Supporting Information

VLA-SMILES: Variable-Length-Array SMILES Descriptors in Neural Network-based QSAR Modeling

Antonina L. Nazarova [1,*,†] and Aiichiro Nakano [2,*]

1 Present: Department of Quantitative & Computational Biology, Bridge Institute, USC Michelson Center for Convergent Bioscience, University of Southern California, Los Angeles, CA 90089, USA
2 Collaboratory of Advanced Computing and Simulations, Department of Computer Science, Department of Physics & Astronomy, Department of Quantitative & Computational Biology, University of Southern California, Los Angeles, CA 90089, USA
* Correspondence: nazarova@usc.edu (A.L.N.); anakano@usc.edu (A.N.)
† Past: Loker Hydrocarbon Research Institute, Bridge Institute, Chemistry Department, University of Southern California, Los Angeles, CA 90089, USA.

Table S1. Internal parameters of designed neural network-based QSAR models.

MLP internal parameters	ATransformedBP-based MLP	iRPROP- based-MLP	Adam-based MLP	Adam-based DNN (AutoEncoder MLP)
Input format	Variable-Length-Array-based SMILES $R_1, R_2, R_4, R_6, R_8, R_{12}, R_{16}$ (Dataset#1) R_1, R_2, R_4, R_8 (Dataset#2)			
Testing/Training set	50/950 (Dataset#1) 70/1300 (Dataset#2)			
Input nodes	d_1=1872/k, with k=1,2,4,6,8,12,16 (Dataset#1) d_2=1192/k, with k=1,2,4,8 (Dataset#2)			
Hidden nodes	d_1=1872/k, with k=1,2,4,6,8,12,16 (Dataset#1) d_2=1192/k, with k=1,2,4,8 (Dataset#2)			
Hidden layers	1	1, 2	1, 2	3
Nof epochs	2000 epochs (k = 6,8,12,16); 1000 epochs (k=4); 200 epochs (k=1,2)			same, but 100 epochs for k=1,2

For each hidden layer, the set of neurons has the same size as the length of the implemented array-featured SMILES representation. The number of the hidden and input layers' neurons was equal to the maximum of the input length of the original database divided by k.

Table S2. Training and prediction results in terms of RMSE for MLPs (single hidden layer) with VLA-SMILES representations, Kennard-Stone-based rational splitting algorithm and ATransformedBP learning, Dataset#1.

ATransformedBP	k	$\gamma_{(opt)}$	RMSE (training set)	RMSE (testing set)	# of epoch for min. RMSE (test)	N of epochs
Tanh(Y)	1	1	0.02	0.84	43	200
	2	2	0.02	0.80	14	200
	4	3	0.004	0.84	59	1000
	6	4	0.0003	0.93	1842	2000
	8	4	0.009	0.90	35	2000
	12	1	0.21	0.96	1159	2000
	16	3	0.05	0.95	375	2000
Sigmoid(Y)	1	2	0.06	0.81	35	200
	2	3	0.07	0.80	143	300
	4	4	0.08	0.85	145	500
	6	4	0.06	0.96	234	1000
	8	2	0.17	0.89	665	1700
	12	3	0.15	0.98	713	2000
	16	3	0.28	0.91	1528	2000
ReLU(Y)	1	1	0.02	0.84	43	200
	2	1	0.04	0.82	14	150
	4	4	0.02	0.84	48	400
	6	4	0.006	0.93	299	500
	8	2	0.04	0.93	201	2000
	12	1	0.08	1.02	219	1300
	16	3	0.04	0.90	143	2000

Table S3. Training and prediction results in terms of RMSE for MLPs (single hidden layer) with VLA-SMILES representations, the Kennard-Stone-based rational splitting procedure and iRPROP$^-$ learning, Dataset#1.

iRPROP-	k	RMSE (training set)	RMSE (testing set)	# of epoch for min. RMSE (test)	N of epochs
	1	0.09	0.788	23	150
	2	0.06	0.796	43	200
	4	0.05	0.848	134	1000
ReLU(Y)	6	0.004	0.874	79	2000
	8	0.03	0.887	46	2000
	12	0.15	0.905	196	2000
	16	0.11	0.875	319	2000
	1	0.09	0.77	70	150
	2	0.13	0.85	102	200
	4	0.08	0.92	490	1000
Tanh(Y)	6	0.04	0.94	297	2000
	8	0.06	0.92	609	2000
	12	0.16	1.01	300	2000
	16	0.15	0.93	866	2000
	1	0.06	0.77	44	150
	2	0.09	0.77	118	200
	4	0.003	0.87	99	1000
Sigmoid(Y)	6	0.00005	0.84	123	2000
	8	0.0008	0.94	137	2000
	12	0.005	0.95	115	2000
	16	0.005	0.89	113	2000

Table S4. Training and prediction results in terms of RMSE for MLPs (single hidden layer) with VLA-SMILES representations, the Ranking by Activity-based rational splitting algorithm and iRPROP^{-} learning, Dataset#1.

iRPROP-	k	RMSE (training set)	RMSE (testing set)	# of epoch for min. RMSE (test)	N of epochs
ReLU(Y)	1	0.10	0.92	27	300
	2	0.10	0.91	20	300
	4	0.14	0.90	163	1000
	6	0.10	0.87	36	2000
	8	0.14	0.86	22	2000
	12	0.19	0.89	115	2000
	16	0.20	0.87	29	2000
Tanh(Y)	1	0.12	0.93	8	300
	2	0.15	0.97	20	300
	4	0.12	0.97	6	1000
	6	0.13	0.97	4	2000
	8	0.13	0.94	262	2000
	12	0.33	0.98	7	2000
	16	0.22	1.03	181	2000
Sigmoid(Y)	1	0.04	0.87	49	300
	2	0.08	0.95	44	300
	4	0.06	0.89	202	1000
	6	0.02	0.87	66	2000
	8	0.10	1.03	56	2000
	12	0.15	0.88	96	2000
	16	0.17	0.94	68	2000

Table S5. Training and prediction results in terms of RMSE for MLPs (single hidden layer) with VLA-SMILES representations, Kennard-Stone-based rational splitting procedure and iRPROP^{-} learning, Dataset#2.

Adam	k	RMSE (training set)	RMSE (testing set)	# of epoch for min. RMSE (test)	N of epochs
ReLU(Y)	1	0.13	0.87	83	1000
	2	0.11	0.85	63	1000
	4	0.24	0.79	464	1000
	8	0.03	0.92	120	2000
Sigmoid (Y)	1	0.001	0.84	67	1000
	2	0.001	0.78	90	1000
	4	0.002	0.83	149	1000
	8	0.006	0.92	220	2000
Tanh(Y)	1	0.001	0.88	118	1000
	2	0.001	0.88	137	1000
	4	0.006	0.87	247	1000
	8	0.030	0.95	383	2000

Table S6. Training and prediction results in terms of RMSE for MLPs (single hidden layer) with VLA-SMILES representations, Kennard-Stone-based rational splitting algorithm and Adam optimizer, Dataset#1.

Adam	k	RMSE (training set)	RMSE (testing set)	# of epoch for min. RMSE (test)	N of epochs
	1	0.05	0.86	3	100
	2	0.08	0.82	6	150
	4	0.16	0.82	77	200
ReLU(Y)	6	0.04	0.95	285	300
	8	0.03	0.92	73	500
	12	0.03	1.00	58	800
	16	0.02	0.90	190	2000
	1	0.10	0.85	6	150
	2	0.04	0.79	7	475
	4	0.04	0.83	82	475
Tanh(Y)	6	0.04	0.92	75	475
	8	0.01	0.91	126	1900
	12	0.02	0.99	236	2000
	16	0.02	0.92	514	2000
	1	0.05	0.82	43	100
	2	0.05	0.79	66	200
	4	0.08	0.84	242	500
Sigmoid(Y)	6	0.01	0.94	816	1000
	8	0.02	0.93	423	2000
	12	0.08	0.99	730	2000
	16	0.36	0.93	1269	2000

Table S7. Training and prediction results in terms of RMSE for MLPs (single hidden layer) with VLA-SMILES representations, Ranking by Activity-based rational splitting algorithm and Adam optimizer, Dataset#1.

Adam	k	RMSE (training set)	RMSE (testing set)	# of epoch for min. RMSE (test)	N of epochs
	1	0.09	0.98	32	100
	2	0.08	1.12	196	200
	4	0.06	1.11	18	1000
ReLU(Y)	6	0.03	1.15	22	2000
	8	0.12	1.26	906	2000
	12	0.16	1.12	66	2000
	16	0.18	1.00	109	2000
	1	0.13	0.93	132	150
	2	0.09	1.15	127	200
	4	0.08	1.19	161	1000
Tanh(Y)	6	0.09	1.19	128	2000
	8	0.13	1.35	28	2000
	12	0.18	1.14	264	2000
	16	0.30	1.29	1	2000
	1	0.07	1.01	38	150
	2	0.11	1.18	193	200
	4	0.10	1.14	11	1000
Sigmoid(Y)	6	0.10	1.21	7	2000
	8	0.16	1.29	2	2000
	12	0.22	1.11	750	2000
	16	0.23	1.26	1	2000

Table S8. Training and prediction results in terms of RMSE for MLPs (two hidden layers) with VLA-SMILES representations, Kennard-Stone-based rational splitting algorithm and iRPROP^{-} learning, Dataset#1.

iRPROP^{-}	**k**	**RMSE (training set)**	**RMSE (testing set)**	**# of epoch for min. RMSE (test)**	**N of epochs**
Tanh(Y)	1	0.05	0.90	32	200
	2	0.05	0.79	82	200
	4	0.003	0.83	134	1000
	6	0.0006	0.89	243	2000
	8	0.001	0.92	128	2000
	12	0.005	0.97	180	2000
	16	0.010	0.95	357	2000
Sigmoid(Y)	1	0.05	0.81	84	200
	2	0.09	0.87	48	200
	4	0.03	0.85	151	1000
	6	0.003	0.89	107	2000
	8	0.003	0.85	115	2000
	12	0.02	0.98	87	2000
	16	0.01	0.90	119	2000

Table S9. Training and prediction results in terms of RMSE for MLPs (two hidden layers) with VLA-SMILES representations, Kennard-Stone-based rational splitting algorithm and Adam learning, Dataset#1.

Adam	**k**	**RMSE (training set)**	**RMSE (testing set)**	**# of epoch for min. RMSE (test)**	**N of epochs**
Tanh(Y)	1	0.15	0.90	4	100
	2	0.03	0.84	19	200
	4	0.01	0.86	167	1000
	6	0.01	0.94	356	2000
	8	0.05	0.95	906	2000
	12	0.14	1.00	655	2000
	16	0.18	0.91	1541	2000
Sigmoid(Y)	1	0.06	0.81	83	150
	2	0.07	0.80	406	650
	4	0.11	0.86	639	1350
	6	0.41	1.01	473	550
	8	0.17	0.94	707	1000
	12	0.22	0.98	458	2000
	16	0.25	0.90	1682	2000

Table S10. Training and prediction results in terms of RMSE for MLPs with Autoencoder (three hidden layers), VLA-SMILES representations, Kennard-Stone-based rational splitting procedure and iRPROP$^-$ learning, Dataset#1.

iRPROP$^-$	**k**	**RMSE (training set)**	**RMSE (testing set)**	**# of epoch for min. RMSE (test)**	**N of epochs**
Sigmoid(Y)	1	0.05	0.85	23	100
	2	0.06	0.84	74	200
	4	0.02	0.88	263	1100
	6	0.05	0.94	255	1100
	8	0.07	0.99	867	2000
	12	0.32	1.03	444	1300
	16	0.18	0.92	1823	2000

