction and not determined experimentally; and the *A* descriptor can be set to 0 in some cases (if no H-bond donors are available in a specific structure of a chemical) without any underlying experiments.

Supplementary Table 1 Number of chemicals in training set, validation set, and test set. Note that the split of the dataset was performed for the *S* descriptor. 10% of the chemicals were assigned to the test set beforehand, the remaining chemicals were split 80/20 into training set and validation set. Due to missing descriptors in datasets of chemicals, the number of chemicals in each set of a descriptor varies. The dataset including all descriptors for the training of the multitask model contains less chemicals. In this case, only complete sets of solute descriptors are included.

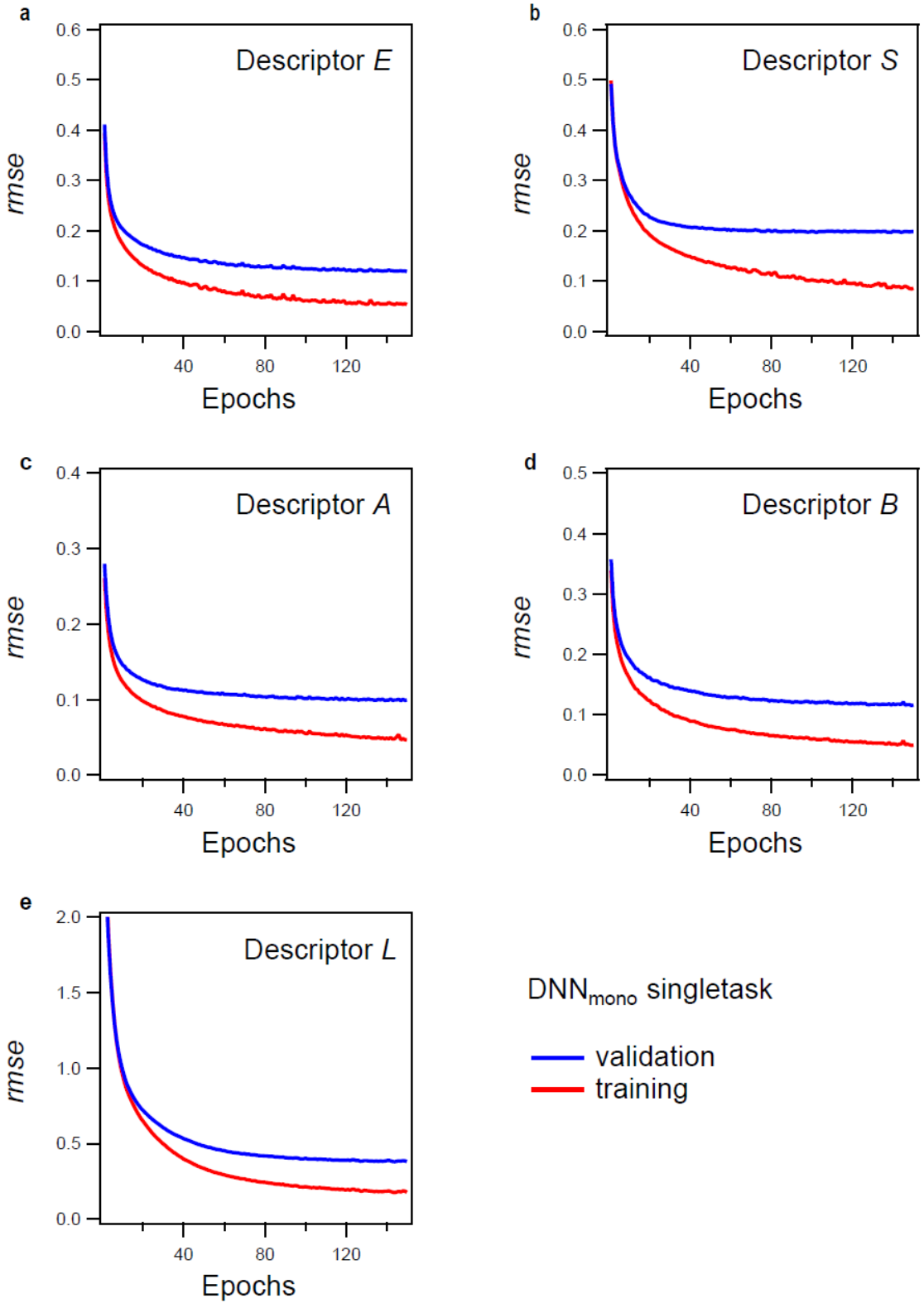
solute descriptor	training set	validation set	test set
E	4568	1144	635
S	4582	1146	636
A	4580	1144	635
B	4396	1108	607
L	4011	1012	549
all (multitask)	3864	986	singletask test sets used

SI 2 Details on the development of the DNN models for the prediction of solute descriptors

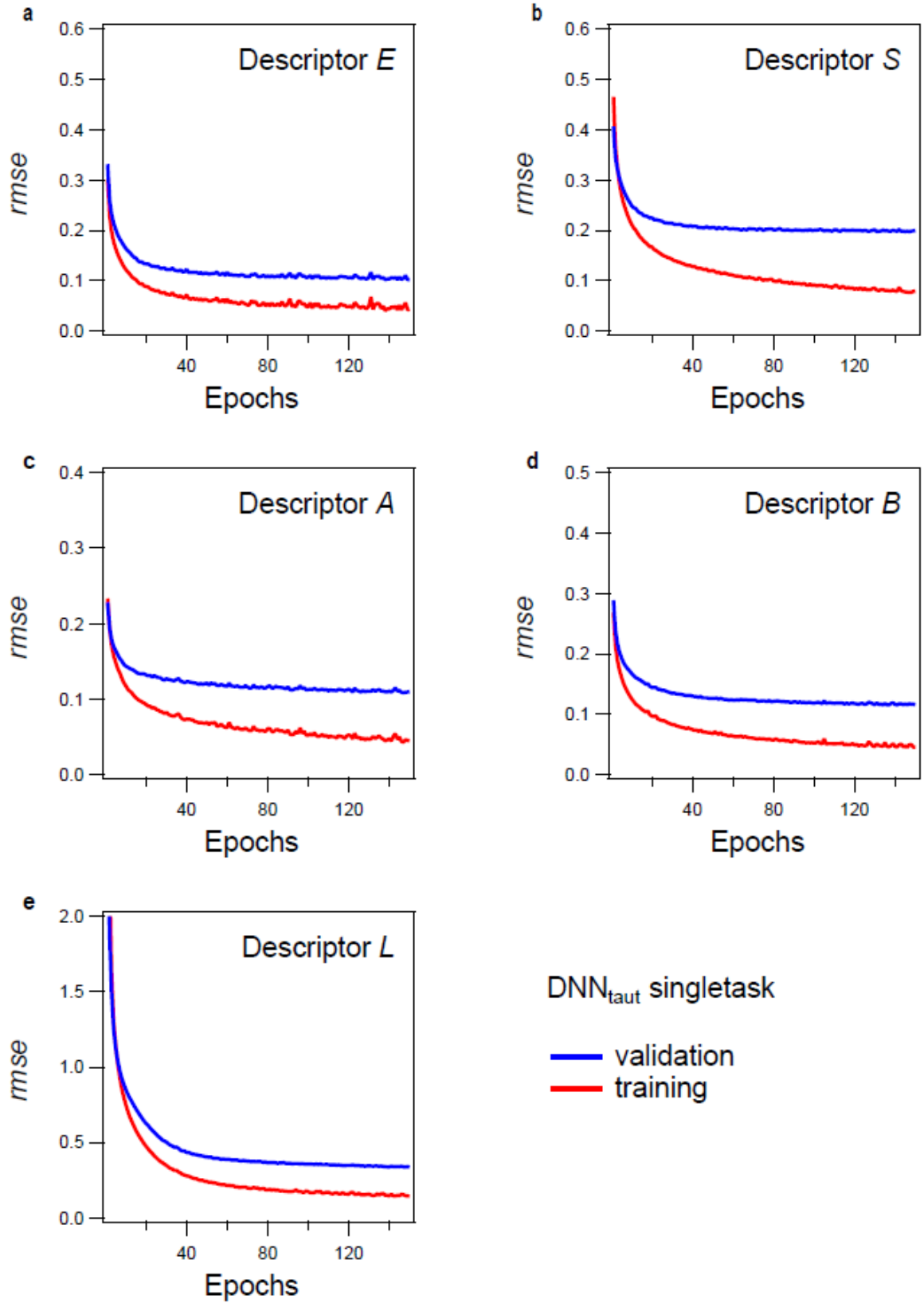
We tested different structures of neural networks with two and three layers and varied the number of neurons in the different layers as well as the learning rate. We used the *rmse* of the validation set to check which net is best performing. To choose the best performing parameters, we searched for the lowest validation *rmse*s while avoiding overfitting and unstable models. If the model had a too high capacity, see Supplementary Figure 1a for the training curves of a three layer model, or if the learning rate was chosen too high, see Supplementary Figure 1b for a fivefold increased learning rate, the model became unstable. We plotted the *rmse* values of the training and the validation set over the epochs (see Supplementary Figure 1 - 4) and stopped training at the point where the validation set *rmse* no longer decreased significantly. Further training at that point would lead to overfitting, increasing the gap between training and validation *rmse*.



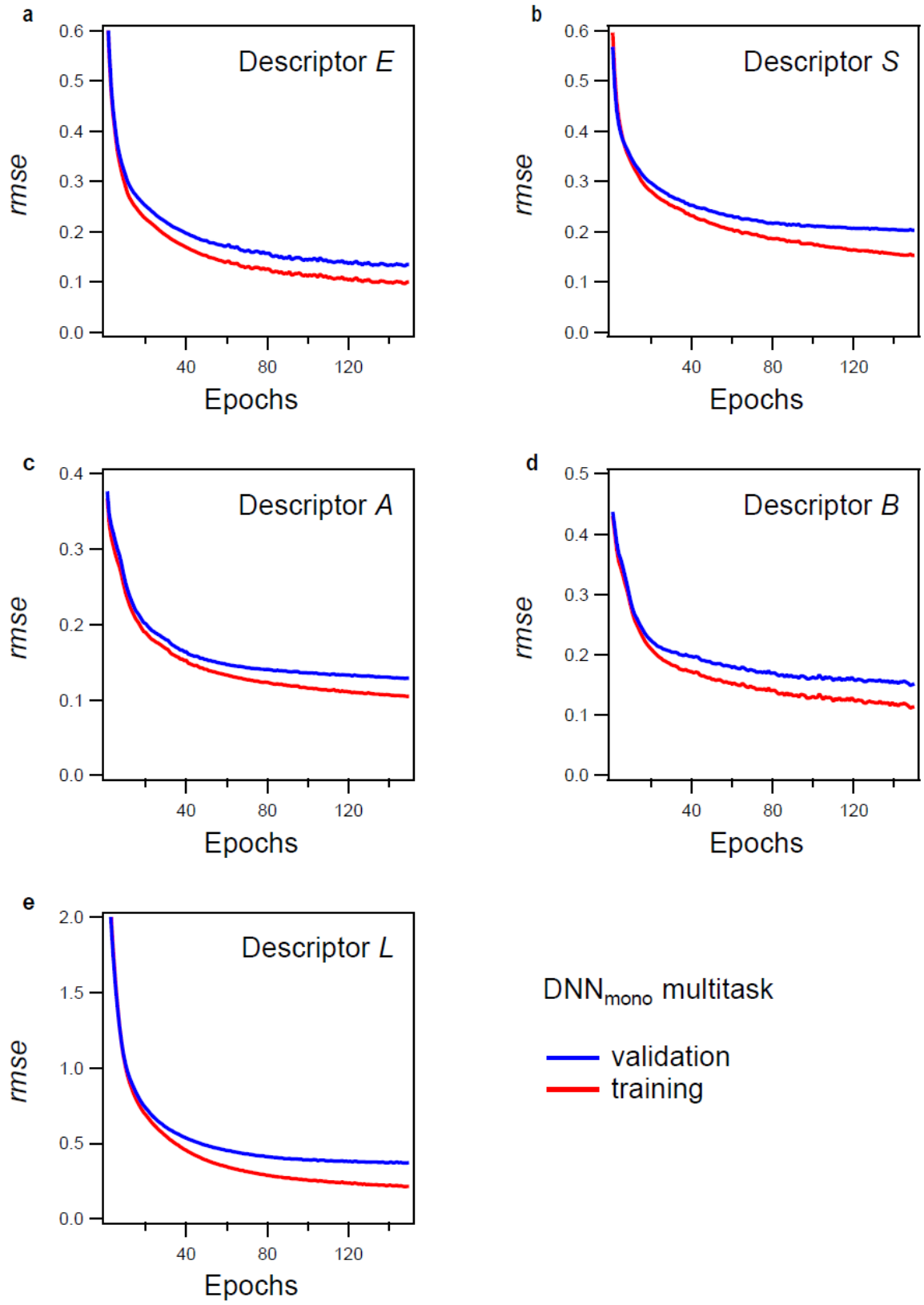
Supplementary Figure 1 Training set and validation set *rmse*s are given over the number of epochs as examples for unstable models. Training was performed using the canonical SMILES as input for graph generation in the singletask DNN models for the *L* descriptor for a high capacity net with three layers **(a)** and a net with a learning rate chosen too high **(b)**.