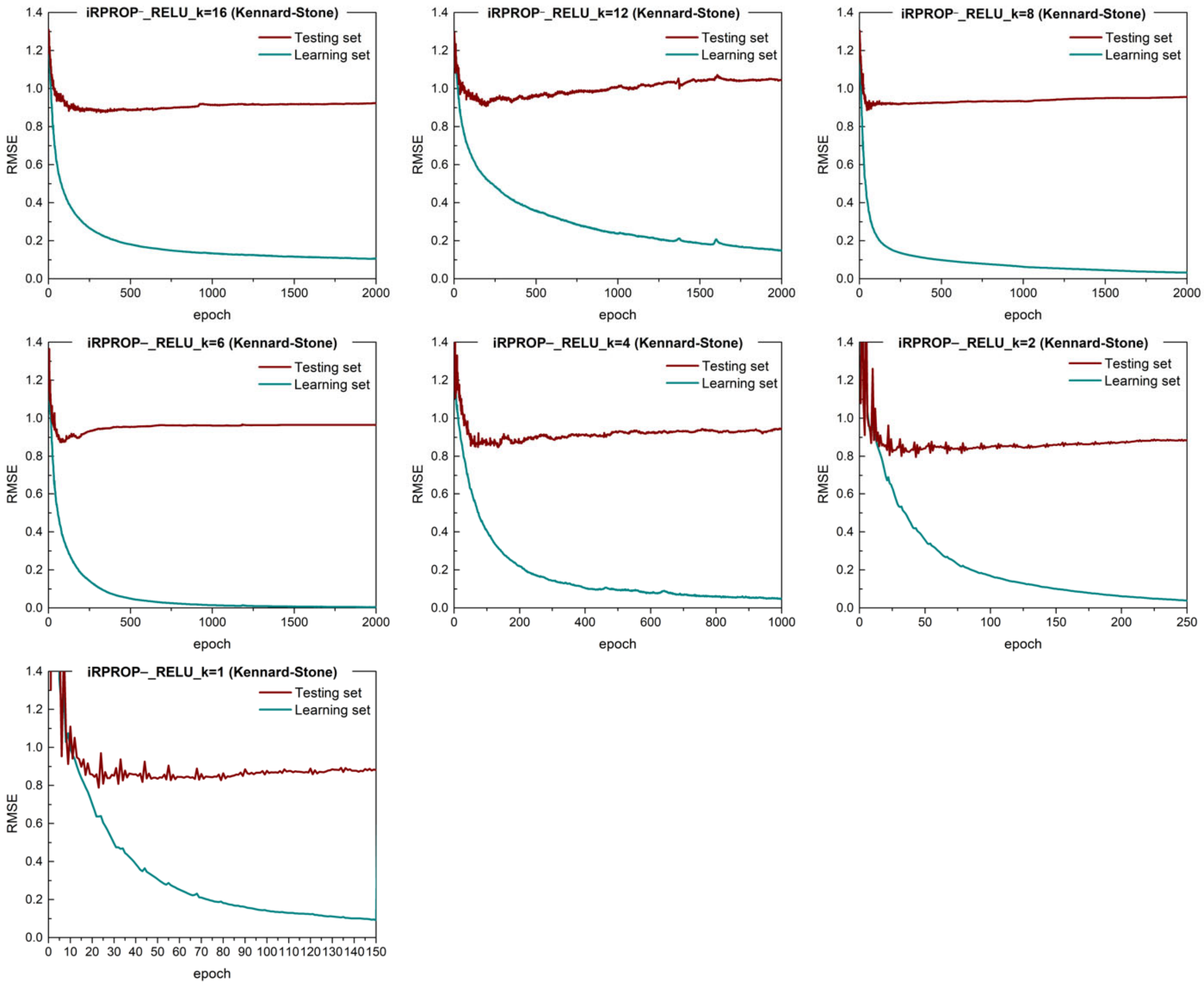

Figure S1. RMSE as a function of epoch for learning (training) and testing sets for an MLP with one hidden layer with iRPROP$^-$ learning algorithm and variable-length-array SMILES representation (ReLU activation, Kennard-Stone-based rational splitting algorithm), Dataset#1.

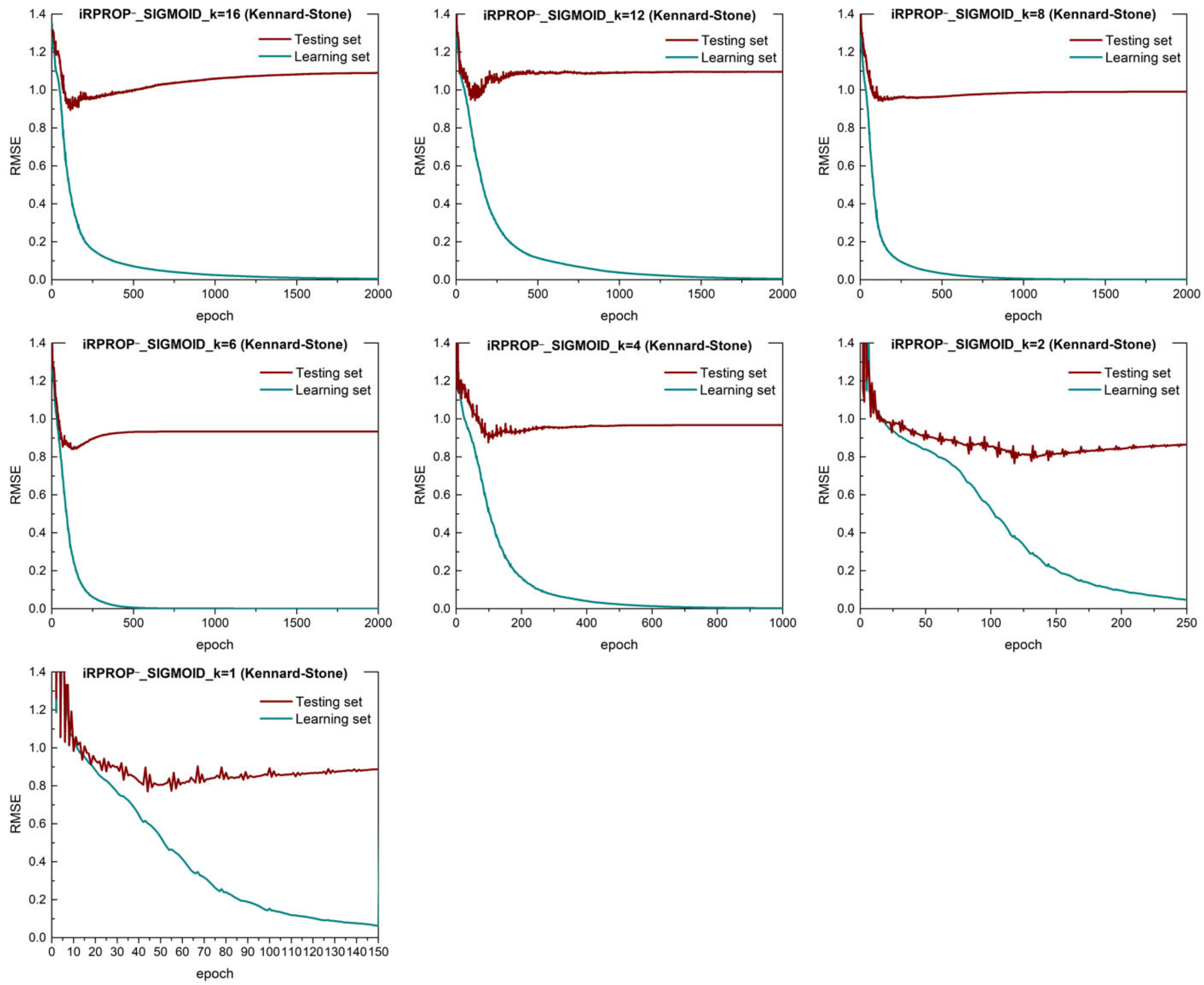

Figure S2. RMSE as a function of epoch for learning (training) and testing sets for an MLP with one hidden layer with iRPROP$^-$ learning algorithm and variable-length-array SMILES representation (Sigmoid activation, Kennard-Stone-based rational splitting algorithm), Dataset#1.

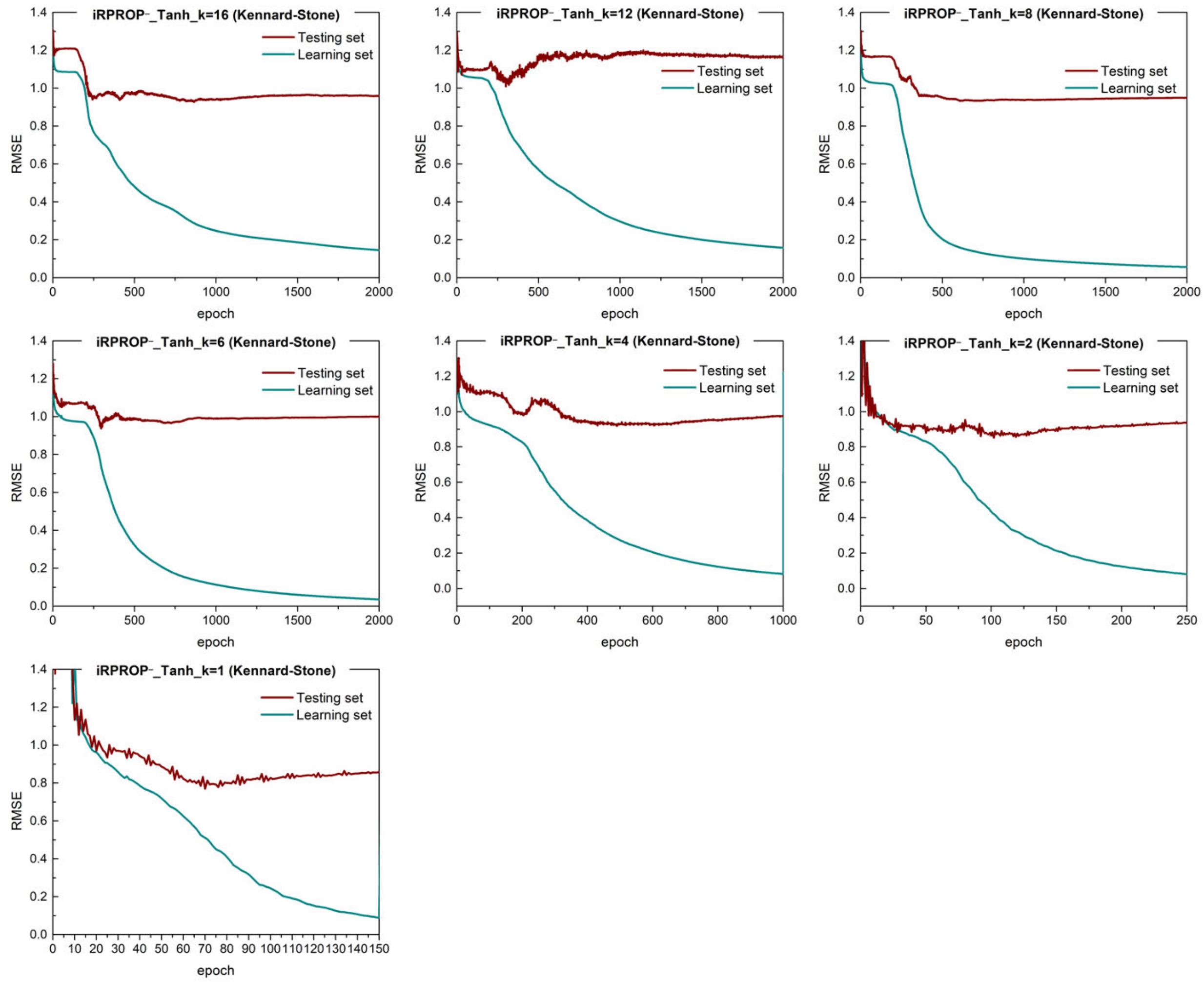

Figure S3. RMSE as a function of epoch for learning (training) and testing sets for an MLP with one hidden layer with iRPROP$^-$ learning algorithm and variable-length-array SMILES representation (Tanh activation, Kennard-Stone-based rational splitting algorithm), Dataset#1.

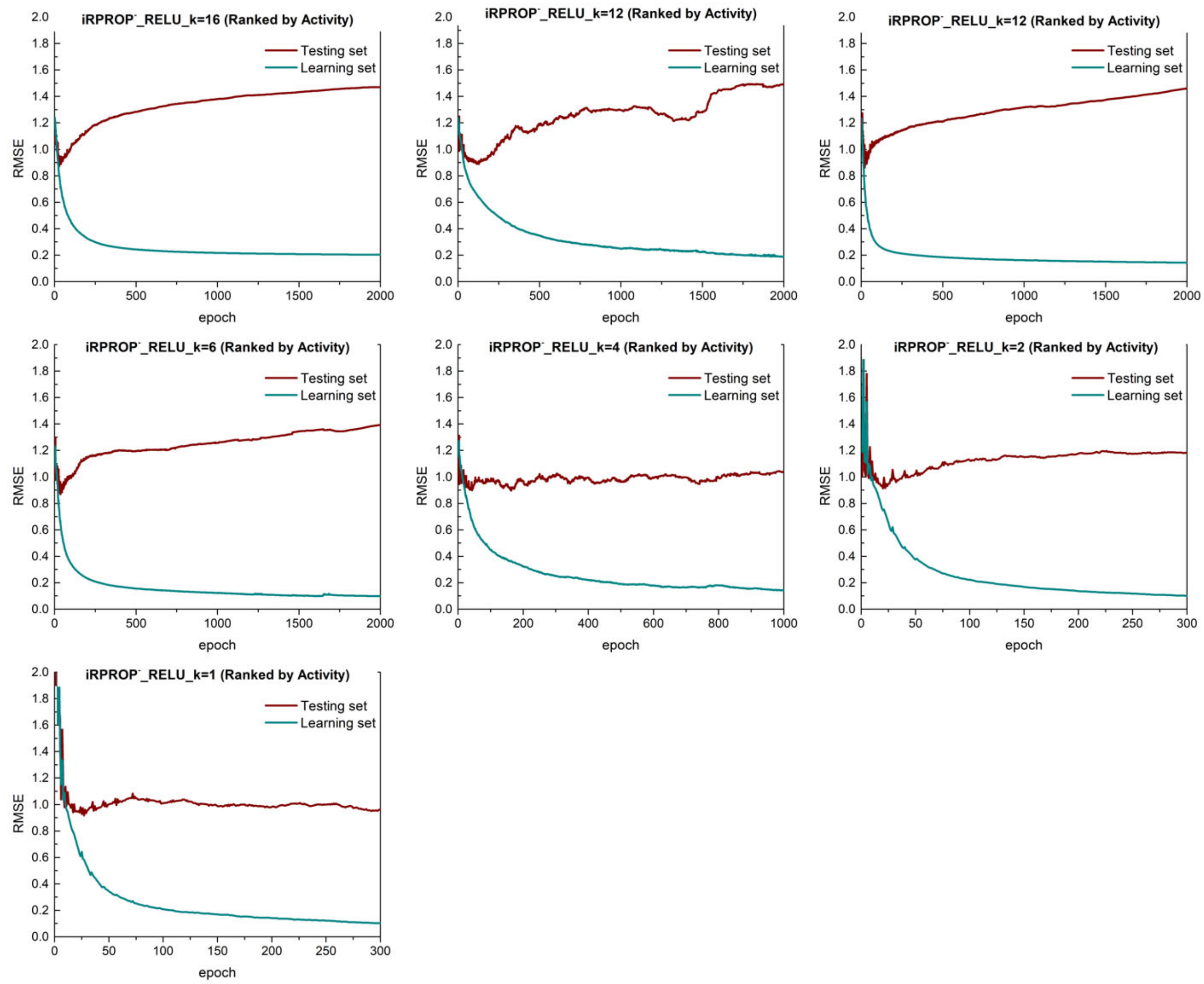

Figure S4. RMSE as a function of epoch for learning (training) and testing sets for an MLP with one hidden layer with iRPROP$^-$ learning algorithm and variable-length-array SMILES representation (ReLU activation, Ranking by Activity-based rational splitting algorithm), Dataset#1.

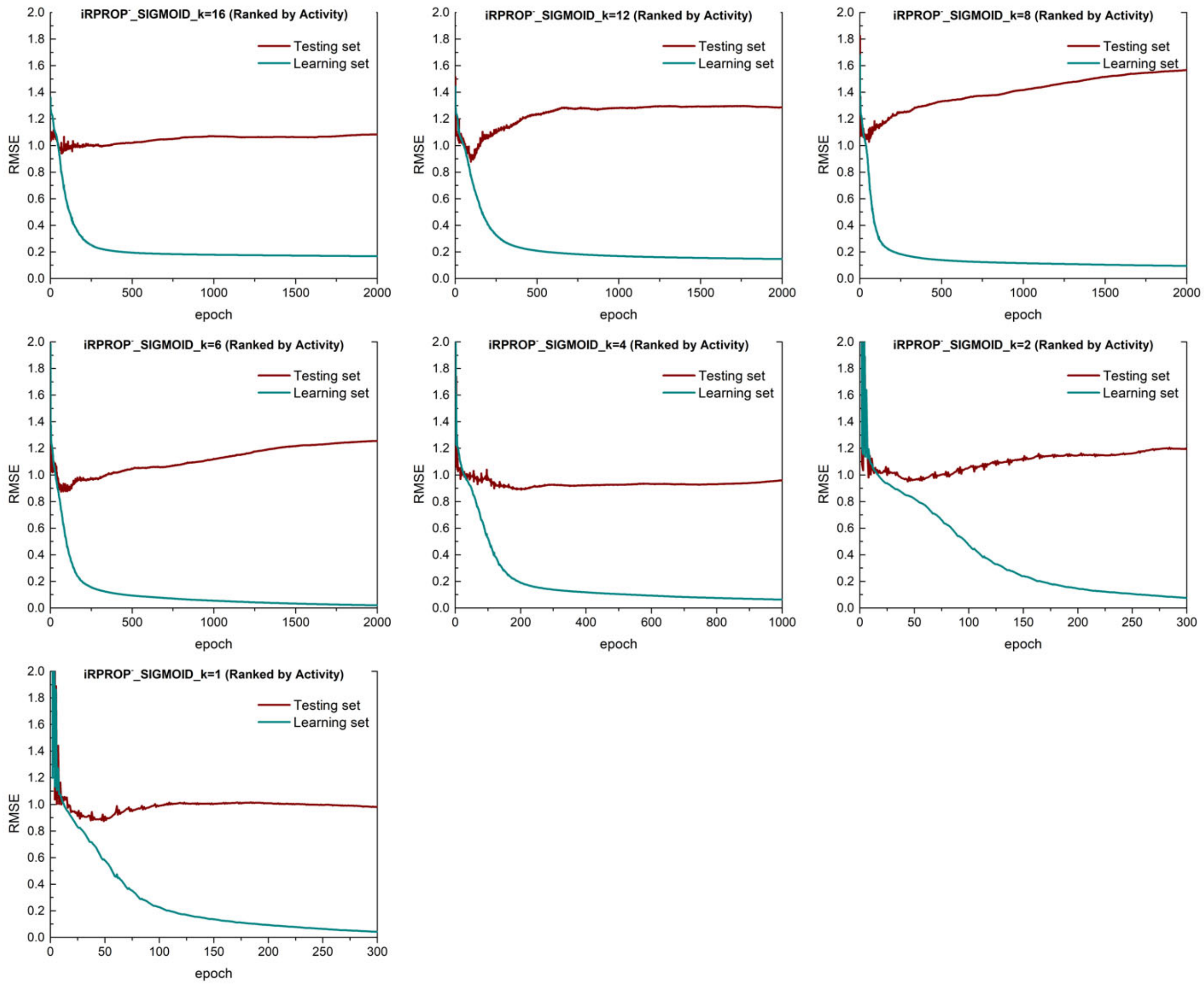

Figure S5. RMSE as a function of epoch for learning (training) and testing sets for an MLP with one hidden layer with iRPROP$^-$ learning algorithm and variable-length-array SMILES representation (Sigmoid activation, Ranking by Activity-based rational splitting algorithm), Dataset#1.

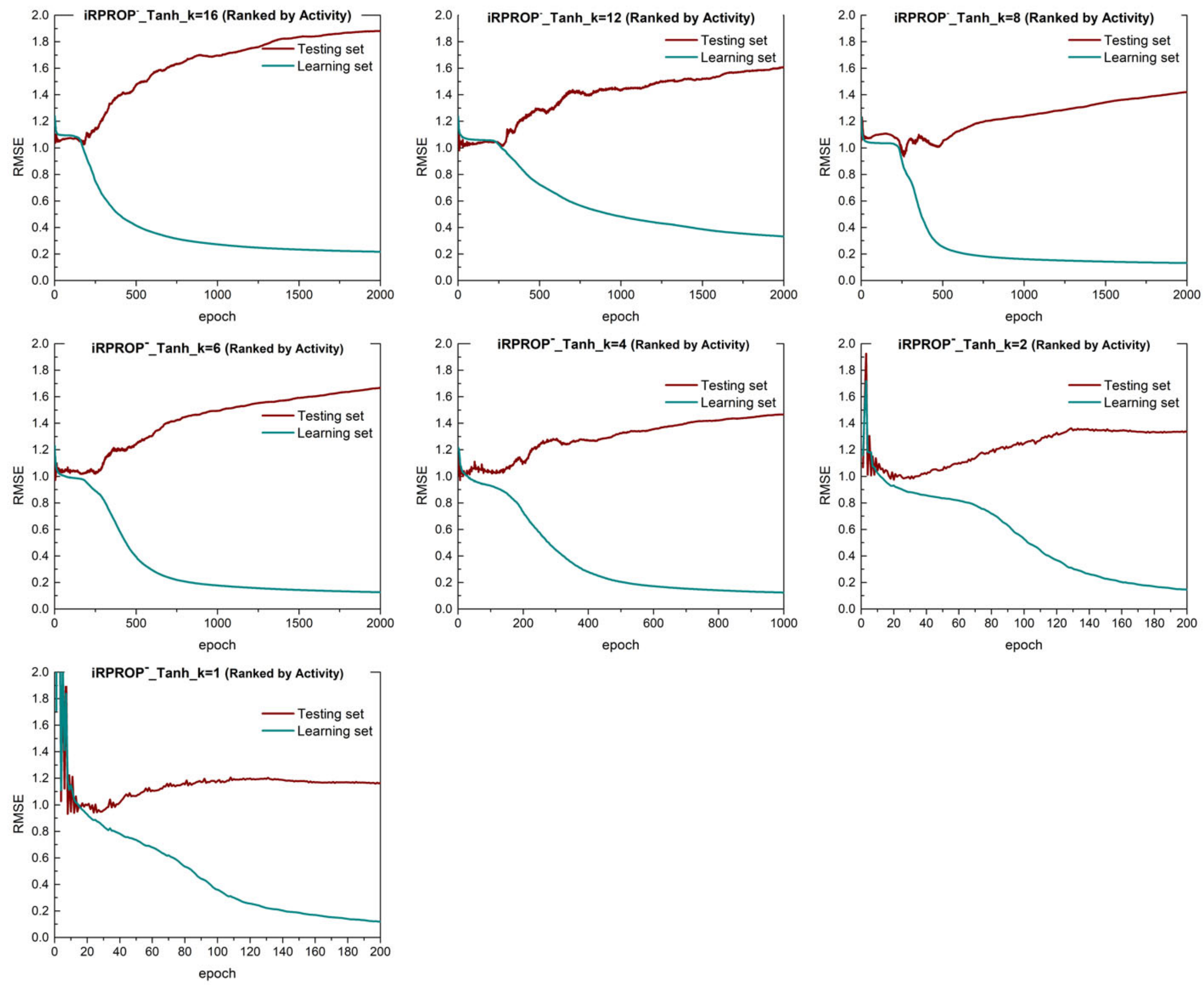

Figure S6. RMSE as a function of epoch for learning (training) and testing sets for an MLP with one hidden layer with iRPROP$^-$ learning algorithm and variable-length-array SMILES representation (Tanh activation, Ranking by Activity-based rational splitting algorithm), Dataset#1.

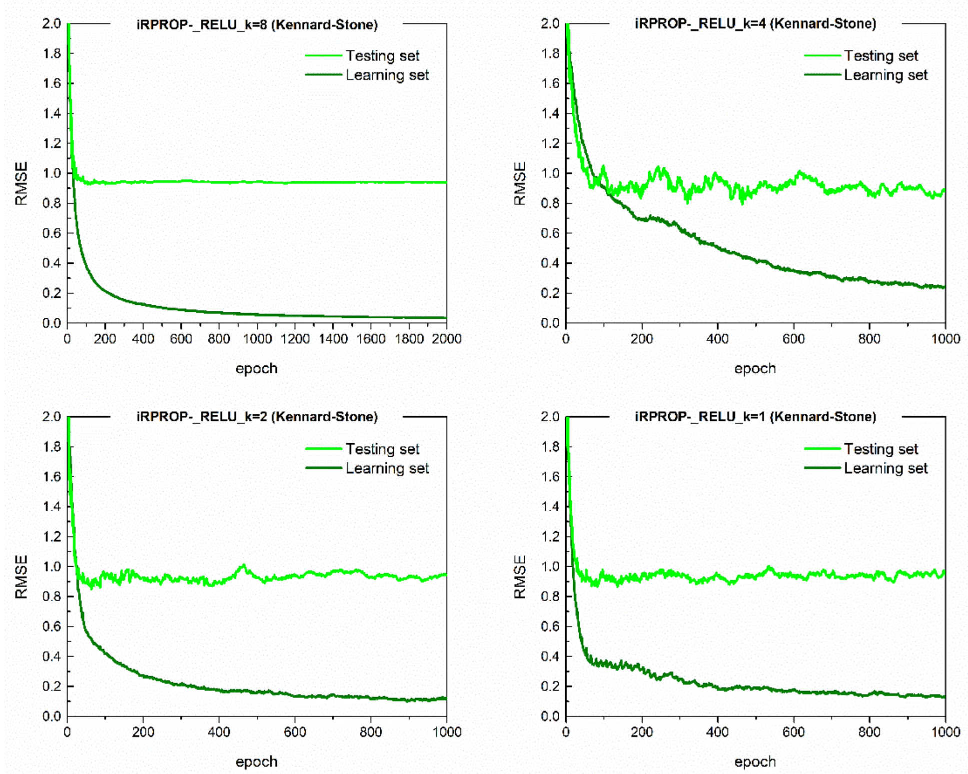

Figure S7. RMSE as a function of epoch for learning (training) and testing sets for an MLP with one hidden layer with iRPROP$^-$ learning algorithm and variable-length-array SMILES representation (RELU activation, Kennard-Stone-based rational splitting algorithm), Dataset#2