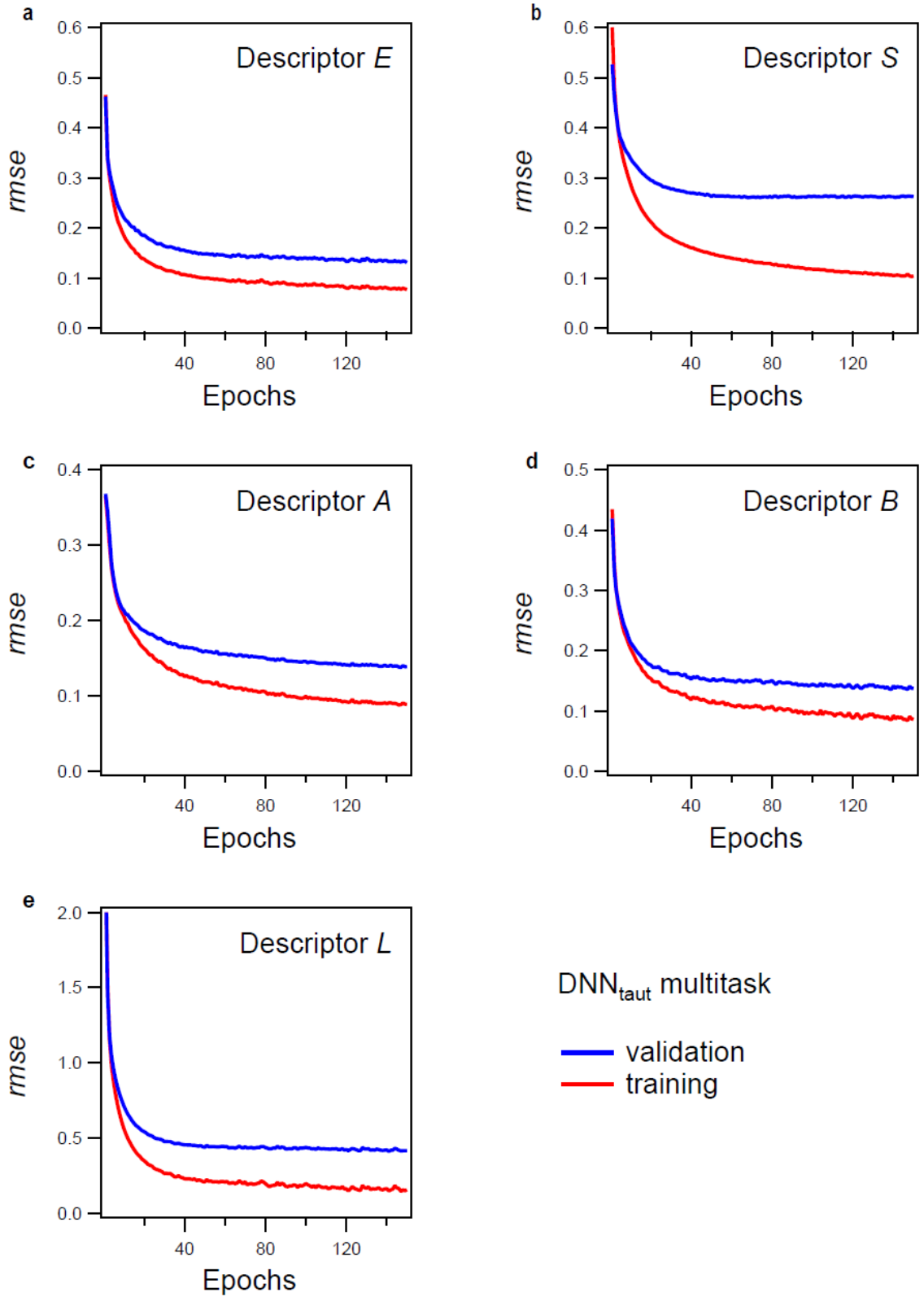
Supplementary Figure 2 Training set and validation set *rmse*s are given for each descriptor over the number of epochs. Training was performed using the canonical SMILES as input for graph generation in the singletask DNN models.



Supplementary Figure 3 Training set and validation set *rmses* are given for each descriptor over the number of epochs. Training was performed using all SMILES variants (including tautomers) as input for graph generation in the singletask DNN models.



Supplementary Figure 4 Training set and validation set *rmse*s are given for each descriptor over the number of epochs. Training was performed using the canonical SMILES as input for graph generation in the multitask DNN models.



Supplementary Figure 5 Training set and validation set *rmse*s are given for each descriptor over the number of epochs. Training was performed using all SMILES variants (including tautomers) as input for graph generation in the multitask DNN models.

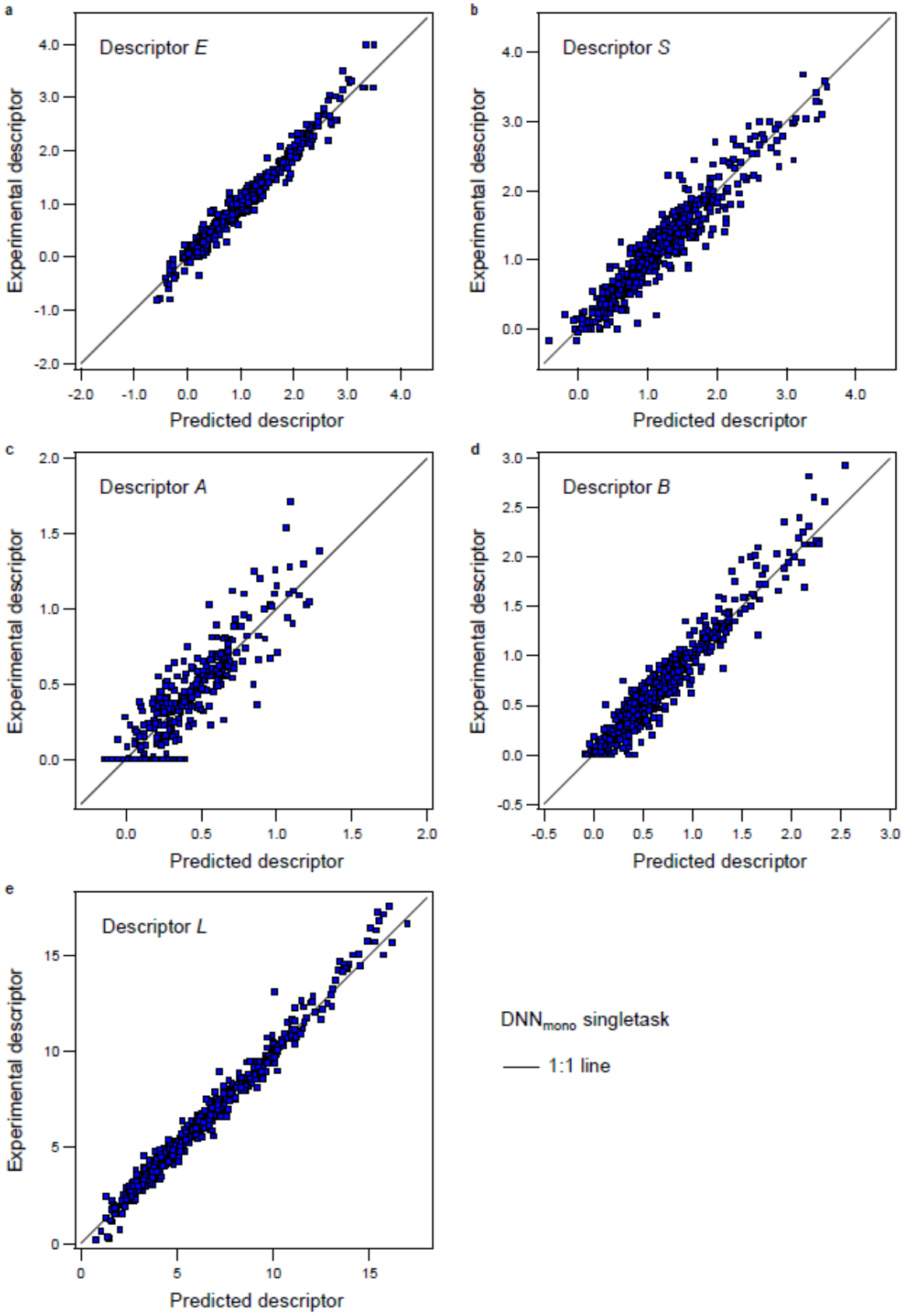
SI 3 Solute descriptor prediction by the singletask and multitask DNN models

Supplementary Table 1 *Rmse*s for the predictions of the test set chemicals of the different DNNs. Additionally singletask DNNs are trained based on the multitask training dataset. Note that the mean value and variance were estimated using bootstrapping according to Vorberg and Tetko[1].

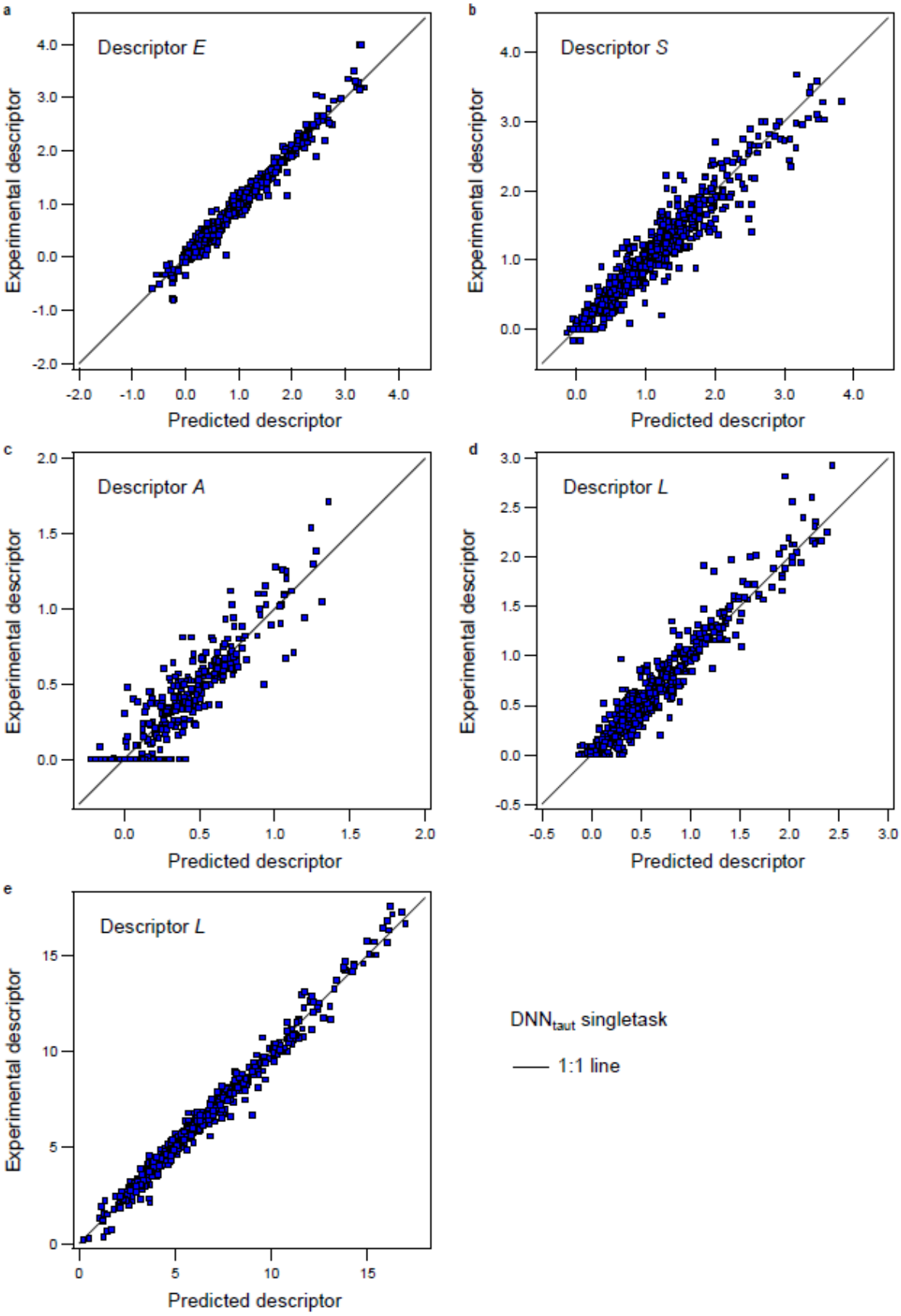
model	number of chemicals in the test set	singletask				multitask				singletask (trained on multitask dataset)			
		DNN_{mono}		DNN_{taut}		DNN_{mono}		DNN_{taut}		DNN_{mono}		DNN_{taut}	
solute descriptor		*rmse*	*sdev*	*rmse*	*sdev*	*rmse*	*sdev*	*rmse*	*sdev*	*rmse*	*sdev*	*rmse*	*sdev*
E	635	0.13	0.01	0.12	0.01	0.16	0.01	0.18	0.01	0.13	0.01	0.15	0.01
S	636	0.21	0.01	0.22	0.01	0.24	0.01	0.26	0.01	0.22	0.01	0.23	0.01
A	635	0.12	0.01	0.11	0.01	0.14	0.01	0.15	0.01	0.11	0.01	0.3	0.01
B	607	0.13	0.01	0.14	0.01	0.18	0.01	0.19	0.01	0.14	0.01	0.15	0.01
L	549	0.46	0.03	0.42	0.02	0.36	0.01	0.38	0.02	0.45	0.03	0.51	0.04

Supplementary Table 2 *Rmse*s for the predictions of the test set chemicals of the different DNNs compared to the QSPR model of LSERD and the QSPR model of ACD/Absolv, for compounds with only one and for compounds with more than one tautomeric form (tautomer count predicted using JChem). Note that the mean value and variance were estimated using bootstrapping according to Vorberg and Tetko[1].