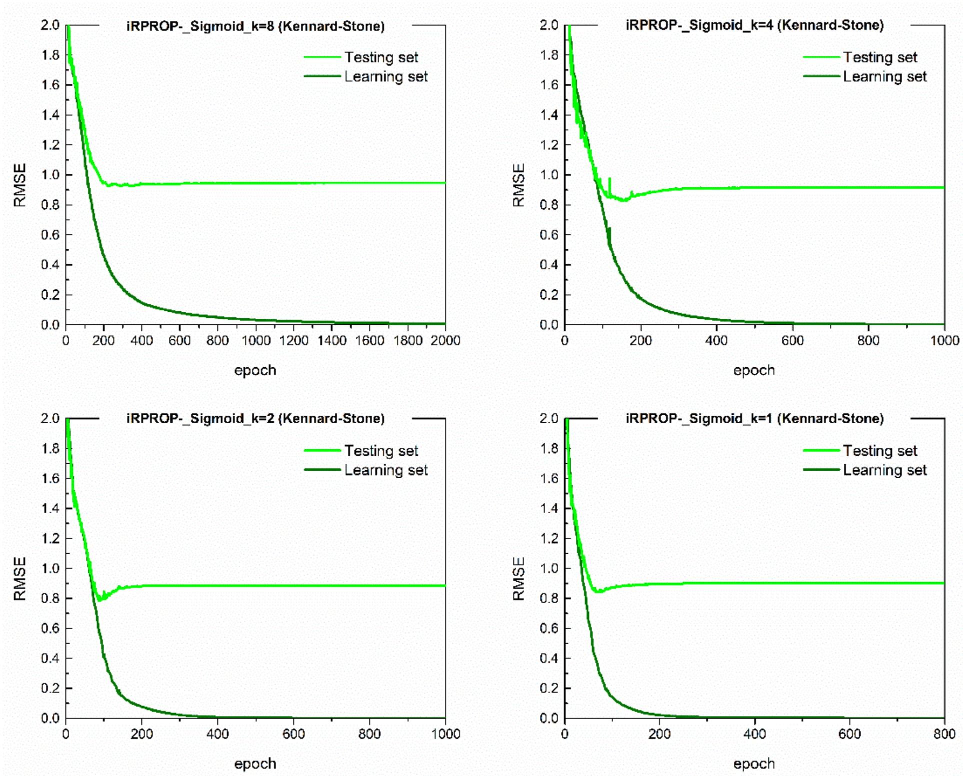

Figure S8. RMSE as a function of epoch for learning (training) and testing sets for an MLP with one hidden layer with iRPROP$^-$ learning algorithm and variable-length-array SMILES representation (Sigmoid activation, Kennard-Stone-based rational splitting algorithm), Dataset#2

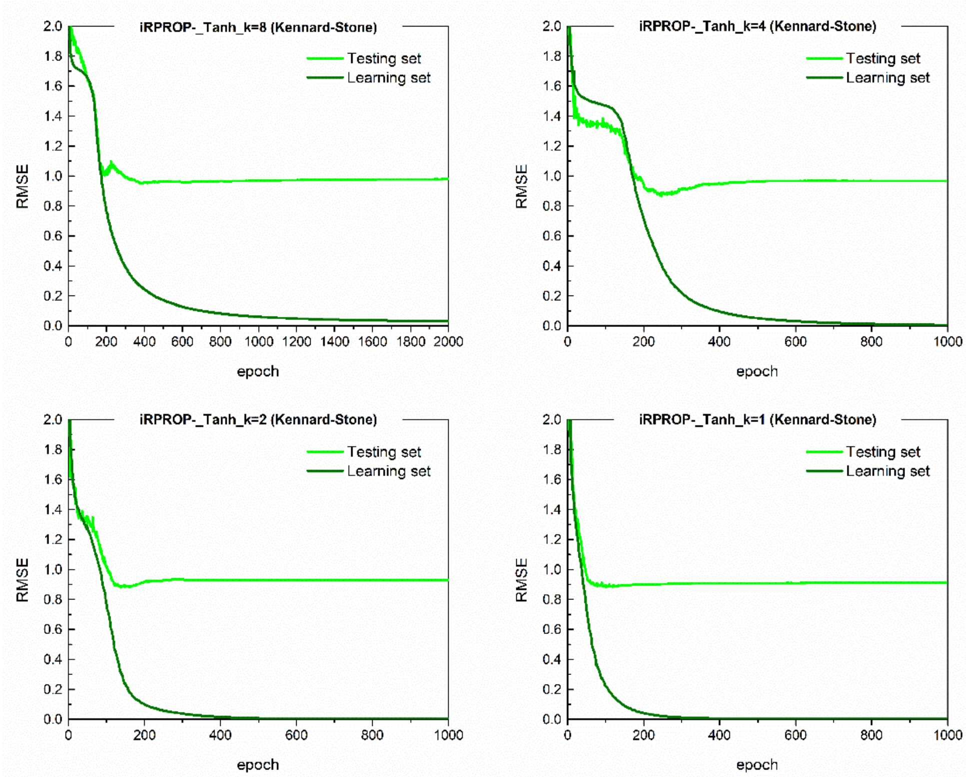

Figure S9. RMSE as a function of epoch for learning (training) and testing sets for an MLP with one hidden layer with iRPROP- learning algorithm and variable-length-array SMILES representation (Tanh activation, Kennard-Stone-based rational splitting algorithm), Dataset#2

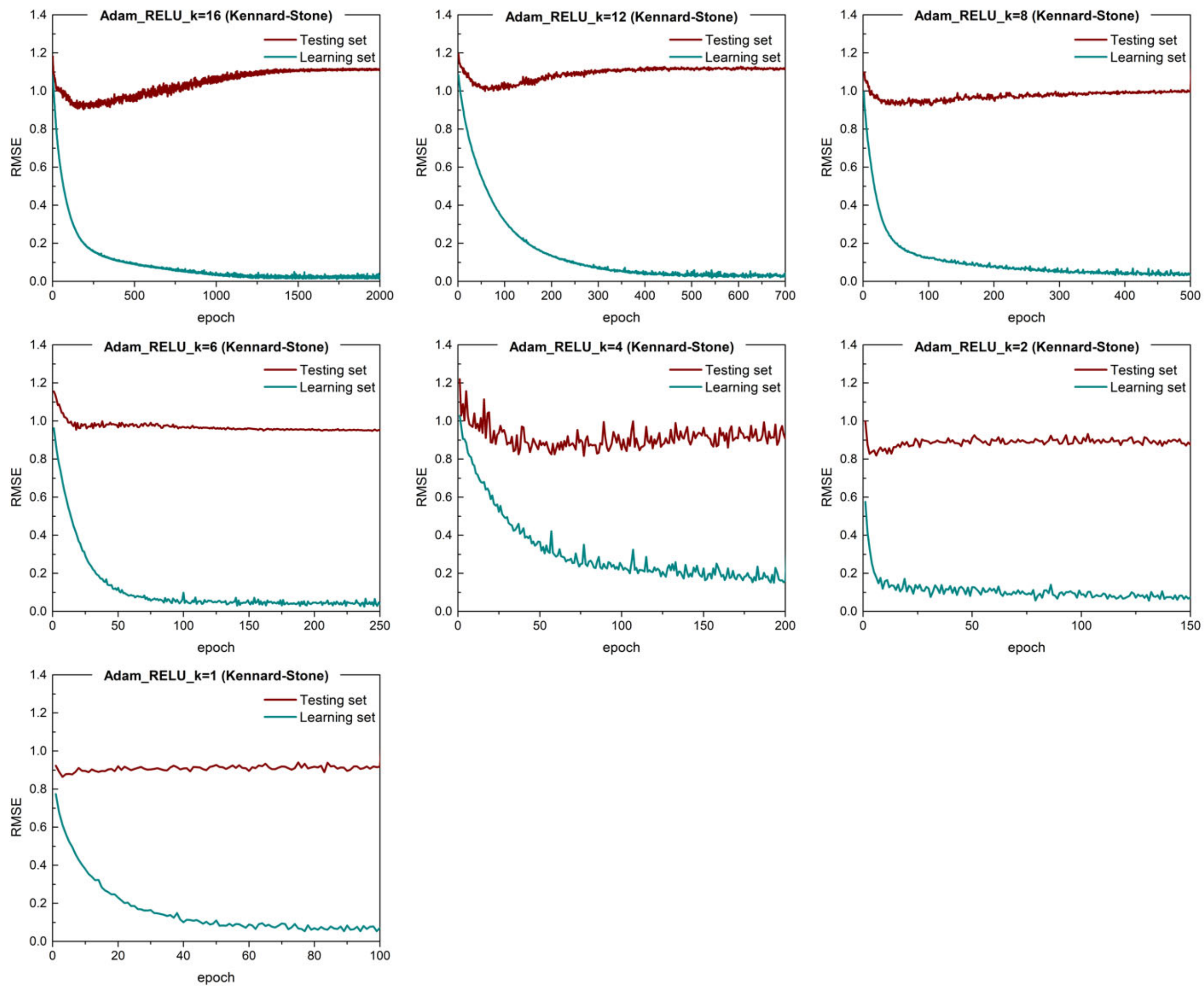

Figure S10. RMSE as a function of epoch for learning (training) and testing sets for an MLP with one hidden layer with Adam learning algorithm and variable-length-array SMILES representation (ReLU activation, Kennard-Stone-based rational splitting algorithm), Dataset#1.

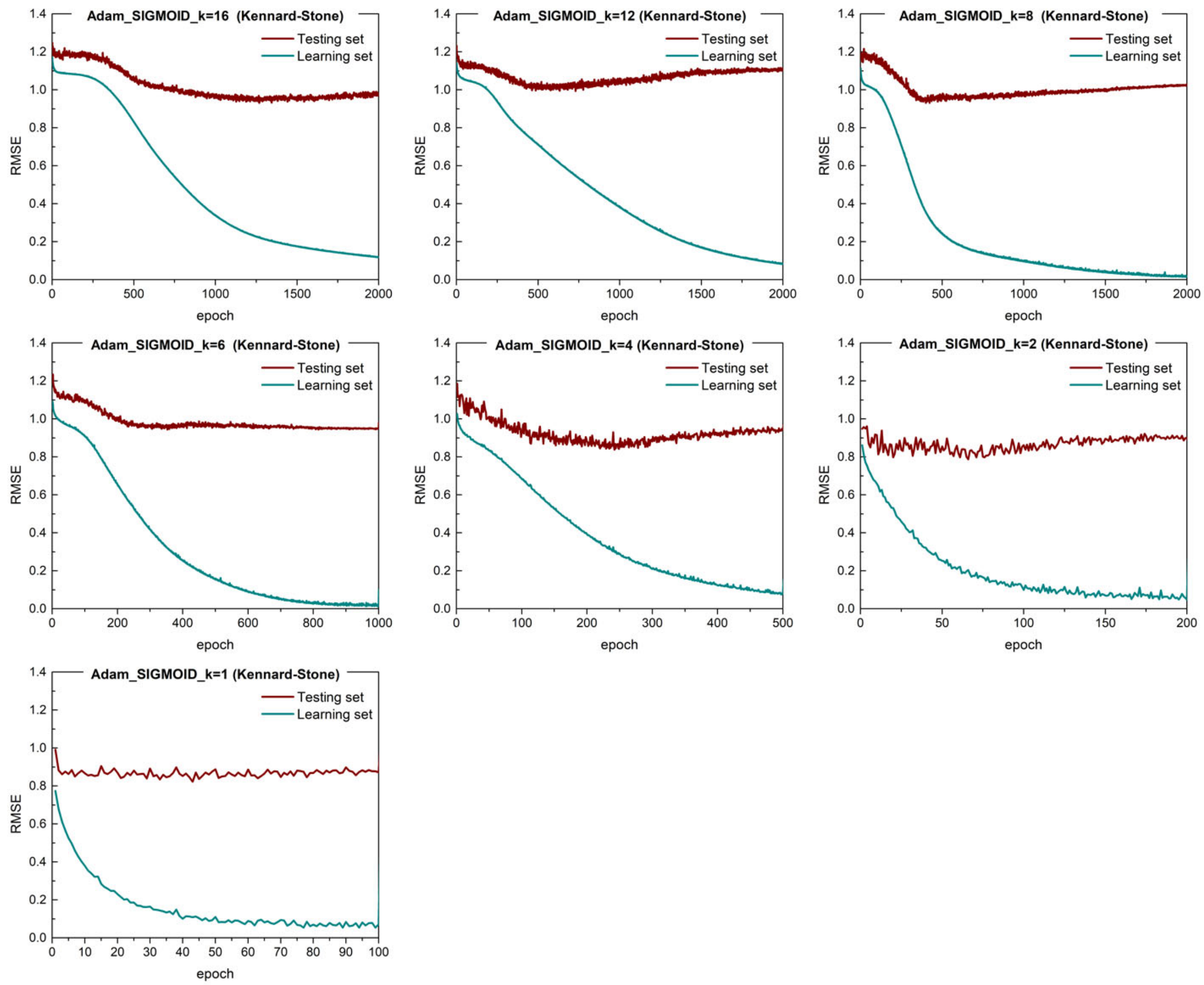

Figure S11. RMSE as a function of epoch for learning (training) and testing sets for an MLP with one hidden layer with Adam learning algorithm and variable-length-array SMILES representation (Sigmoid activation, Kennard-Stone-based rational splitting algorithm), Dataset#1.