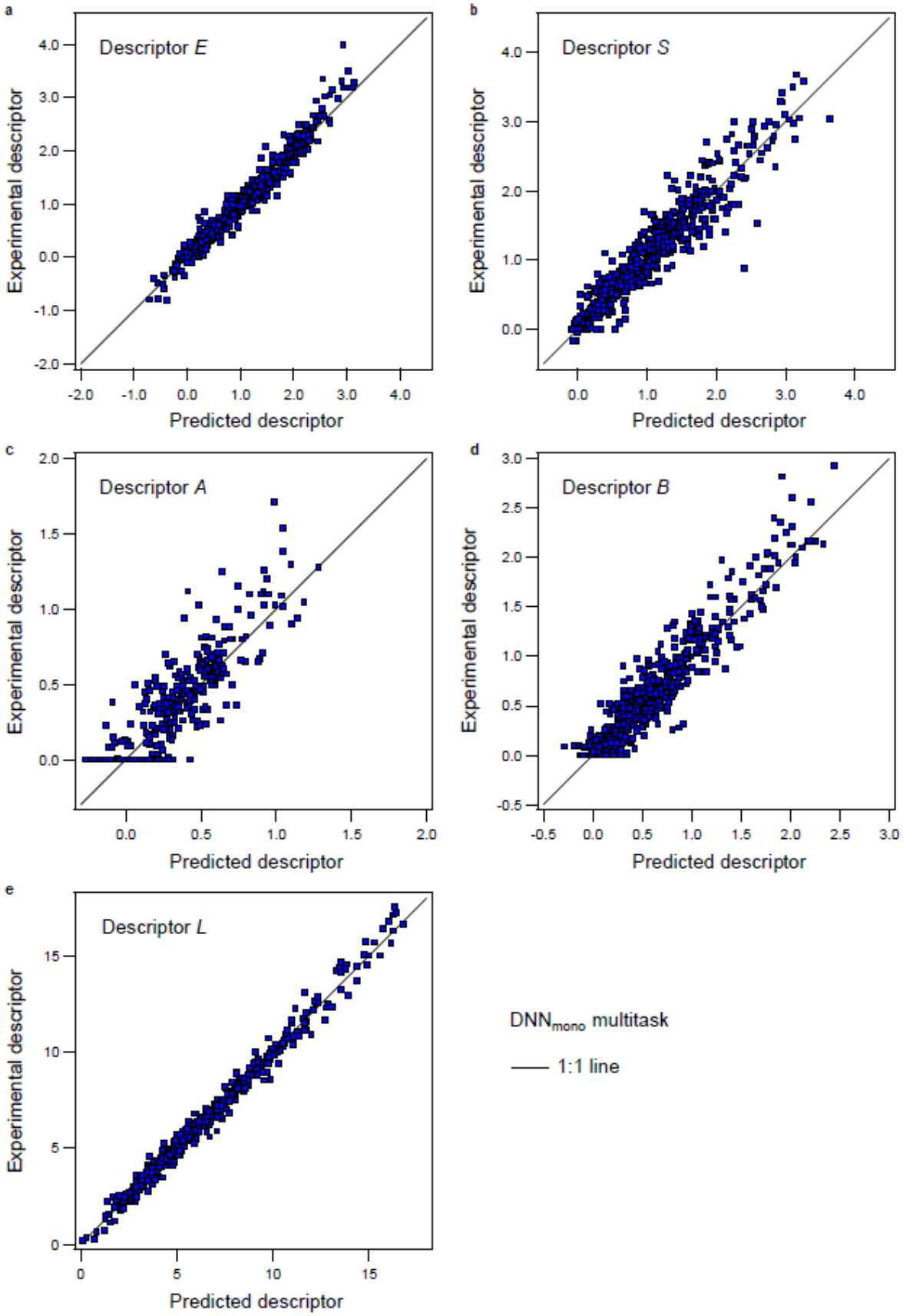
model	tautomer count	number of chemicals in the test set	singletask				multitask				LSERD		ACD/Absolv	
			DNN_{mono}		DNN_{taut}		DNN_{mono}		DNN_{taut}		QSPR			
solute descriptor			*rmse*	*sdev*	*rmse*	*sdev*	*rmse*	*sdev*	*rmse*	*sdev*	*rmse*	*sdev*	*rmse*	*sdev*
E	> 1	177	0.13	0.01	0.13	0.02	0.17	0.01	0.17	0.01	0.11	0.01	0.10	0.01
E	1	458	0.13	0.01	0.12	0.01	0.16	0.01	0.18	0.02	0.15	0.05	0.10	0.01
S	> 1	177	0.23	0.02	0.26	0.02	0.31	0.03	0.31	0.03	0.32	0.03	0.28	0.02
S	1	459	0.20	0.01	0.21	0.01	0.20	0.01	0.23	0.01	0.26	0.02	0.21	0.01
A	> 1	177	0.16	0.01	0.14	0.01	0.19	0.01	0.20	0.01	0.12	0.01	0.11	0.01
A	1	458	0.09	0.01	0.09	0.01	0.11	0.01	0.12	0.01	0.08	0.01	0.07	0.01
B	>1	175	0.15	0.01	0.15	0.01	0.20	0.01	0.20	0.01	0.15	0.01	0.10	0.01
B	1	432	0.12	0.01	0.14	0.01	0.17	0.01	0.18	0.01	0.12	0.01	0.16	0.01
L	>1	139	0.54	0.05	0.45	0.06	0.44	0.03	0.43	0.03	0.73	0.05	0.62	0.07
L	1	410	0.43	0.03	0.40	0.02	0.34	0.02	0.36	0.02	0.42	0.06	0.36	0.02



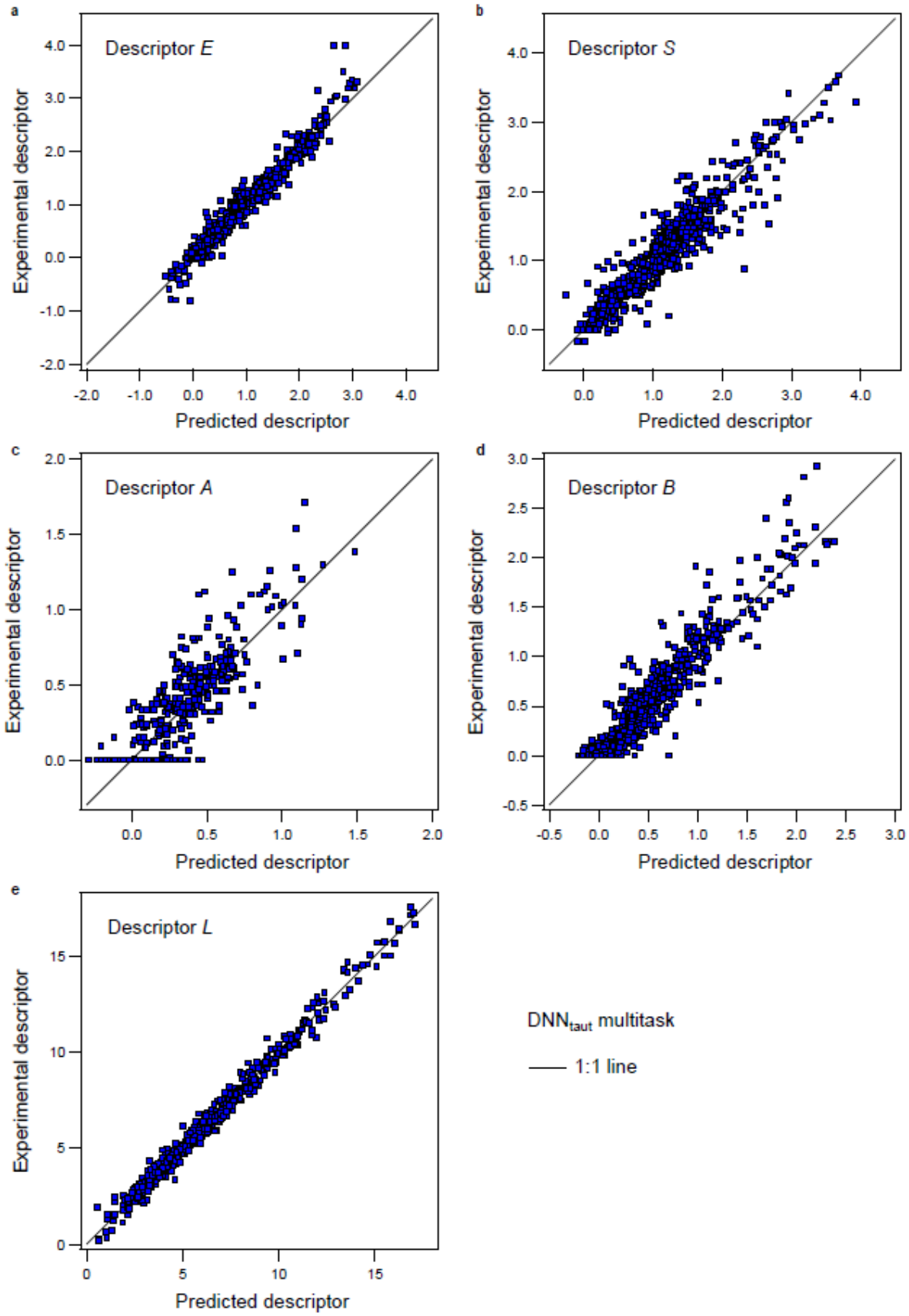
Supplementary Figure 6 Predictions of the solute descriptors *E* (**a**), *S* (**b**), *A* (**c**), *B* (**d**), and *L* (**e**) for the test set by the singletask model DNN_{mono}.



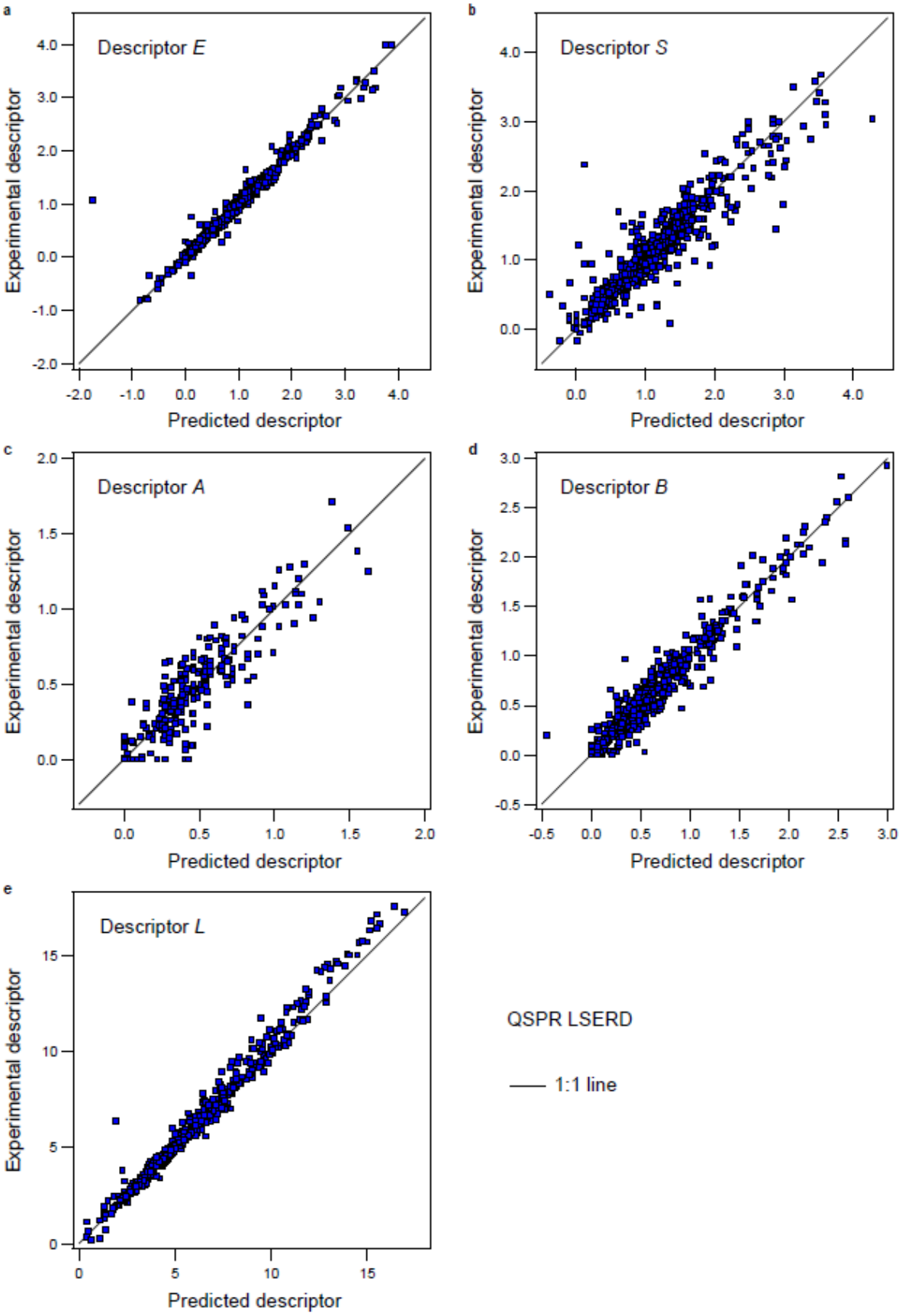
Supplementary Figure 7 Predictions of the solute descriptors *E* (**a**), *S* (**b**), *A* (**c**), *B* (**d**), and *L* (**e**) for the test set by the singletask model DNN_{taut}.



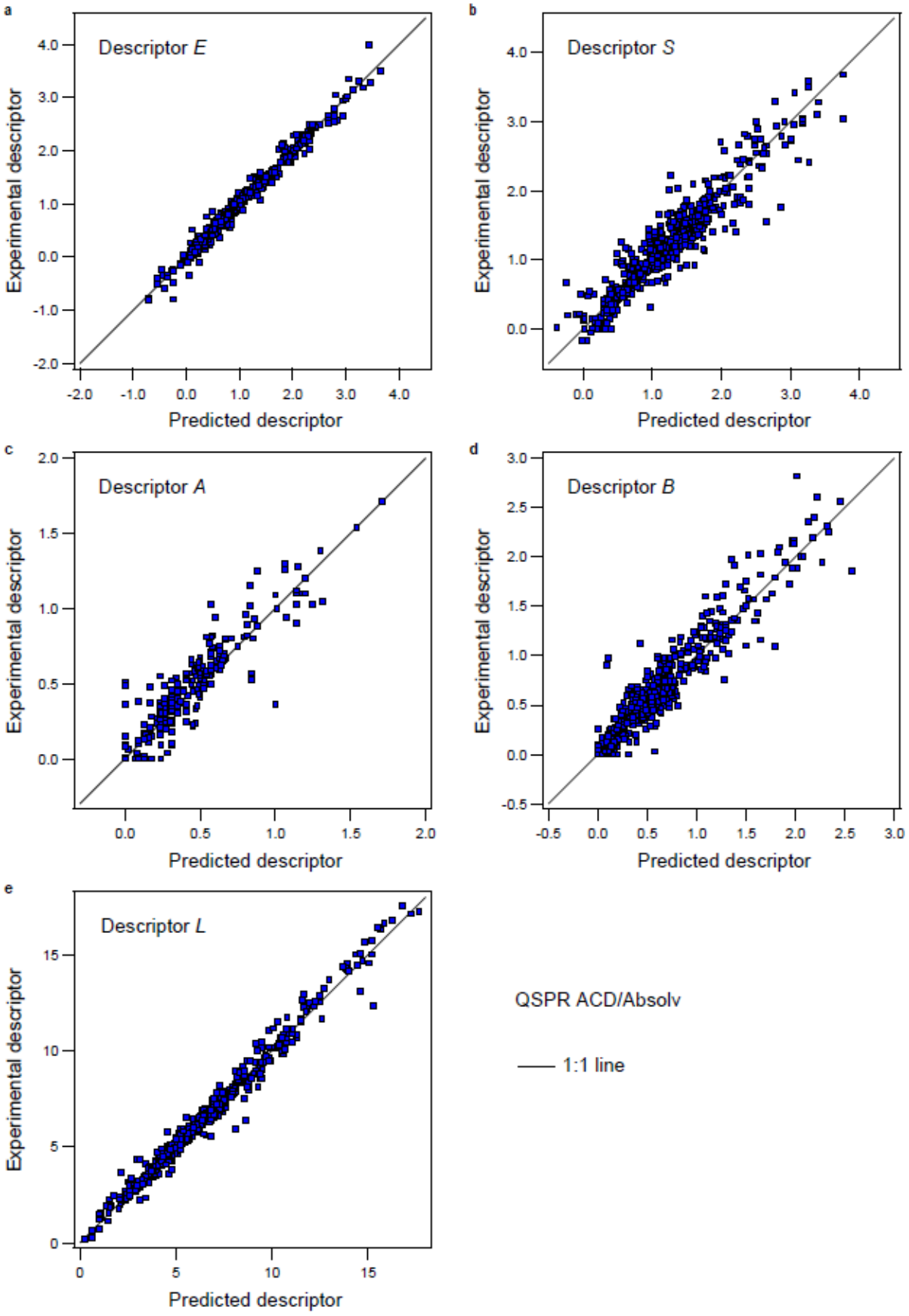
Supplementary Figure 8 Predictions of the solute descriptors *E* (**a**), *S* (**b**), *A* (**c**), *B* (**d**), and *L* (**e**) for the test set by the multitask model DNN$_{mono}$.



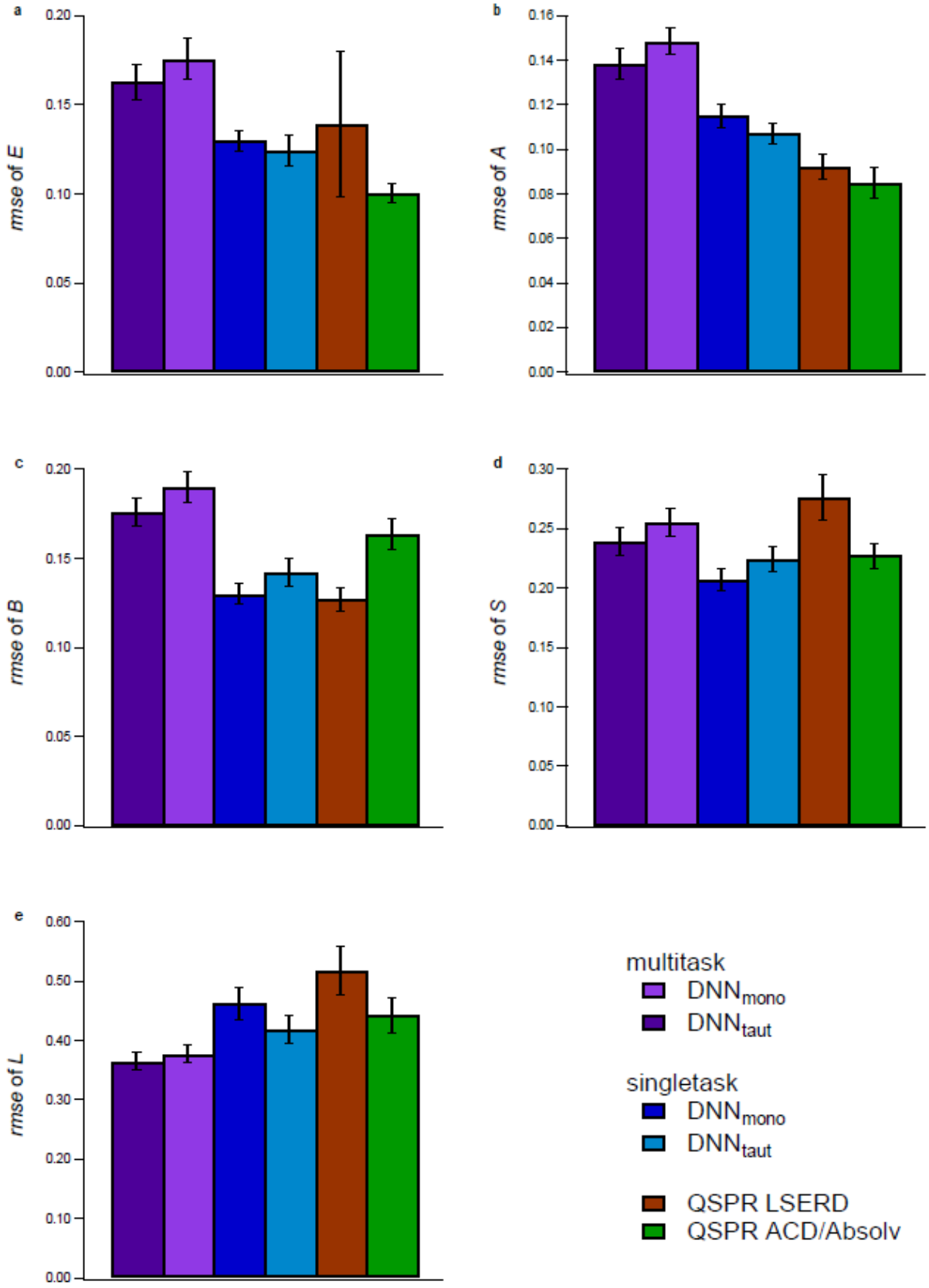
Supplementary Figure 9 Predictions of the solute descriptors *E* (**a**), *S* (**b**), *A* (**c**), *B* (**d**), and *L* (**e**) for the test set by the multitask model DNN_{taut}.



Supplementary Figure 10 Predictions of the solute descriptors *E* (**a**), *S* (**b**), *A* (**c**), *B* (**d**), and *L* (**e**) for the test set by the QSPR of LSERD.



Supplementary Figure 11 Predictions of the solute descriptors *E* (**a**), *S* (**b**), *A* (**c**), *B* (**d**), and *L* (**e**) for the test set by the QSPR of ACD/Absolv.

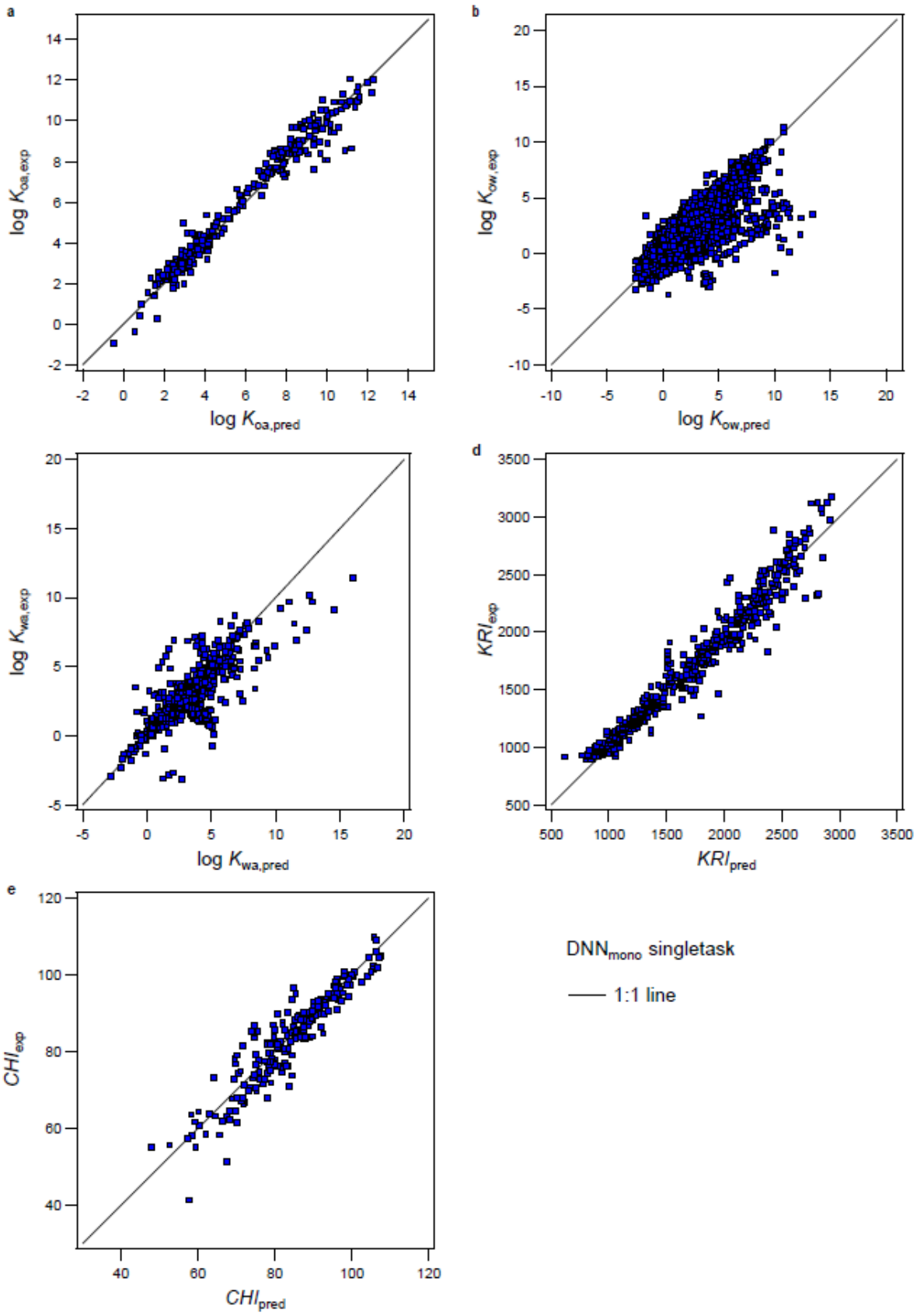


Supplementary Figure 12 Comparison of the *rmse*s for the solute descriptors *E* (**a**), *A* (**b**), *B* (**c**), *S* (**d**), and *L* (**e**) of the test set of the four different prediction tools DNN_{mono}, DNN_{taut}, QSPR from LSERD, and the QSPR ACD/Absolv.

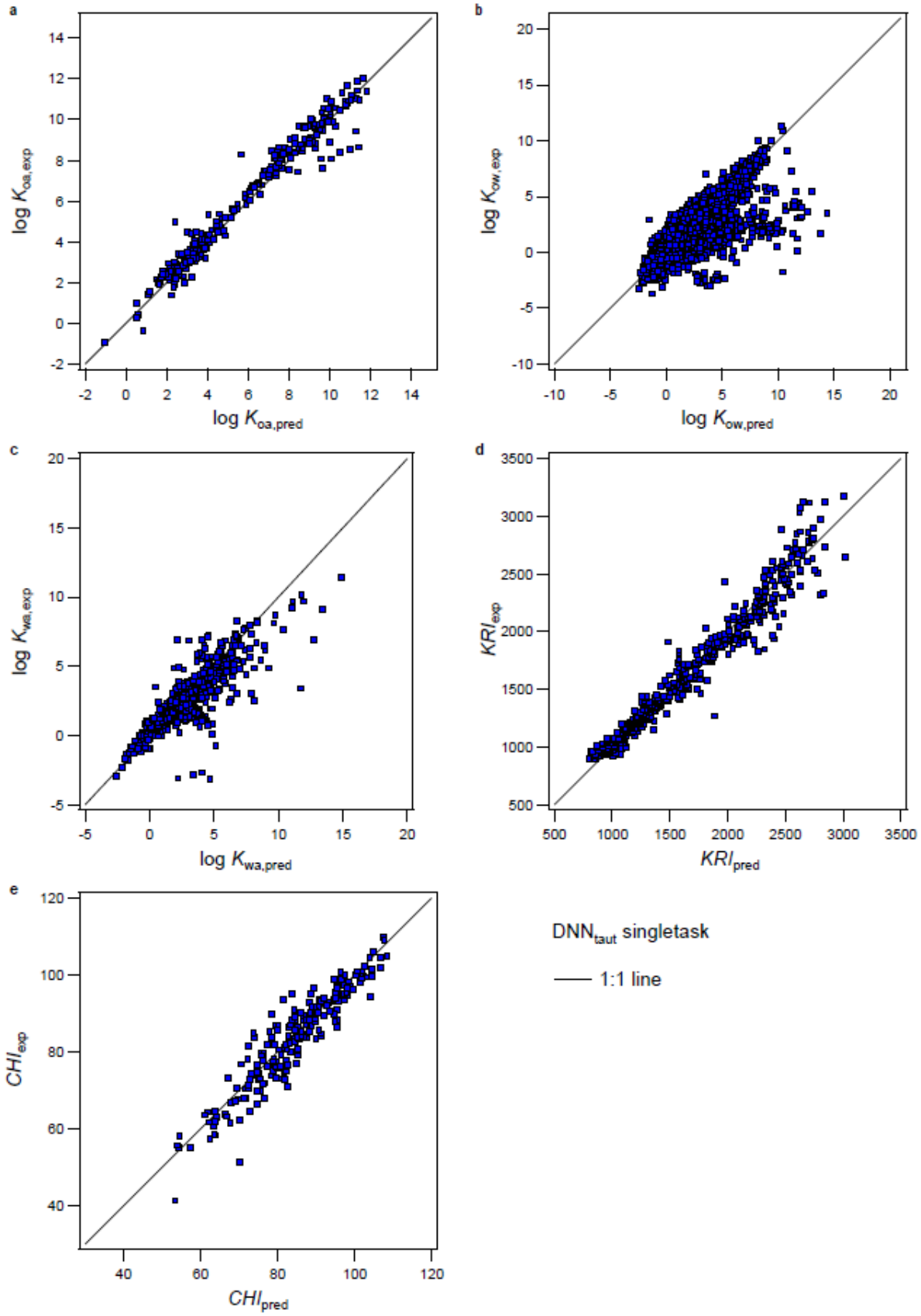
SI 4 Prediction performance of the four DNN models and the two QSPR models for various datasets

Supplementary Table 3 The root mean square error (*rmse*) and corresponding variance (*sdev*) are given for the predictions of partition coefficients and retention data based on LSER equations of the different DNNs compared to the QSPR model of LSERD and the QSPR model of ACD/Absolv. Note that the units of K_{ow}, K_{oa}, and K_{wa} are [$L_{water}/L_{octanol}$], [$L_{air}/L_{octanol}$], and [L_{air}/L_{water}], respectively. K_{wa} values were originally given as HL values (air-water, [m^3 atm/mole]) and converted according to Sander[2]. Note that the mean value and variance were estimated using bootstrapping according to Vorberg and Tetko[1].