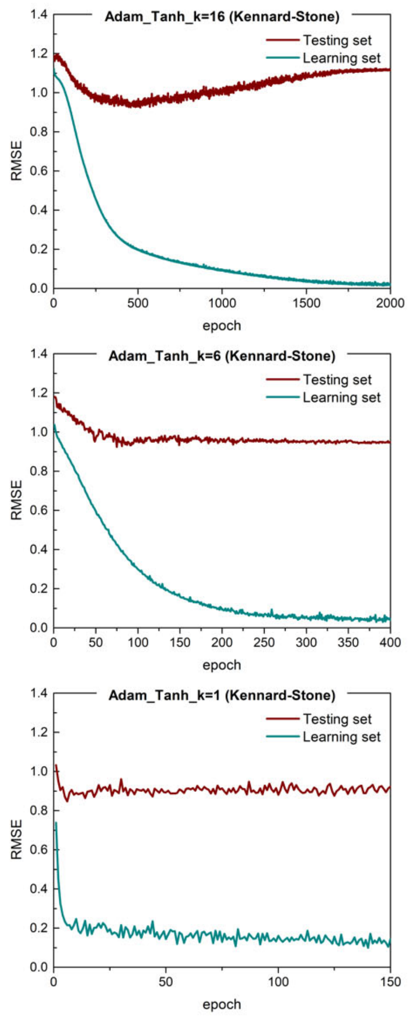

Figure S12. RMSE as a function of epoch for learning (training) and testing sets for an MLP with one hidden layer with Adam learning algorithm and variable-length-array SMILES representation (Tanh activation, Kennard-Stone-based rational splitting algorithm), Dataset#1.

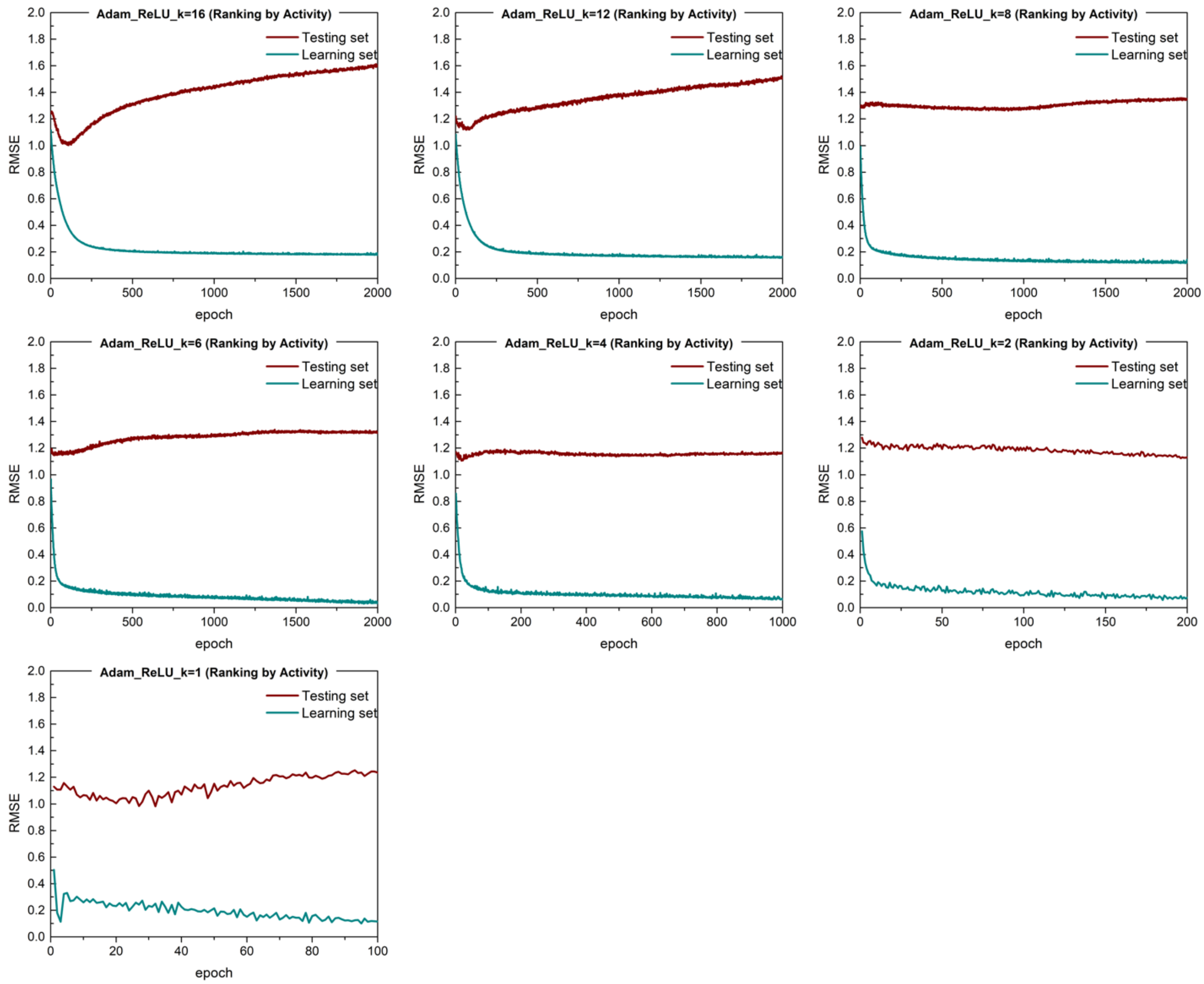

Figure S13. RMSE as a function of epoch for learning (training) and testing sets for an MLP with one hidden layer with Adam learning algorithm and variable-length-array SMILES representation (ReLU activation, Ranking by Activity-based rational splitting algorithm), Dataset#1.

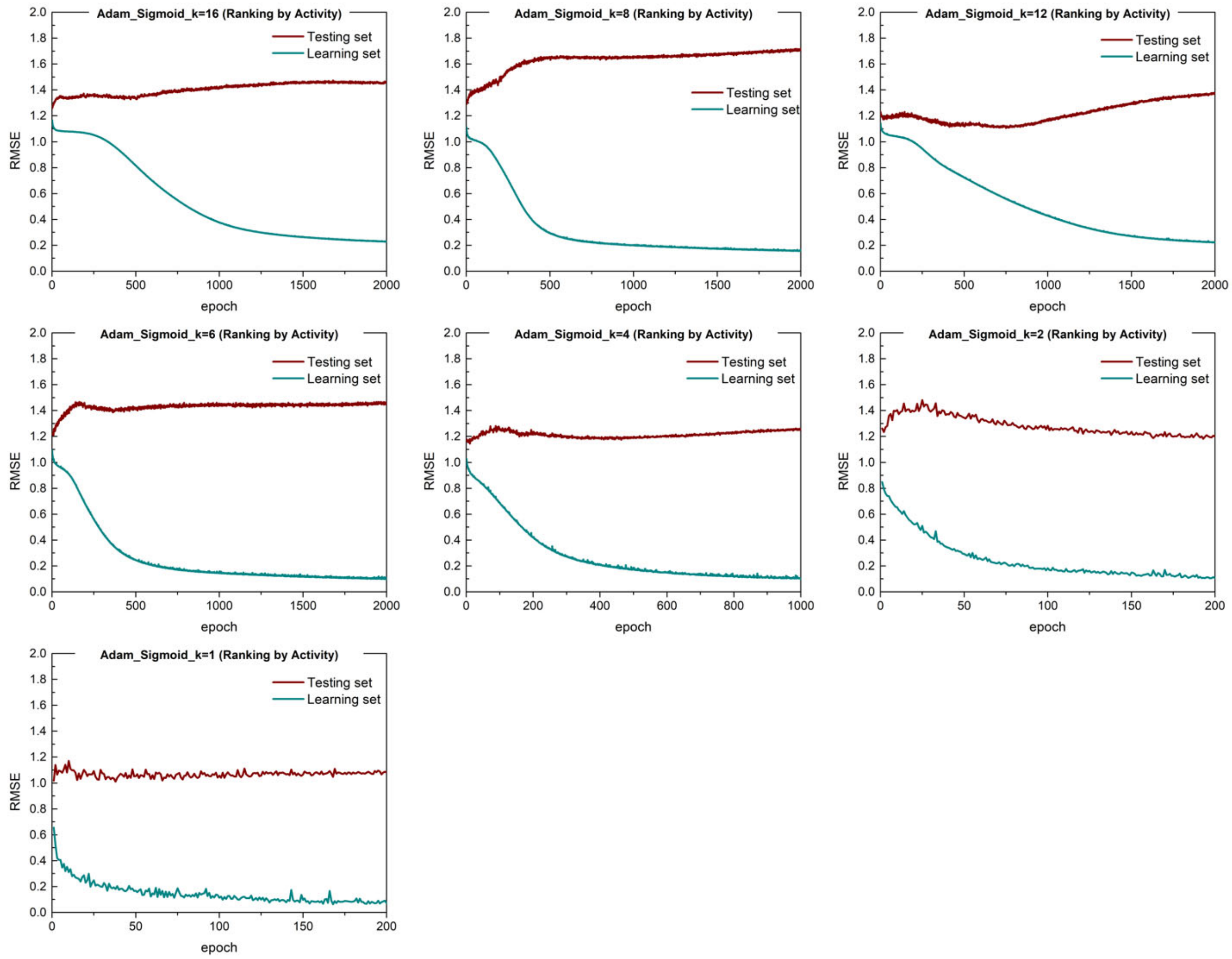

Figure S14. RMSE as a function of epoch for learning (training) and testing sets for an MLP with one hidden layer with Adam learning algorithm and variable-length-array SMILES representation (Sigmoid activation, Ranking by Activity-based rational splitting algorithm), Dataset#1.