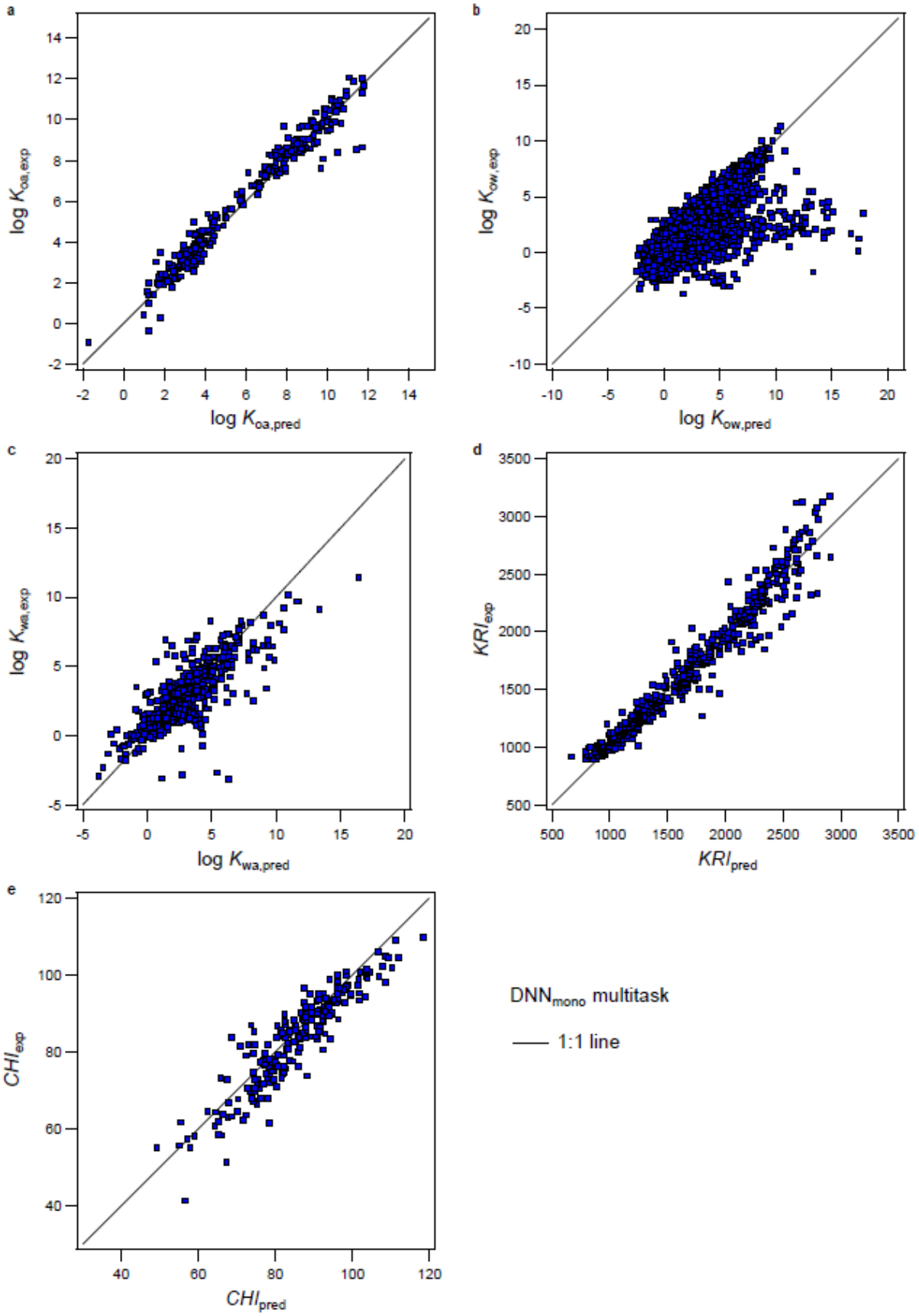
model	number of chemicals in the test set	singletask				multitask				LSERD		ACD/Absolv	
		DNN_{mono}		DNN_{taut}		DNN_{mono}		DNN_{taut}		QSPR			
dataset		*rmse*	*sdev*	*rmse*	*sdev*	*rmse*	*sdev*	*rmse*	*sdev*	*rmse*	*sdev*	*rmse*	*sdev*
log K_{ow} Eq. (1)	12010	1.01	0.02	1.04	0.02	1.36	0.03	1.41	0.04	0.91	0.01	0.87	0.01
log K_{ow} Eq. (3)	12010	0.86	0.01	0.89	0.01	1.20	0.02	1.25	0.03	1.11	0.01	0.90	0.01
log K_{oa} Eq. (1)	270	0.60	0.04	0.63	0.05	0.63	0.05	0.68	0.05	0.49	0.07	0.59	0.07
log K_{oa} Eq. (3)	270	0.59	0.04	0.61	0.05	0.62	0.05	0.68	0.05	0.50	0.08	0.59	0.07
log K_{wa} Eq. (1)	696	1.46	0.07	1.36	0.08	1.46	0.07	1.84	0.07	1.28	0.07	1.39	0.09
log K_{wa} Eq. (3)	696	1.36	0.05	1.24	0.07	1.37	0.07	1.76	0.07	1.20	0.07	1.34	0.12
KRI	454	127	7	129	7	131	7	134	8	104	10	120	7
CHI	204	4.58	0.29	4.49	0.29	5.42	0.29	5.14	0.28	5.52	0.55	3.19	0.14



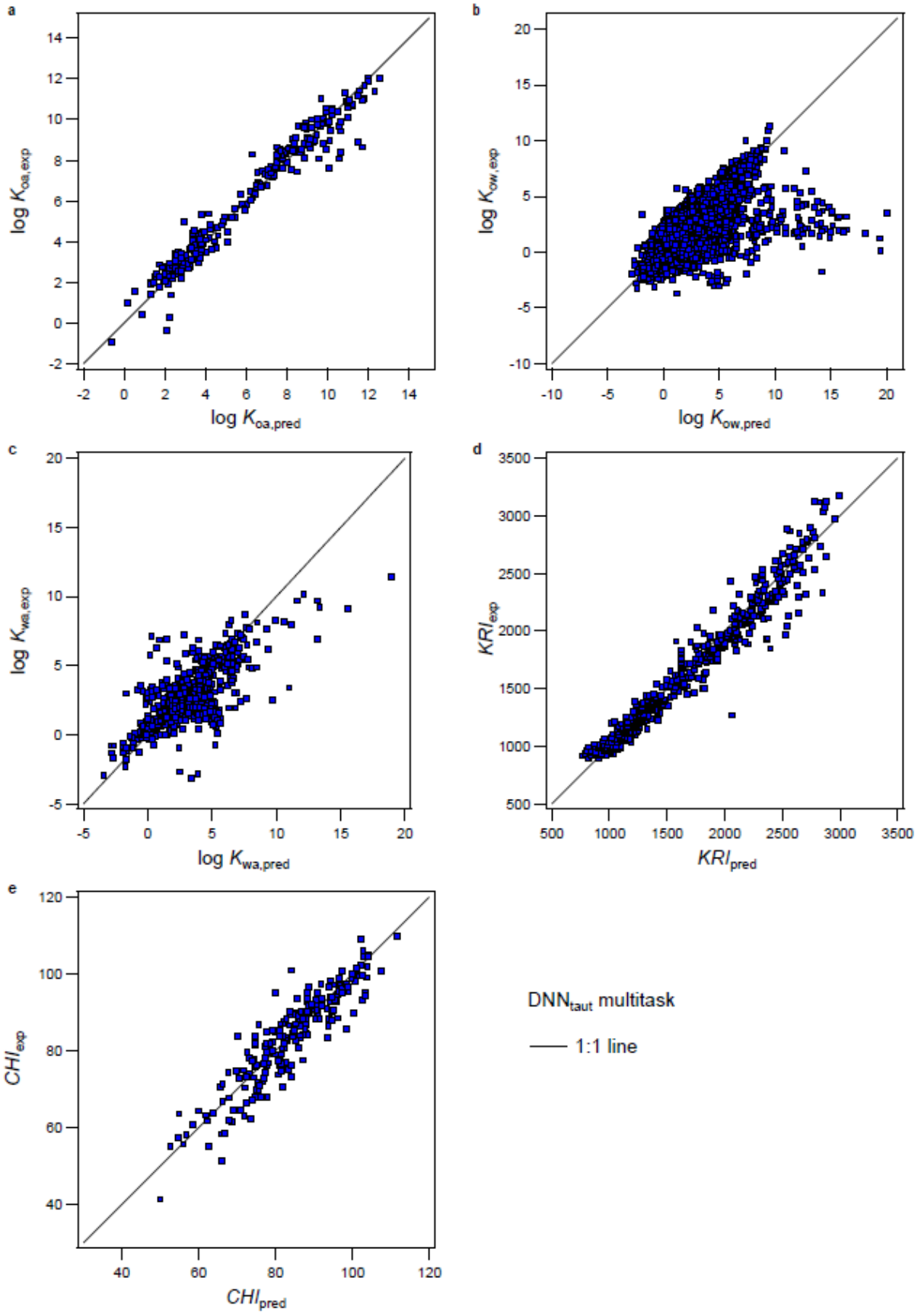
Supplementary Figure 13 Predictions of the partition coefficients log K_{oa} (**a**), log K_{ow} (**b**), log K_{wa} (**c**), and the retention indices *KRI* (GC) (**d**) and *CHI* (LC) (**e**) by the singletask model DNN_{mono}.



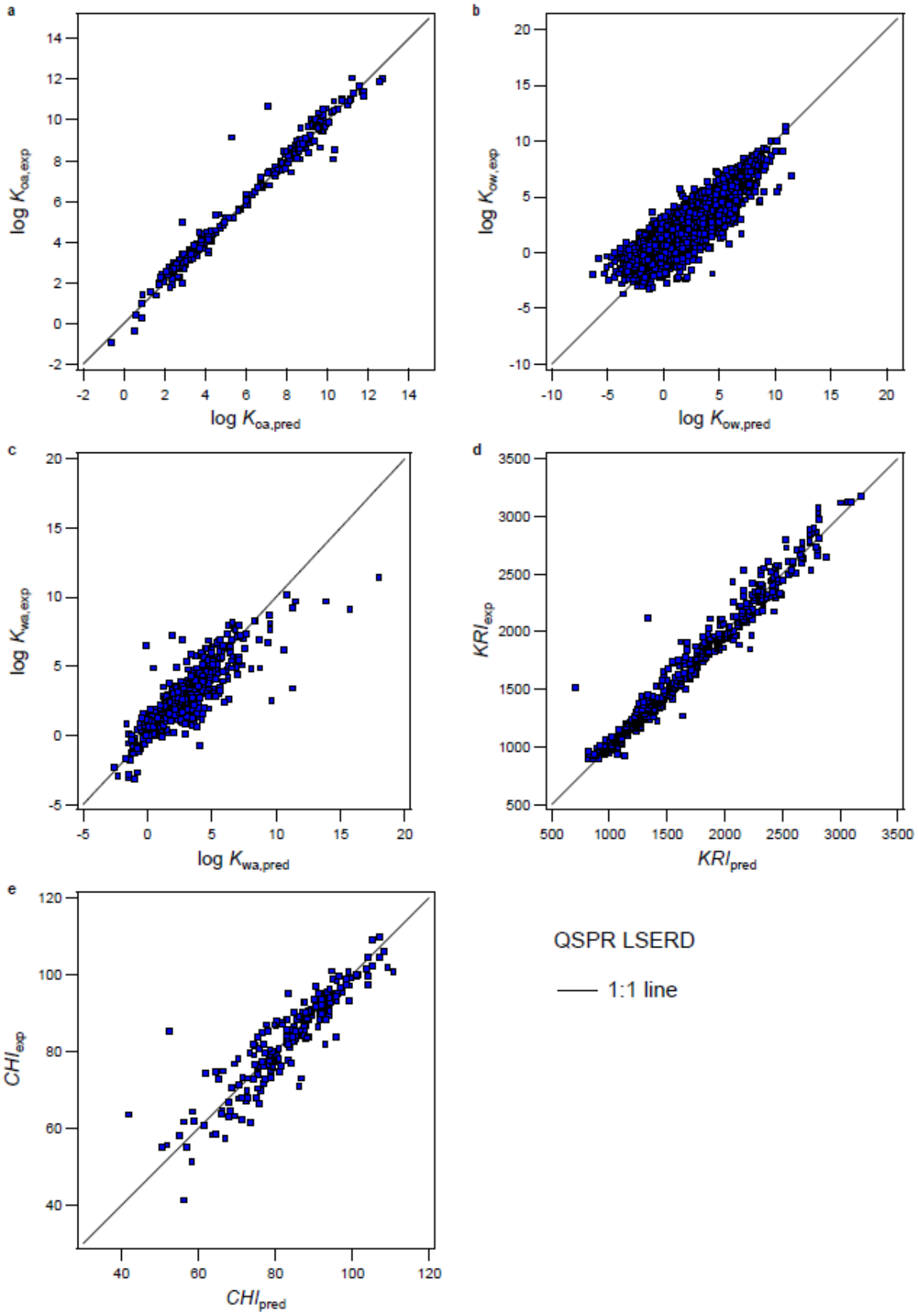
Supplementary Figure 14 Predictions of the partition coefficients log K_{oa} (**a**), log K_{ow} (**b**), log K_{wa} (**c**), and the retention indices *KRI* (GC) (**d**) and *CHI* (LC) (**e**) by the singletask model DNN_{taut}.



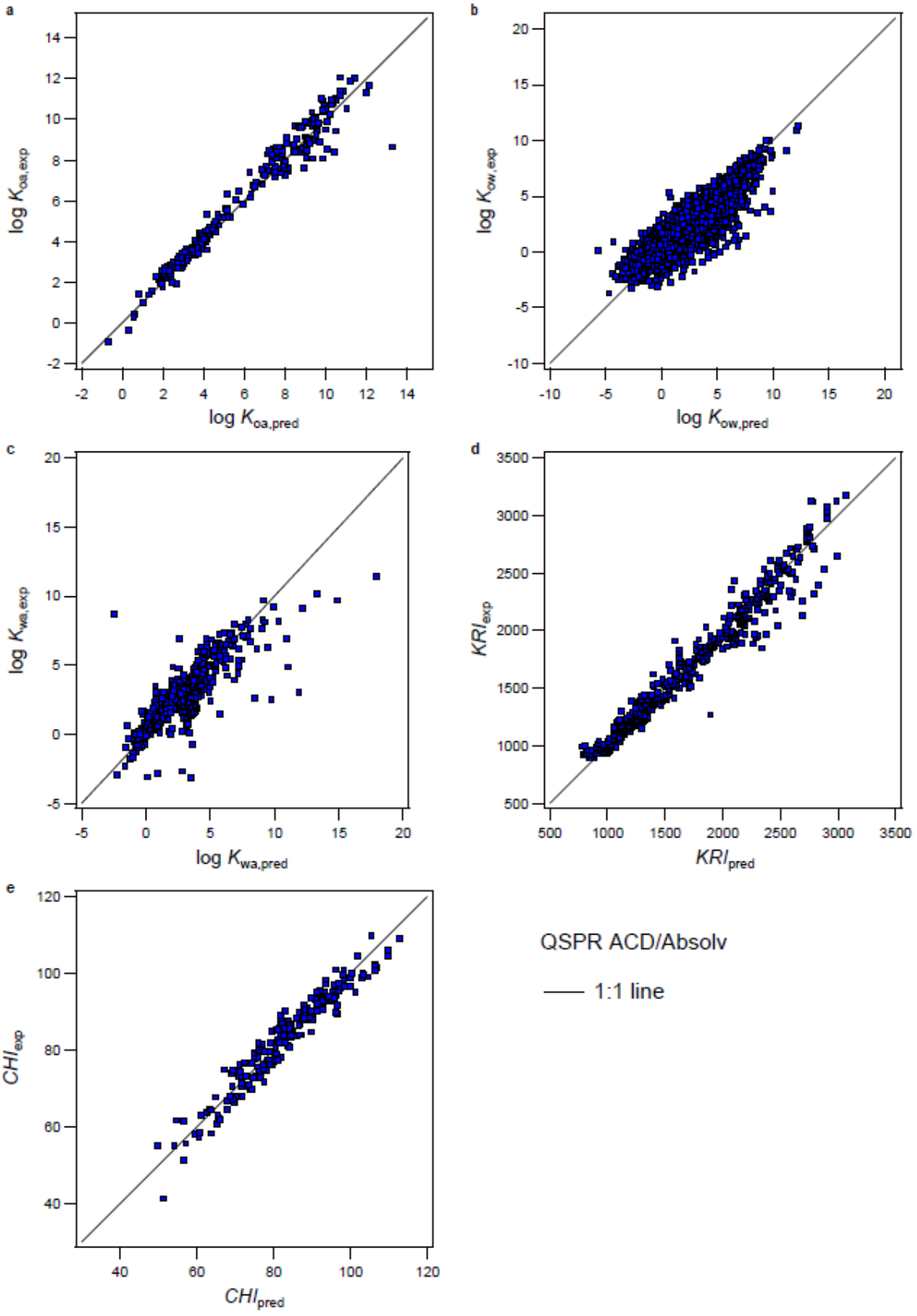
Supplementary Figure 15 Predictions of the partition coefficients log K_{oa} (**a**), log K_{ow} (**b**), log K_{wa} (**c**), and the retention indices *KRI* (GC) (**d**) and *CHI* (LC) (**e**) by the multitask model DNN_{mono}.



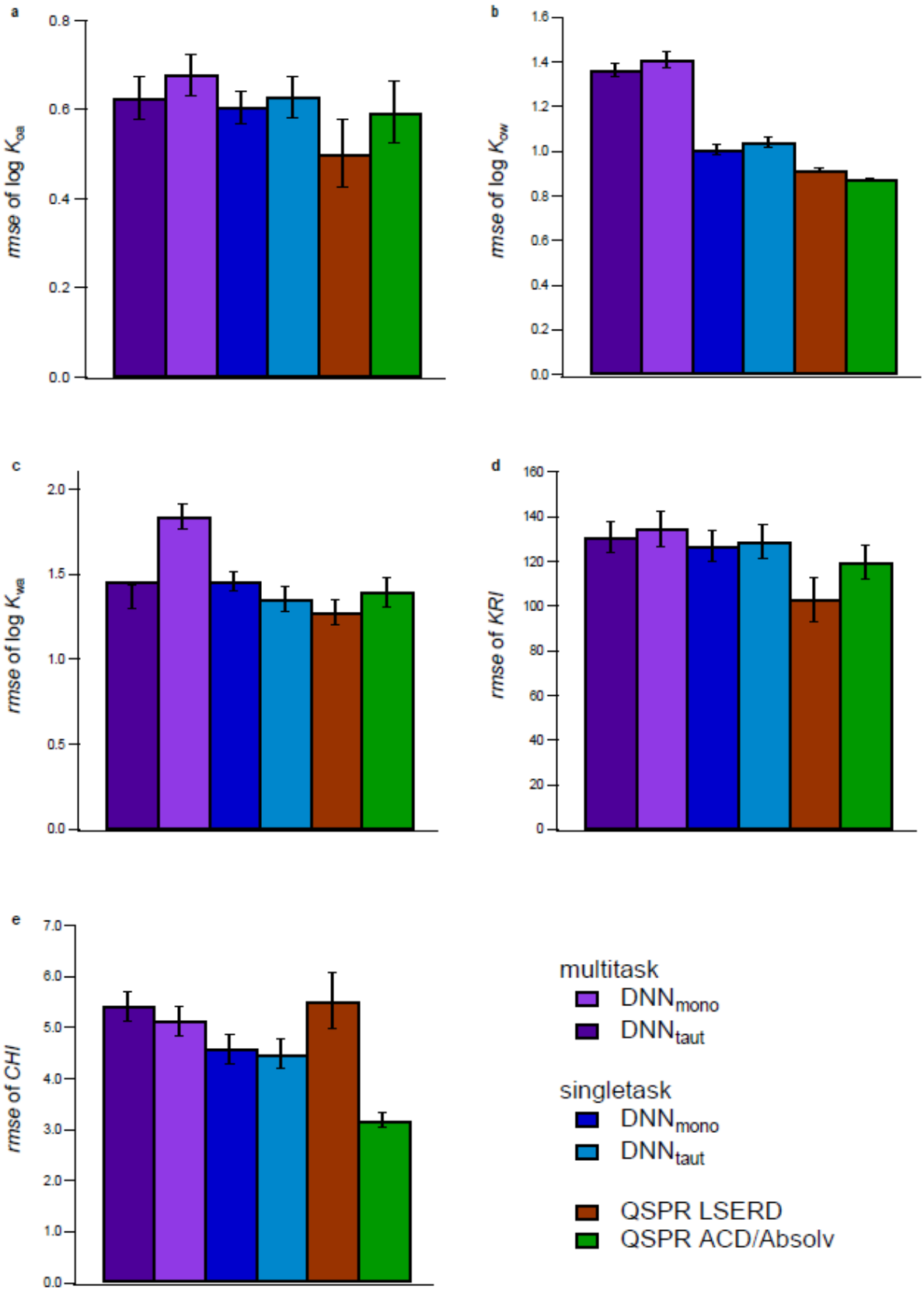
Supplementary Figure 16 Predictions of the partition coefficients log K_{oa} (**a**), log K_{ow} (**b**), log K_{wa} (**c**), and the retention indices *KRI* (GC) (**d**) and *CHI* (LC) (**e**) by the multitask model DNN_{taut}.



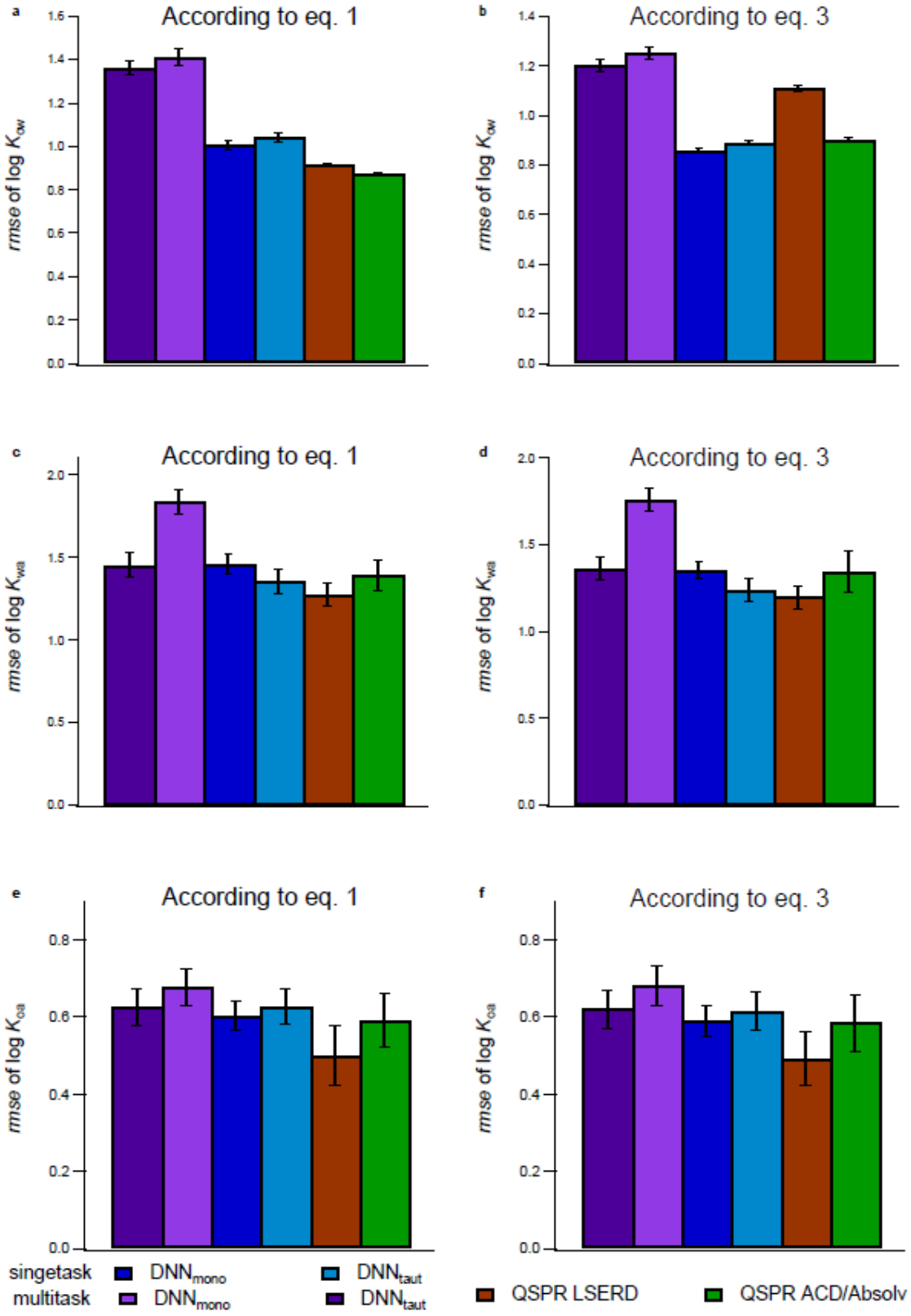
Supplementary Figure 17 Predictions of the partition coefficients log K_{oa} (**a**), log K_{ow} (**b**), log K_{wa} (**c**), and the retention indices *KRI* (GC) (**d**) and *CHI* (LC) (**e**) by the QSPR model of LSERD.



Supplementary Figure 18 Predictions of the partition coefficients log K_{oa} (**a**), log K_{ow} (**b**), log K_{wa} (**c**), and the retention indices *KRI* (GC) (**d**) and *CHI* (LC) (**e**) by the QSPR model ACD/Absolv.

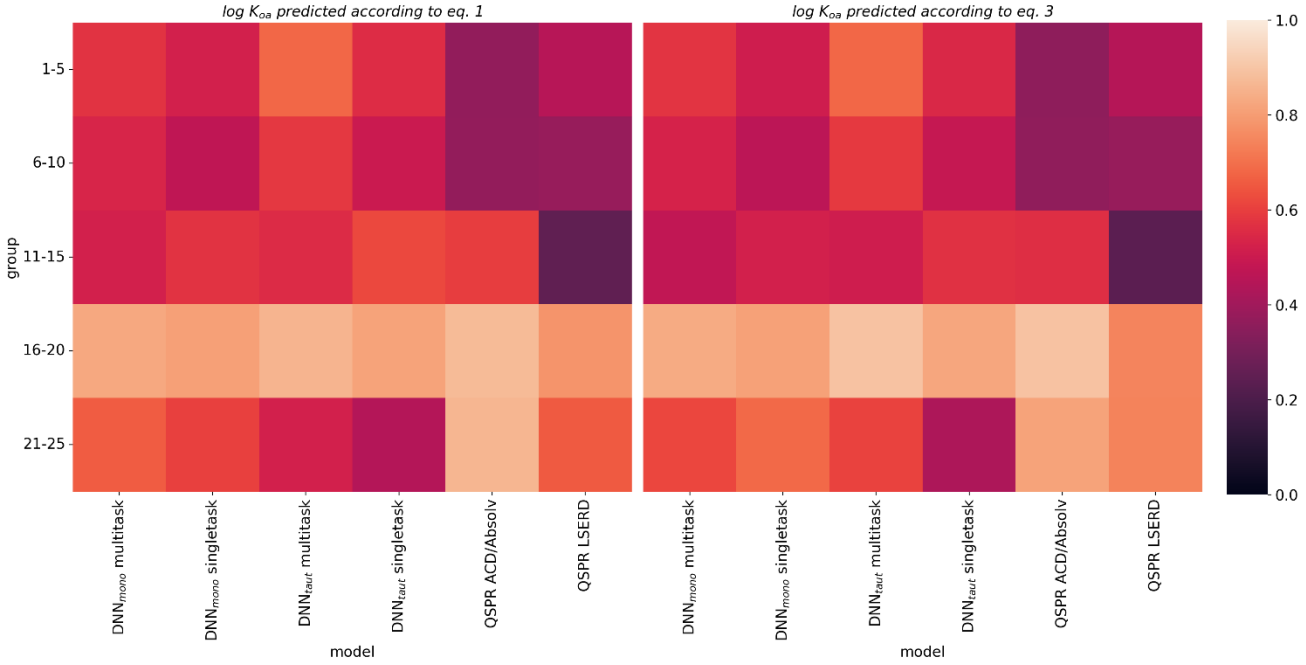


Supplementary Figure 19 Comparison of the *rmse*s for the predictions of the partition coefficients log K_{oa} (**a**), log K_{ow} (**b**), log K_{wa} (**c**), and the retention indices *KRI* (GC) (**d**) and *CHI* (LC) (**e**) by the four different prediction tools DNN_{mono}, DNN_{taut}, QSPR from LSERD, and the QSPR ACD/Absolv.