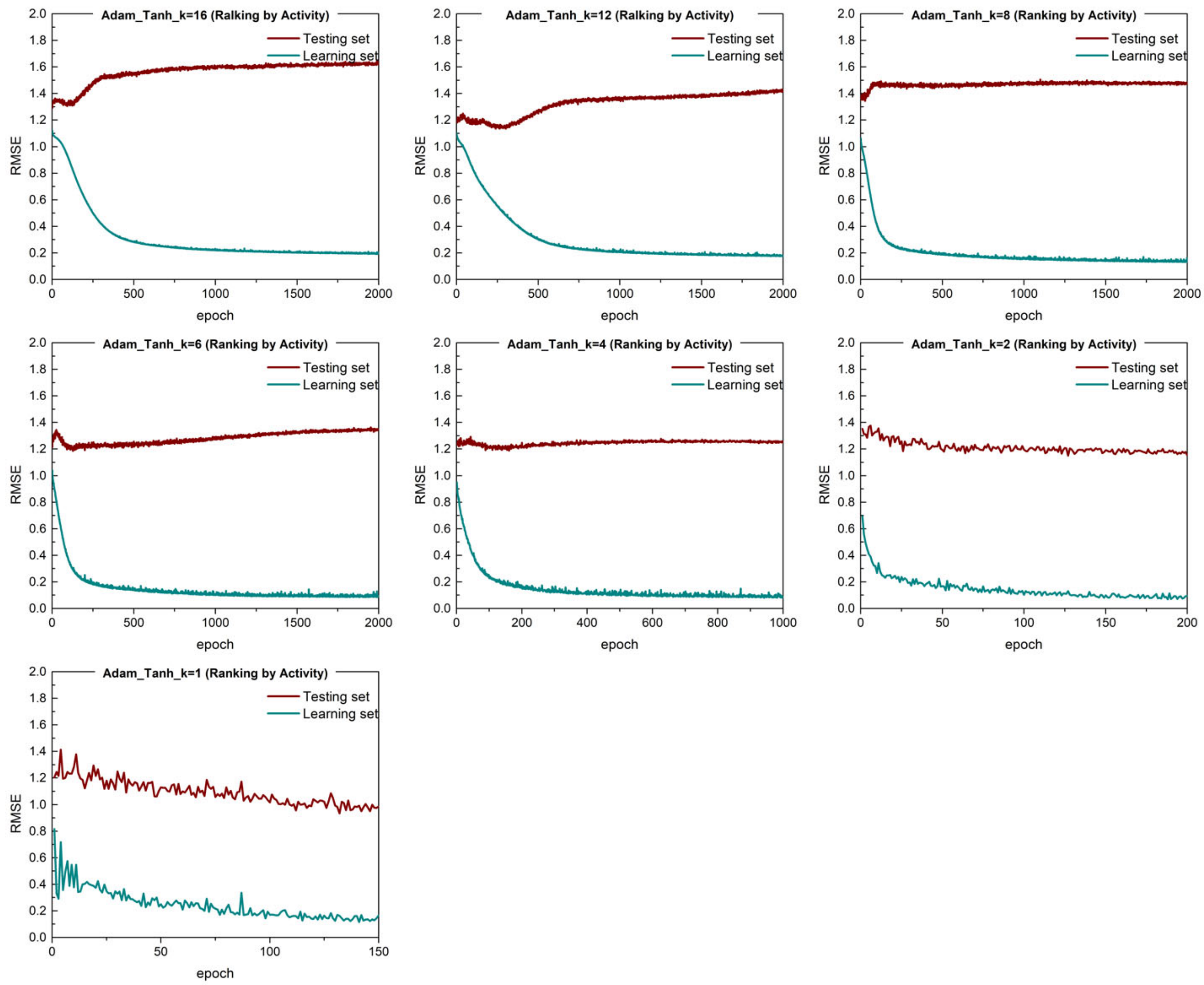

Figure S15. RMSE as a function of epoch for learning (training) and testing sets for an MLP with one hidden layer with Adam learning algorithm and variable-length-array SMILES representation (Tanh activation, Ranking by Activity-based rational splitting algorithm), Dataset#1.

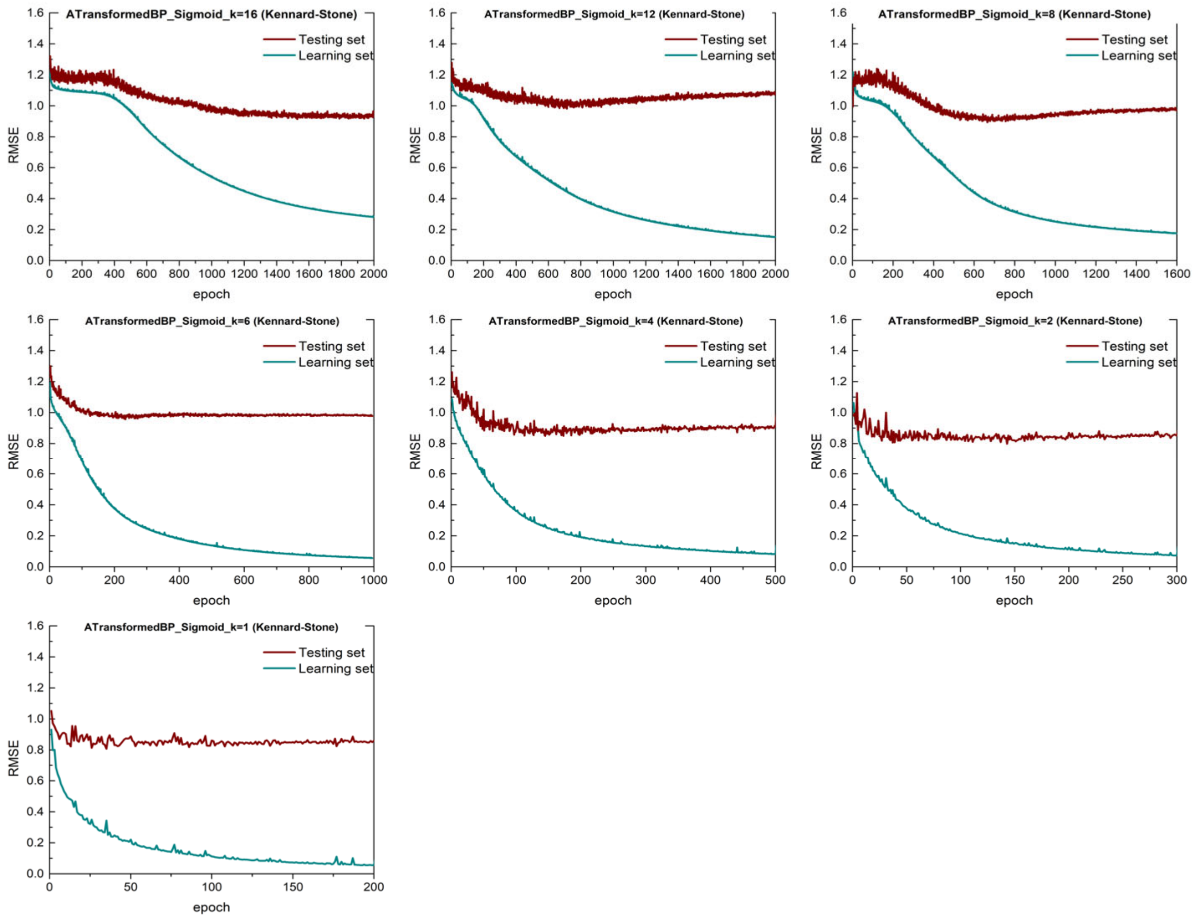

Figure S16. RMSE as a function of epoch for learning (training) and testing sets for an MLP with one hidden layer with ATransformedBP learning algorithm and variable-length-array SMILES representation (Sigmoid activation, Kennard-Stone-based rational splitting algorithm), Dataset#1.

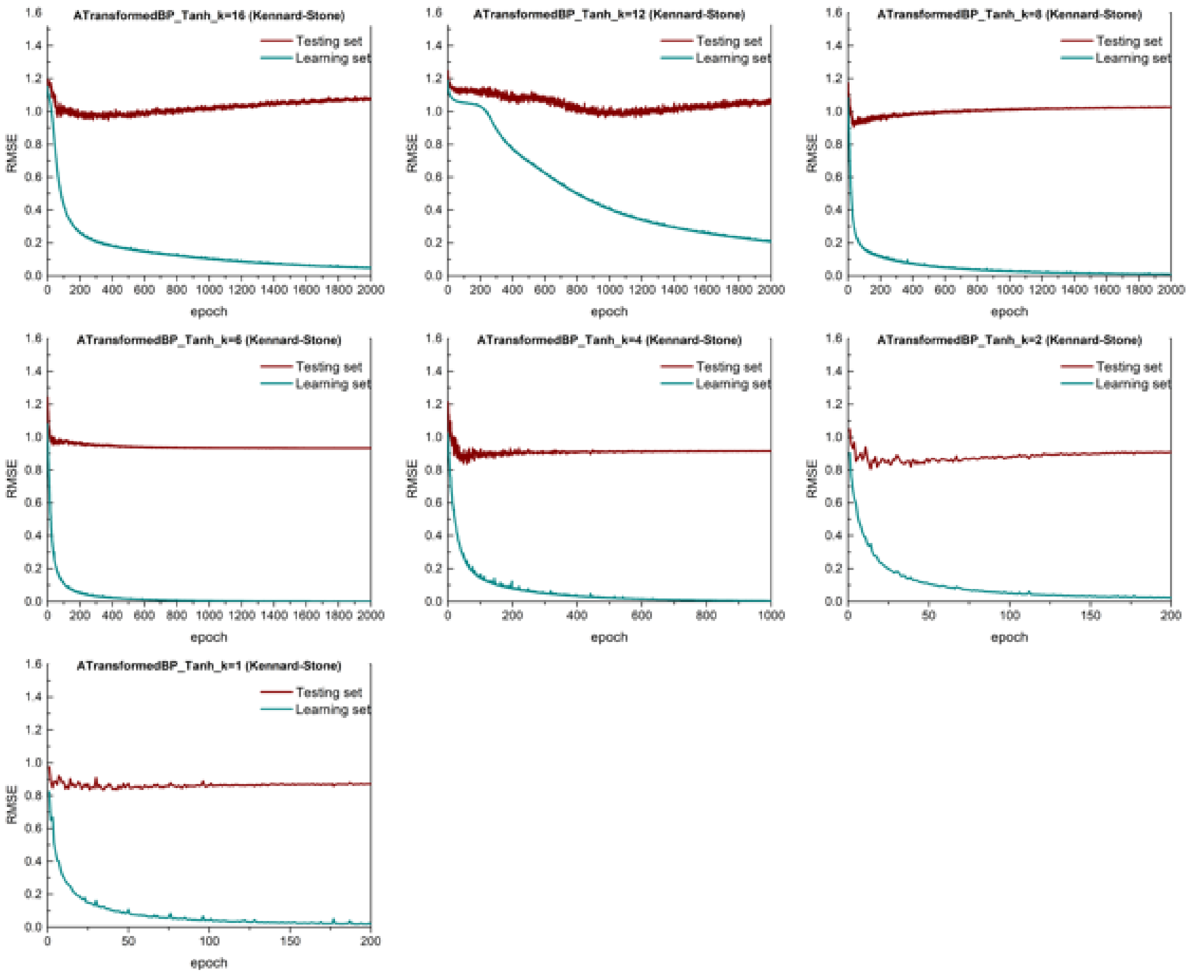

Figure S17. RMSE as a function of epoch for learning (training) and testing sets for an MLP with one hidden layer with ATransformedBP learning algorithm and variable-length-array SMILES representation (Tanh activation, Kennard-Stone-based rational splitting algorithm), Dataset#1.

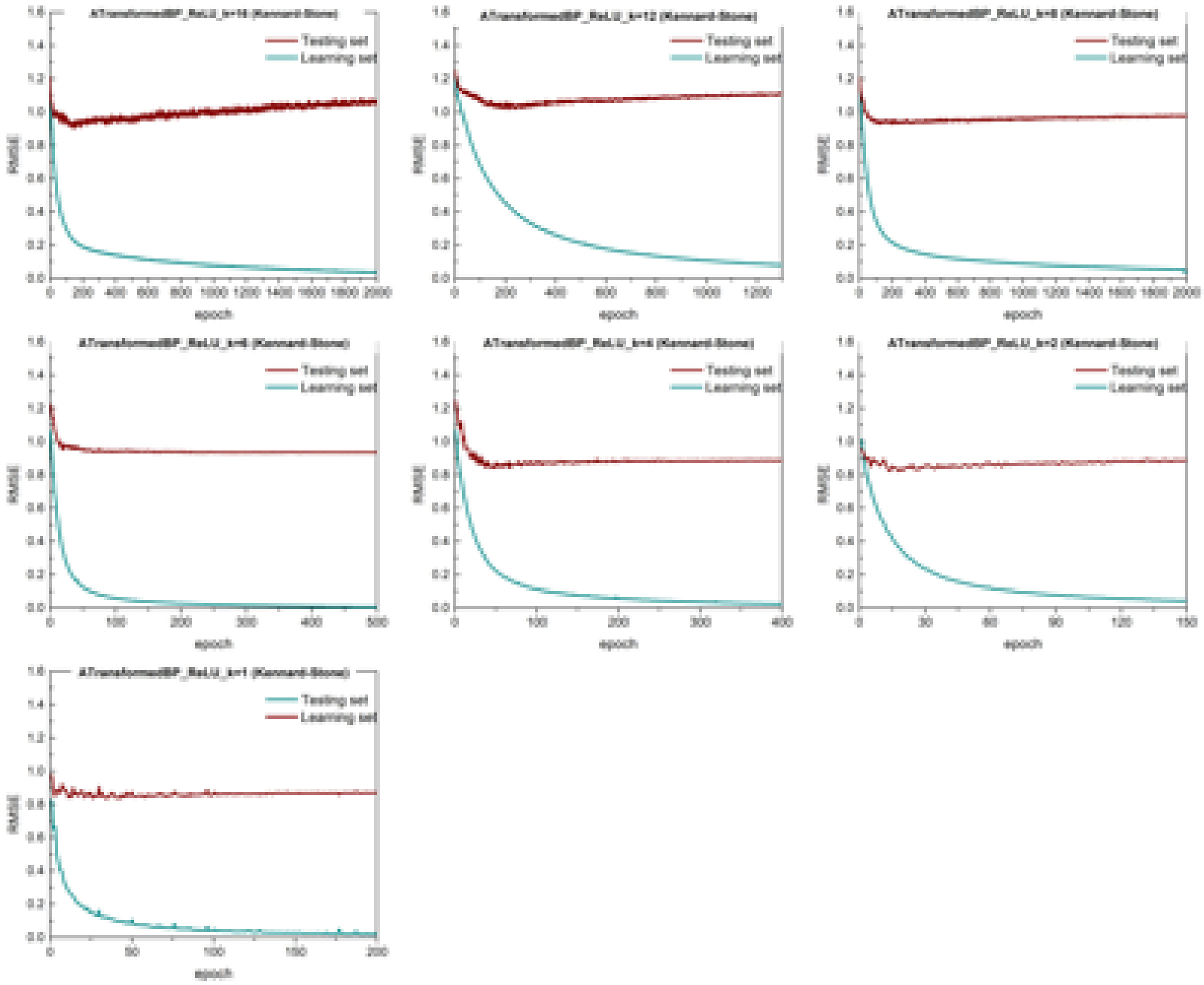

Figure S18. RMSE as a function of epoch for learning (training) and testing sets for an MLP with one hidden layer with ATransformedBP learning algorithm and variable-length-array SMILES representation (Tanh activation, Kennard-Stone-based rational splitting algorithm), Dataset#1.

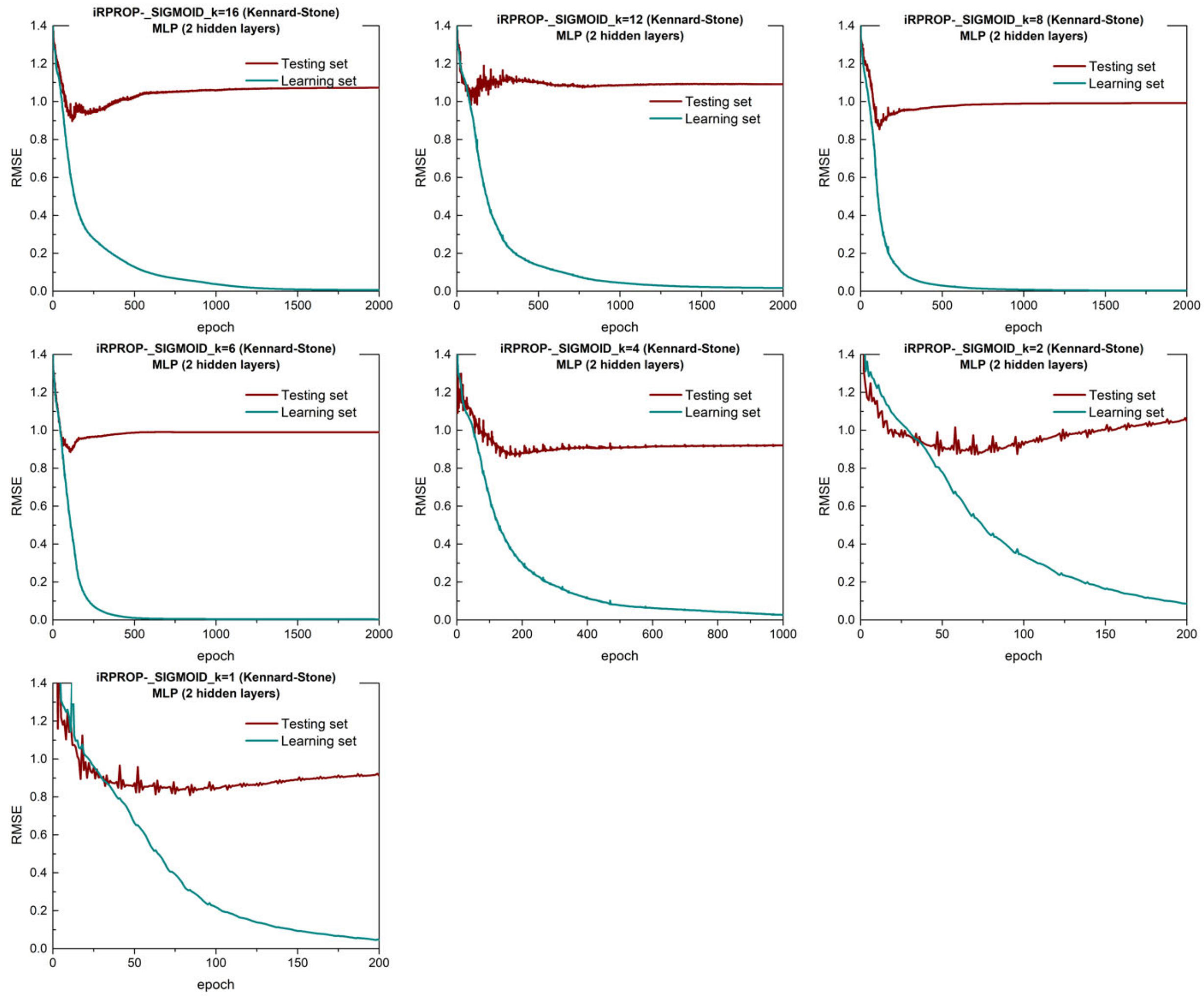

Figure S19. RMSE as a function of epoch for learning (training) and testing sets for an MLP with two hidden layers with iRPROP$^-$ learning algorithm and variable-length-array SMILES representation (Sigmoid activation, Kennard-Stone-based rational splitting algorithm), Dataset#1.

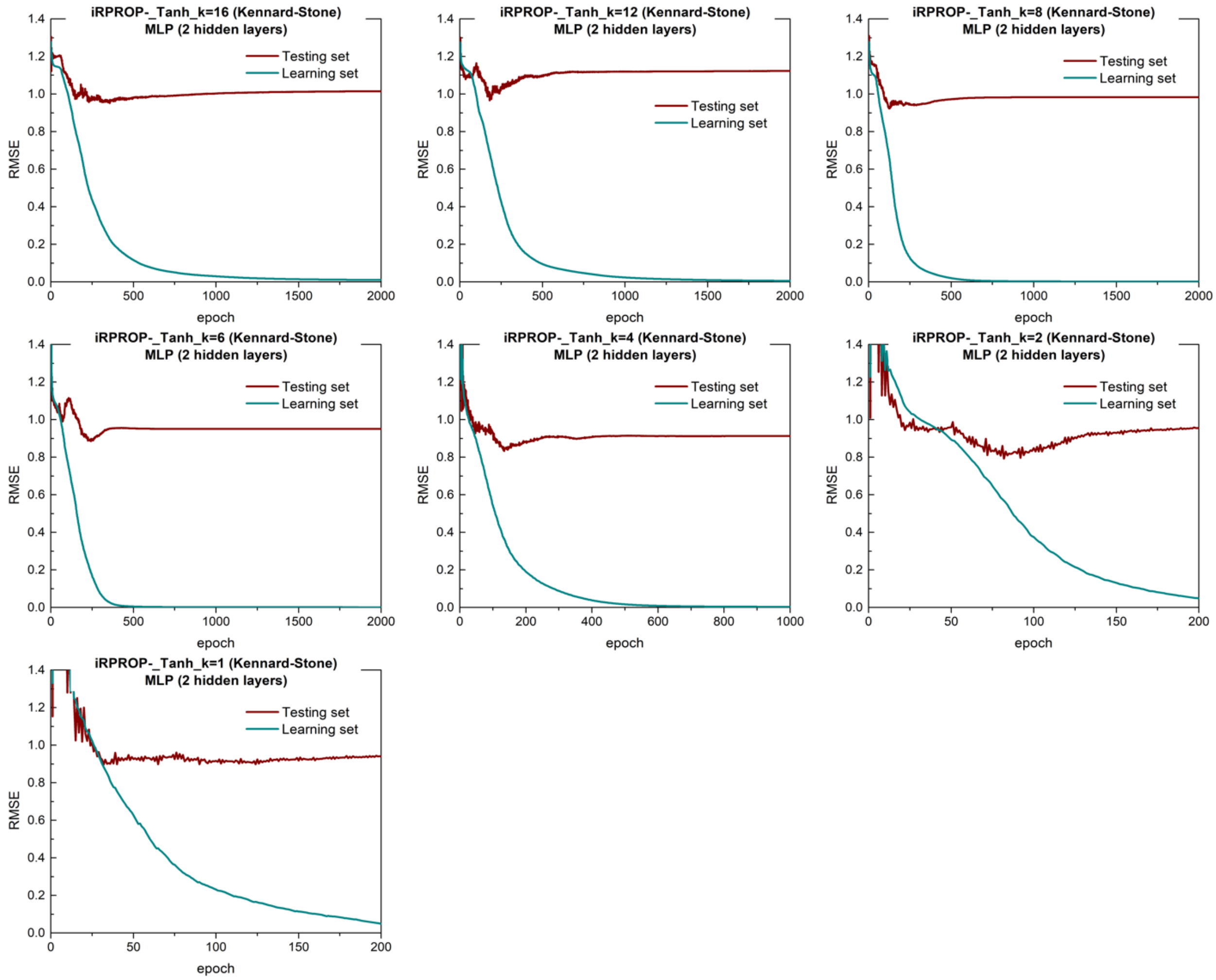

Figure S20. RMSE as a function of epoch for learning (training) and testing sets for an MLP with two hidden layers with iRPROP^{-} learning algorithm and variable-length-array SMILES representation (Tanh activation, Kennard-Stone-based rational splitting algorithm), Dataset#1.

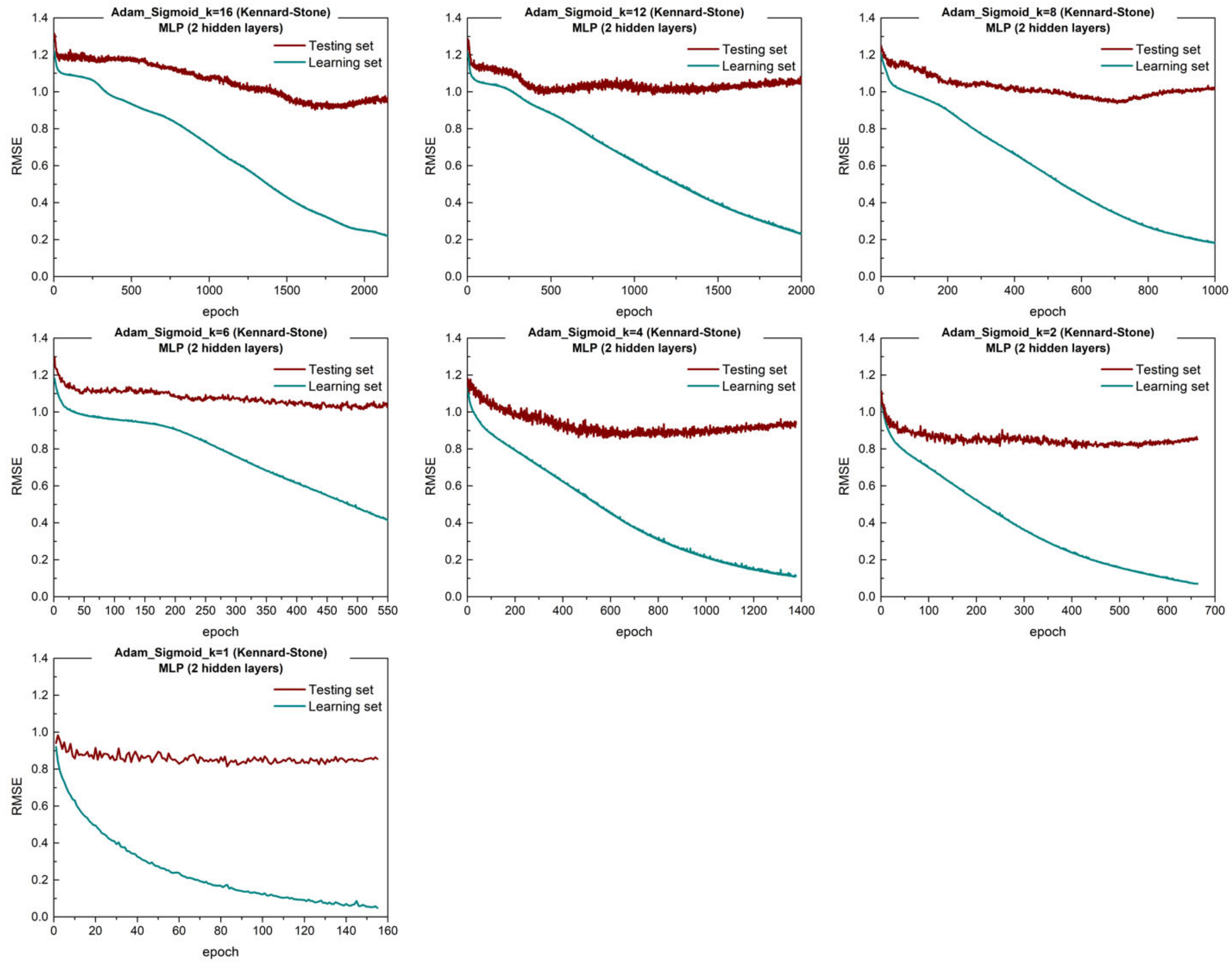

Figure S21. RMSE as a function of epoch for learning (training) and testing sets for an MLP with two hidden layers with Adam learning algorithm and variable-length-array SMILES representation (Sigmoid activation, Kennard-Stone-based rational splitting algorithm), Dataset#1.