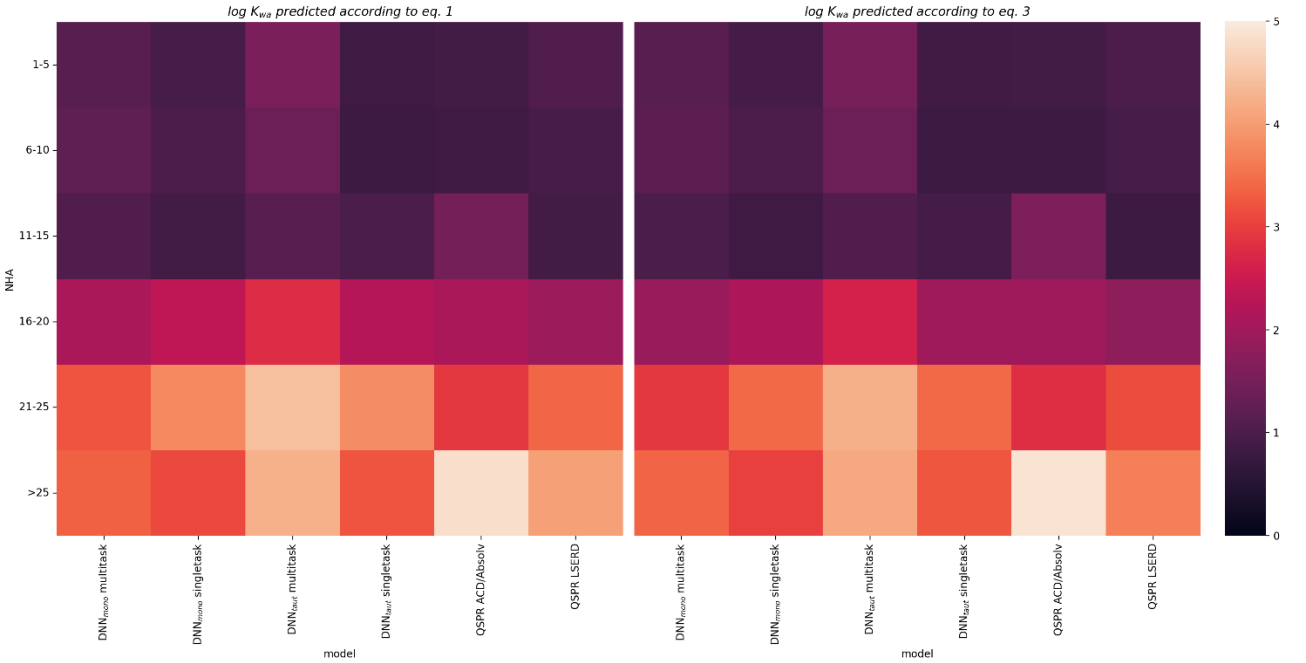


Supplementary Figure 20 Comparison of the *rmse*s for the predictions of the partition coefficients log K_{ow} (**a** and **b**), log K_{wa} (**c** and **d**), log K_{oa} (**e** and **f**) according to the different LSER equations by the four different prediction tools DNN_{mono}, DNN_{taut}, QSPR from LSERD, and the QSPR ACD/Absolv.

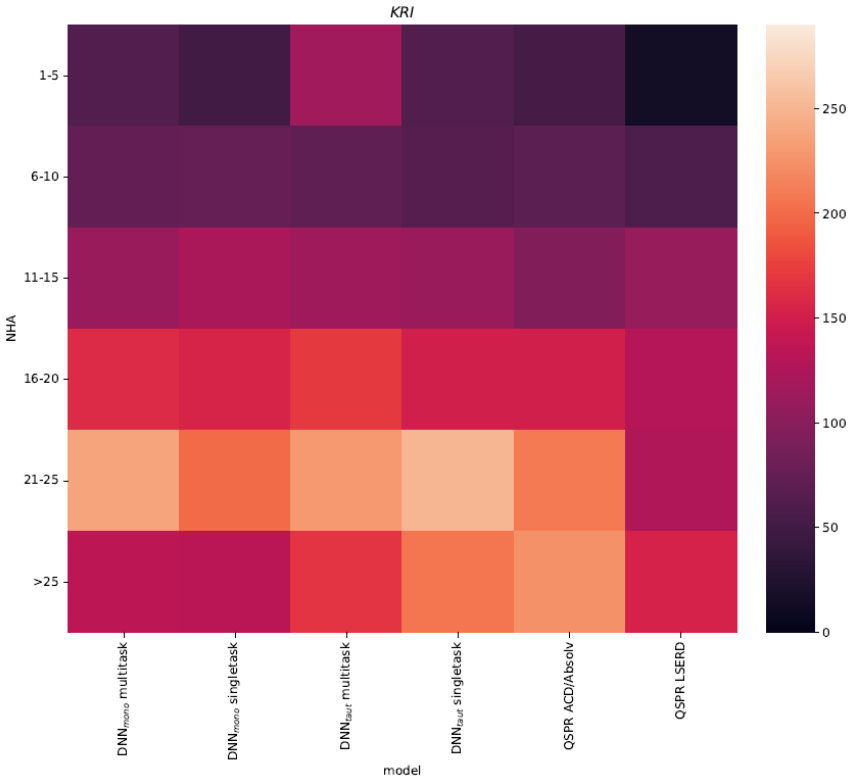
SI 5 *Rmse*s of the dataset predictions depending on the number of non-hydrogen atoms NHA



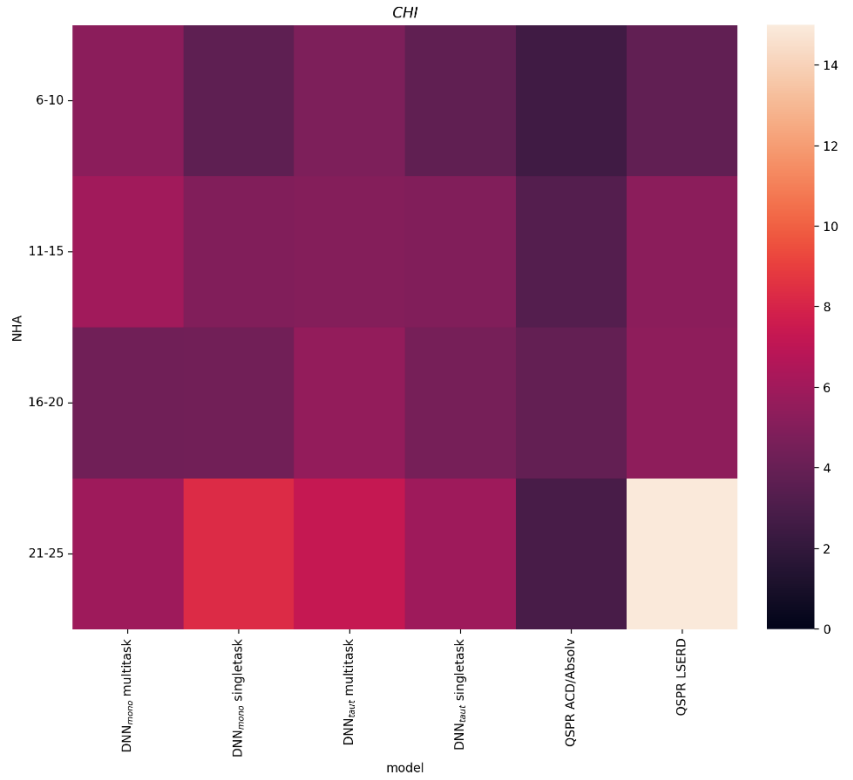
Supplementary Figure 21 *Rmse*s of the log K_{oa} predictions based on the different DNN approaches and QSPR depending on the number of non-hydrogen atoms NHA. The log K_{oa} is predicted based on two different LSER equations (1 – left heat map, 2 – right heat map).



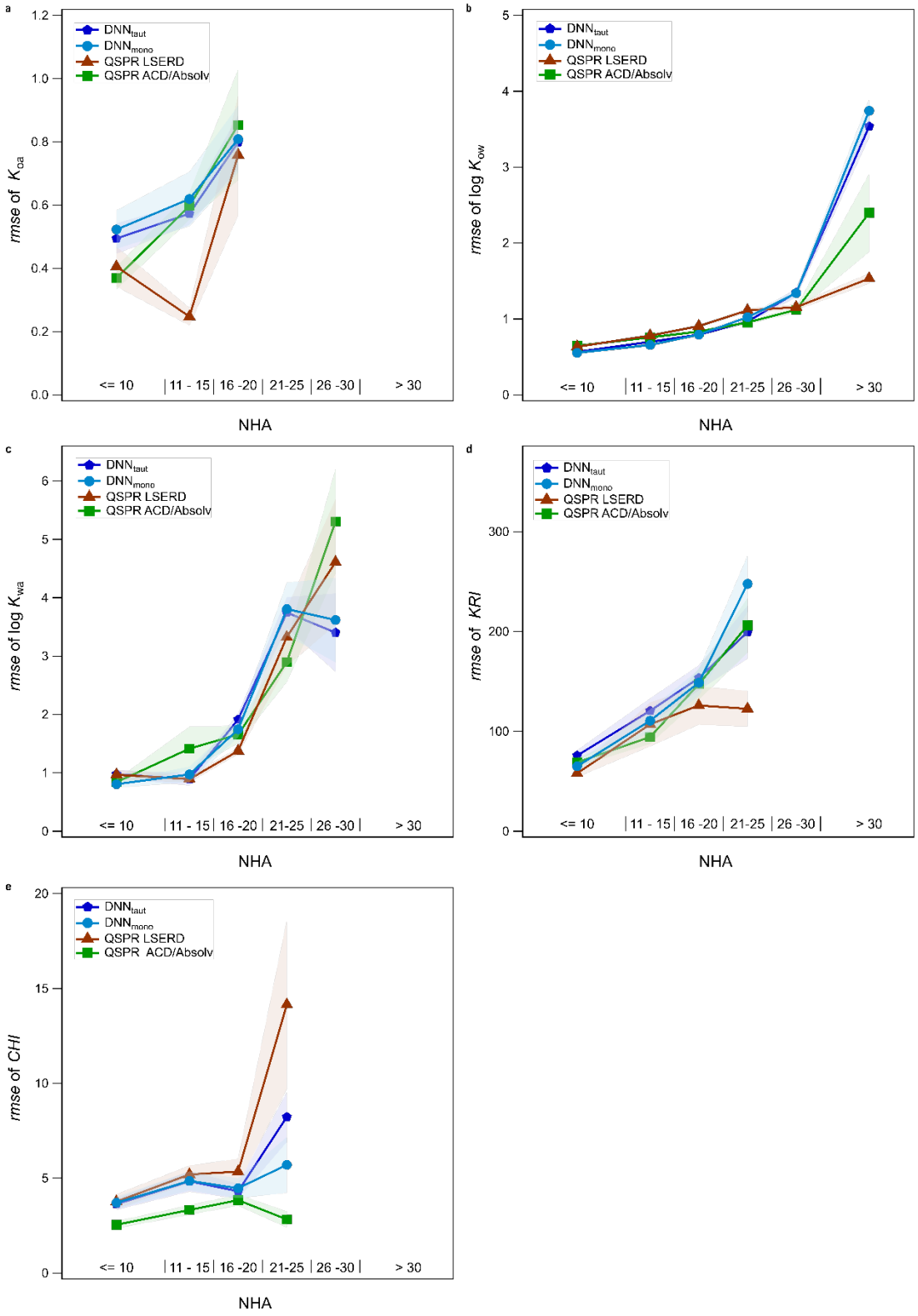
Supplementary Figure 22 *Rmse*s of the log K_{wa} predictions based on the different DNN approaches and QSPR depending on the number of non-hydrogen atoms NHA. The log K_{wa} is predicted based on two different LSER equations (1 – left heat map, 2 – right heat map).



Supplementary Figure 23 *Rmse*s of the *KRI* predictions based on the different DNN approaches and QSPR depending on the number of non-hydrogen atoms NHA.

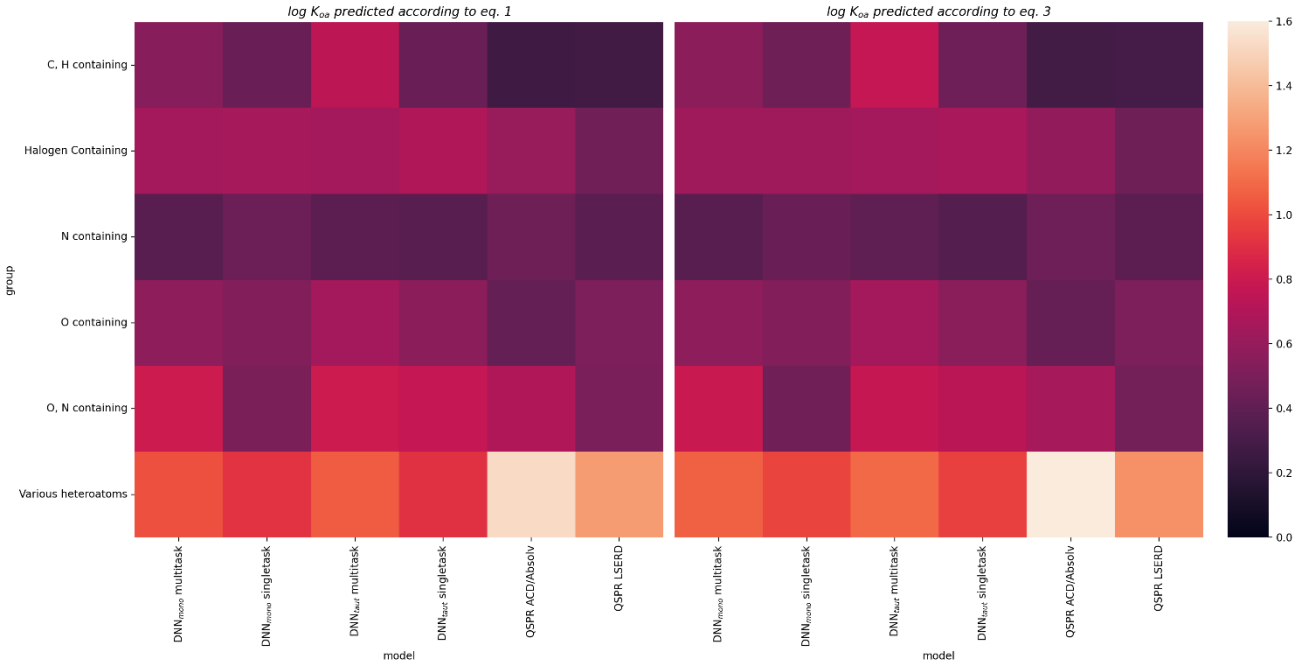


Supplementary Figure 24 *Rmse*s of the *CHI* predictions based on the different DNN approaches and QSPR depending on the number of non-hydrogen atoms NHA.

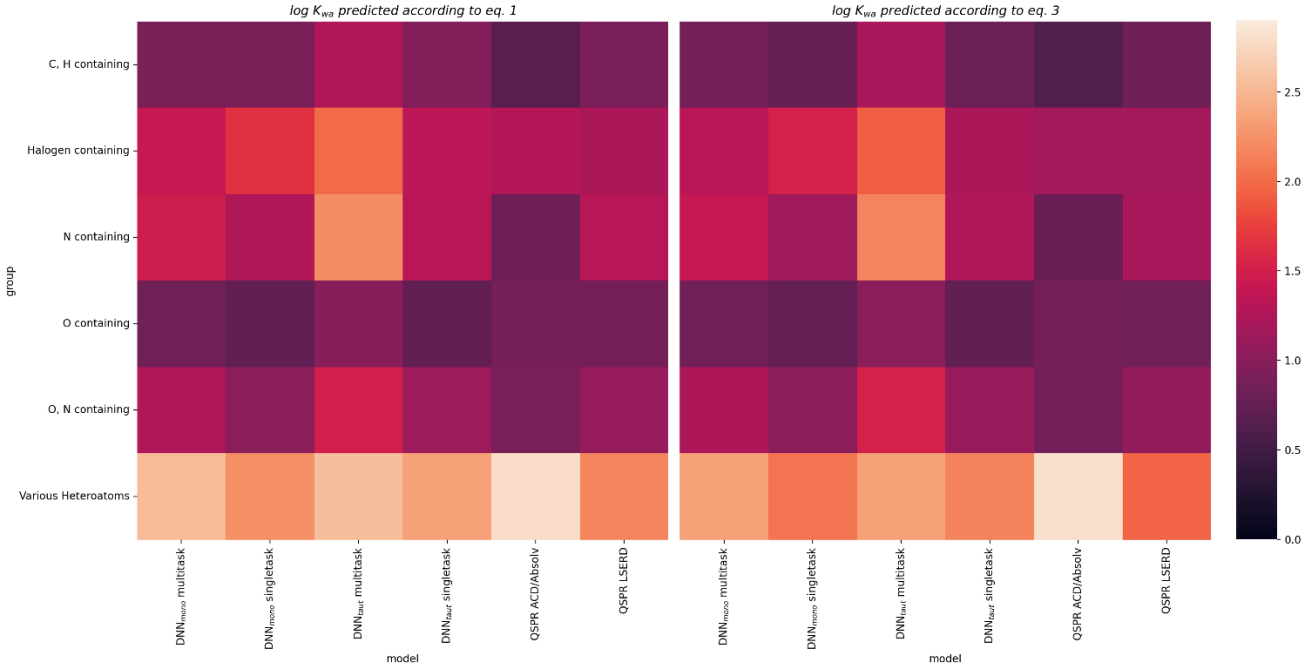


Supplementary Figure 25 *Rmse*s of the predictions of the partition coefficients log K_{oa} (**a**), log K_{ow} (**b**), log K_{wa} (**c**), and the retention indices *KRI* (GC) (**d**) and *CHI* (LC) (**e**) based on the two different DNN approaches and QSPR approaches depending on the number of non-hydrogen atoms NHA. The respective *sdev*s are given as shadow in the figure.

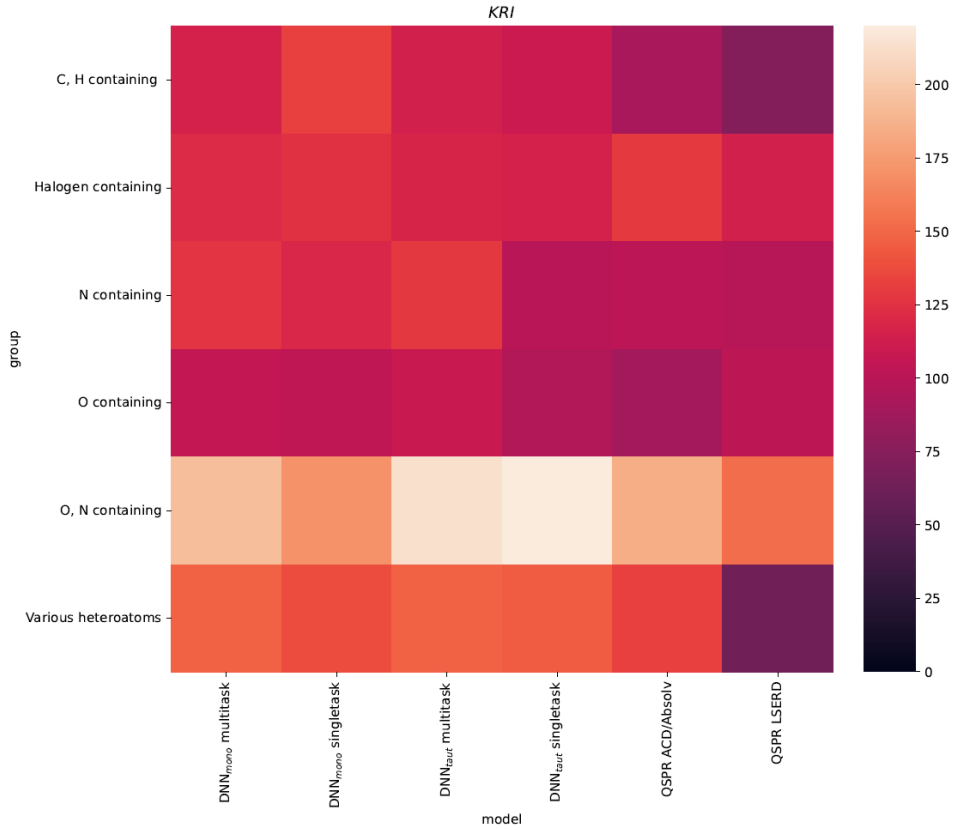
SI 6 *Rmse*s of the dataset predictions depending on the composition of the chemicals



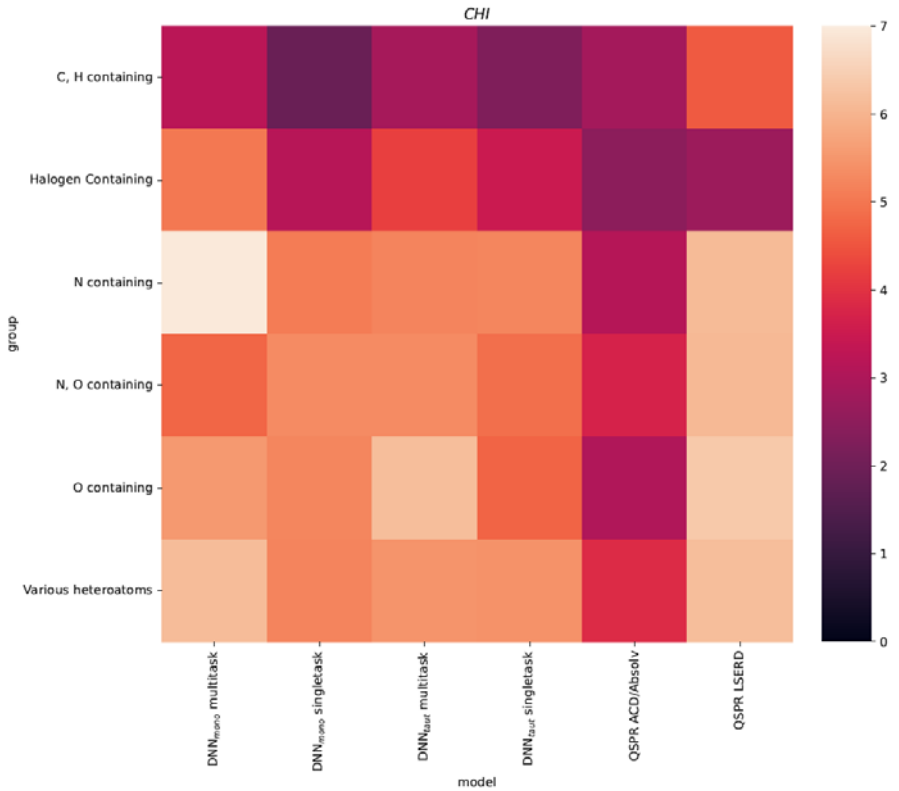
Supplementary Figure 26 *Rmse*s of the log K_{oa} predictions based on the different DNN approaches and QSPR depending on the composition of the chemicals. The log K_{oa} is predicted based on two different LSER equations (1 – left heat map, 2 – right heat map).



Supplementary Figure 27 *Rmse*s of the log K_{wa} predictions based on the different DNN approaches and QSPR depending on the composition of the chemicals. The log K_{wa} is predicted based on two different LSER equations (1 – left heat map, 2 – right heat map).

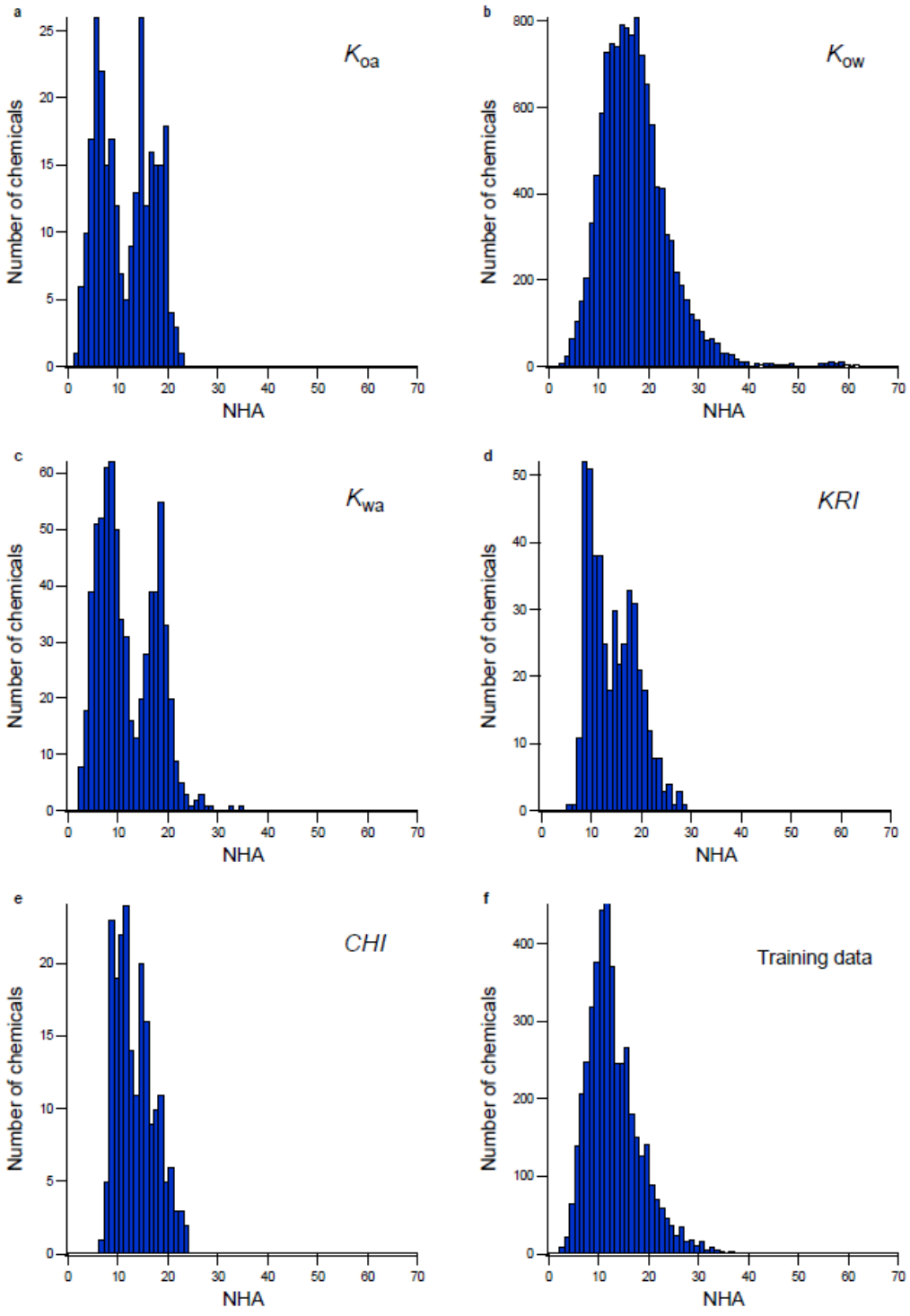


Supplementary Figure 28 *Rmse*s of the *KRI* predictions based on the different DNN approaches and QSPR depending on the composition of the chemicals.



Supplementary Figure 29 *Rmse*s of the *CHI* predictions based on the different DNN approaches and QSPR depending on the composition of the chemicals.

SI 7 Histograms of chemicals according to their NHA for the different datasets



Supplementary Figure 30 Histograms of chemicals according to their NHA for the datasets of the partition coefficients K_{oa} (**a**), K_{ow} (**b**), K_{wa} (**c**), and the retention indices *KRI* (GC) (**d**) and *CHI* (LC) (**e**) and our training set for DNN model development (**f**).

REFERENCES

1. Vorberg S, Tetko IV. Modeling the Biodegradability of Chemical Compounds Using the Online CHEmical Modeling Environment (OCHEM). *Mol Inform* **33**, 73-85 (2014).

2. Sander R. Compilation of Henry's law constants (version 4.0) for water as solvent. *Atmos Chem Phys* **15**, 4399-4981 (2015).