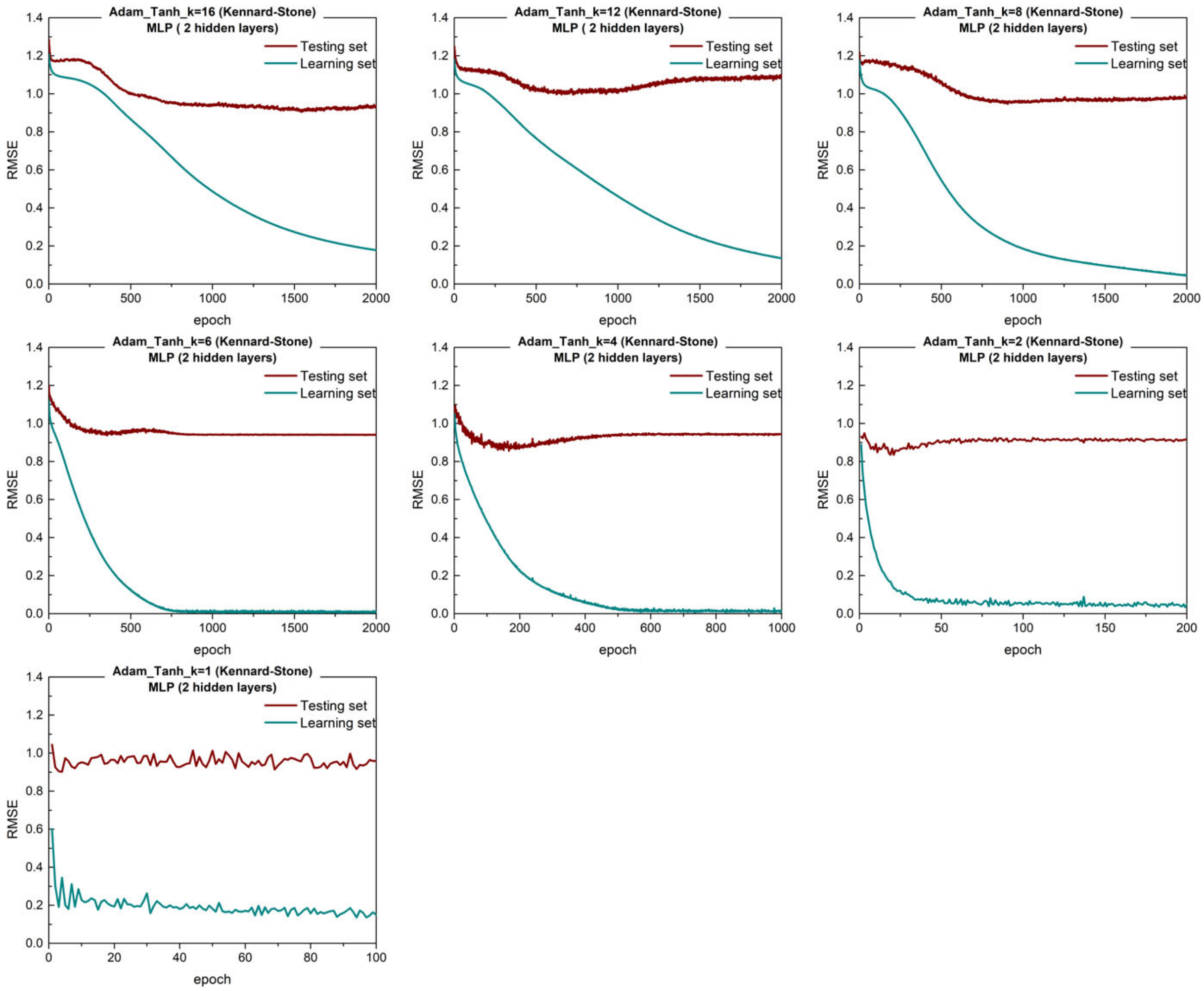

Figure S22. RMSE as a function of epoch for learning (training) and testing sets for an MLP with two hidden layers with Adam learning algorithm and variable-length-array SMILES representation (Tanh activation, Kennard-Stone-based rational splitting algorithm), Dataset#1.

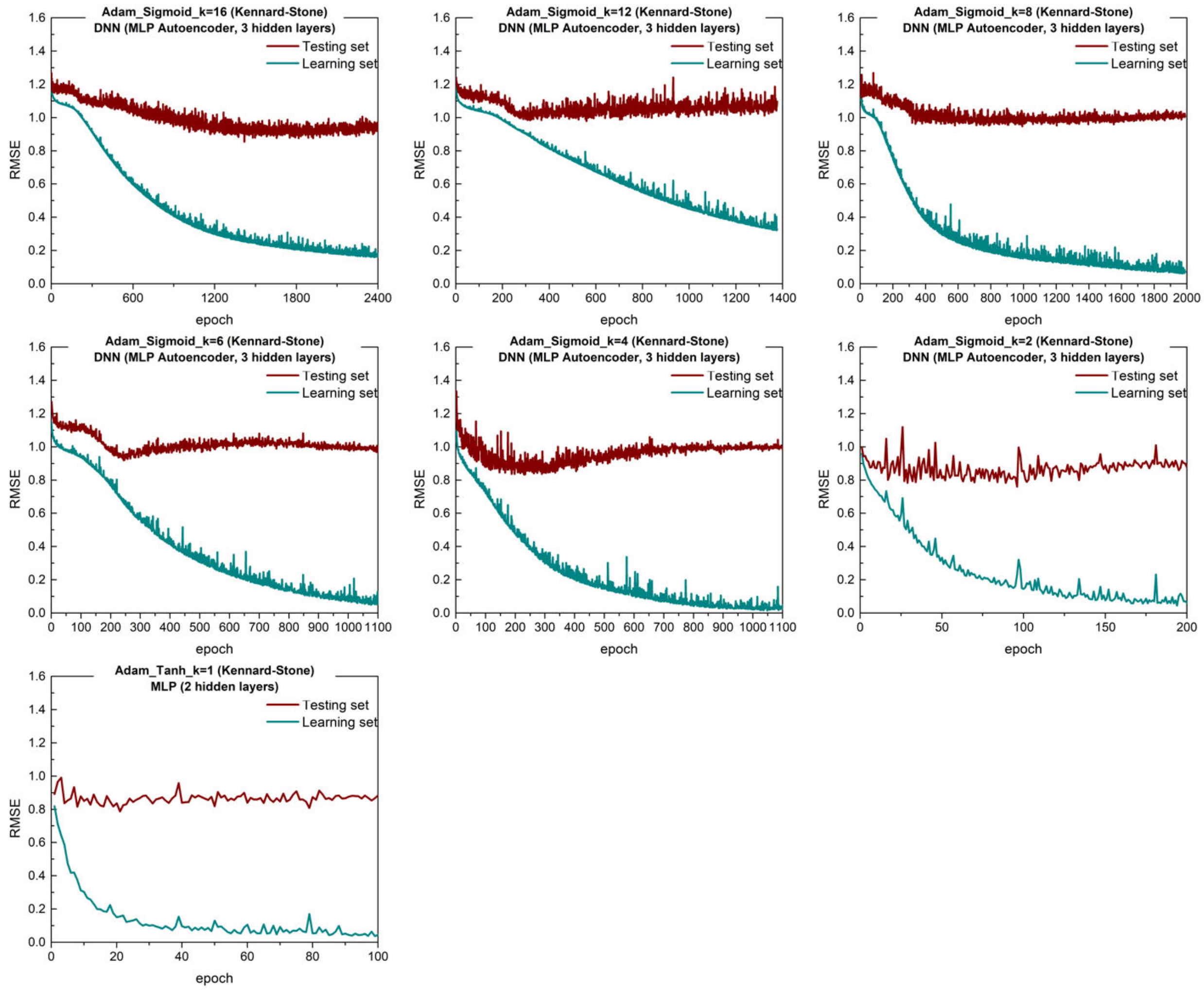

Figure S23. RMSE as a function of epoch for learning (training) and testing sets for an MLP with Autoencoder (three hidden layers with Adam learning algorithm and variable-length-array SMILES representation, Tanh activation, Kennard-Stone-based rational splitting algorithm), Dataset#1.

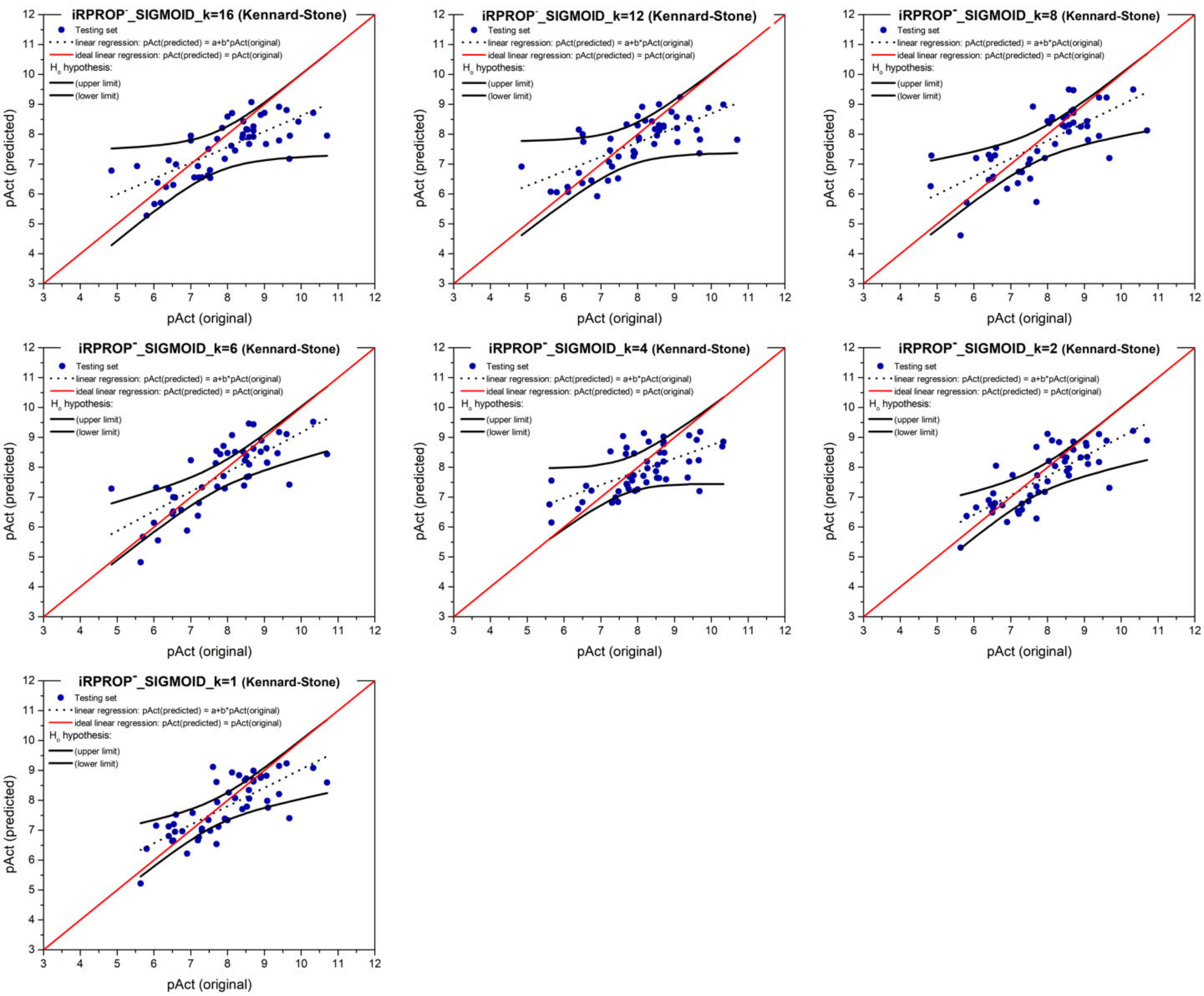

Figure S24. Parity plot and H_0 hypothesis testing for a single-layer MLP prediction model with iRPROP^{-} learning algorithm and variable-length-array SMILES representation (*Sigmoid(Y)* activation, Kennard-Stone-based rational splitting algorithm), Dataset#1.