Electronic supplementary materials *Mendeleev Commun.*, 2026, **36**, 257–259

Overfitting in machine learning models for predicting partial atomic charges of drug-like molecules

Denis V. Zverev, Polina K. Nikiforova, Arslan R. Shaimardanov, Dmitry A. Shulga and Vladimir A. Palyulin

S1. Methods

The virtual library contained 6,878 molecules and was constructed from PDBBind 2020[S1] with the addition of extra charged and neutral molecular forms. For each structure, MEP calculation was performed using the RHF/6-31G* level of theory in Firefly QC 8.2.0.[S2] RESP charges were then computed from the resulting electrostatic potential with AmberTools22.[S3] For each molecular structure, AtomPair fingerprints[S4] with path length 4 were generated using RDKit.[S5] Clustering was carried out in RDKit using Tanimoto similarity[S6] and the Butina algorithm,[S7] yielding 1,260 clusters, of which 342 were singletons (i.e., molecules lacking structurally similar neighbors above the similarity threshold of 0.2).

The RF used parameters reported in the literature.[S8] The MLP comprised an input layer with 2,048 neurons with ReLu activation functions, two hidden layers with 64 neurons each, also with ReLu activation functions, and an output neuron (Figure S1). The inputs were the 2,048-bit AtomPair fingerprint vectors (RDKit) for all atoms of a given element present in the dataset; the outputs were the corresponding partial atomic charges.

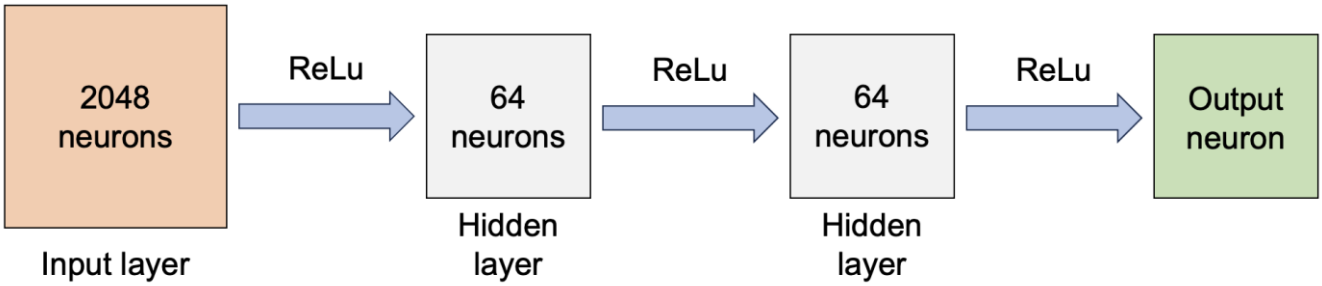
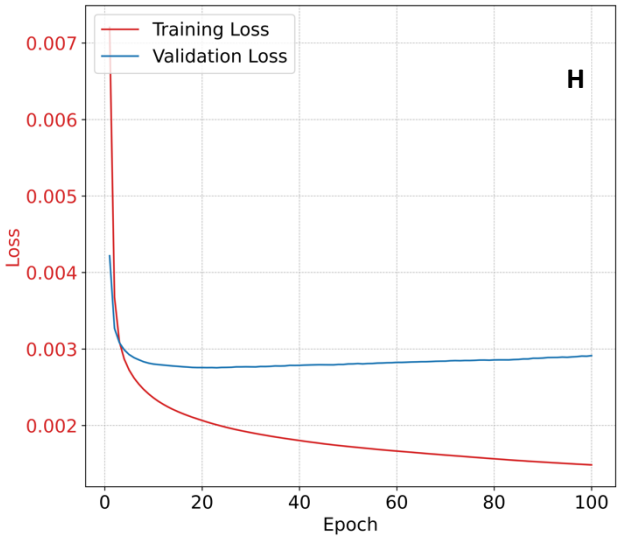
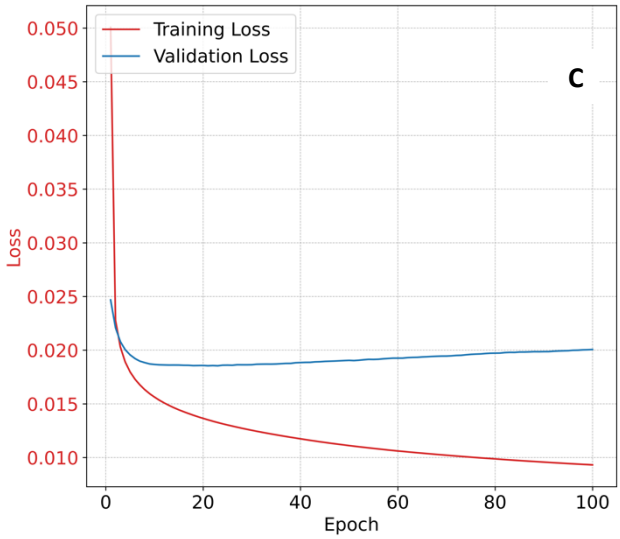
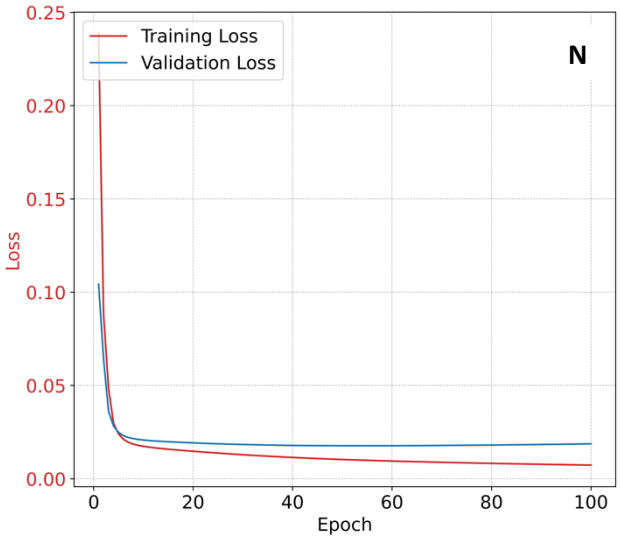


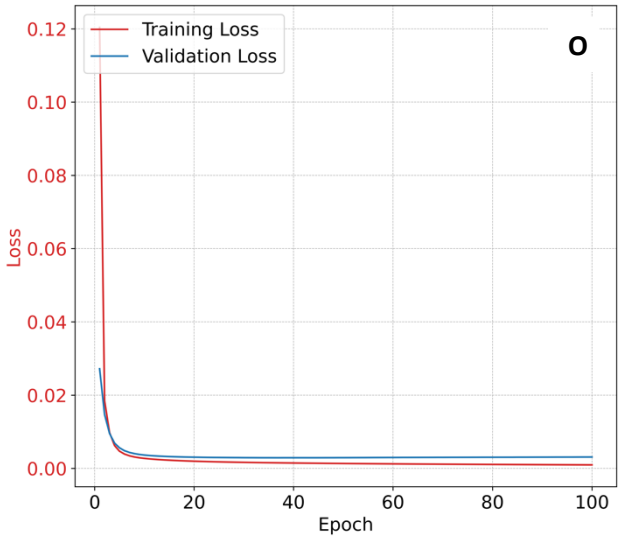
Figure S1 Schematic representation of the MLP model architecture.

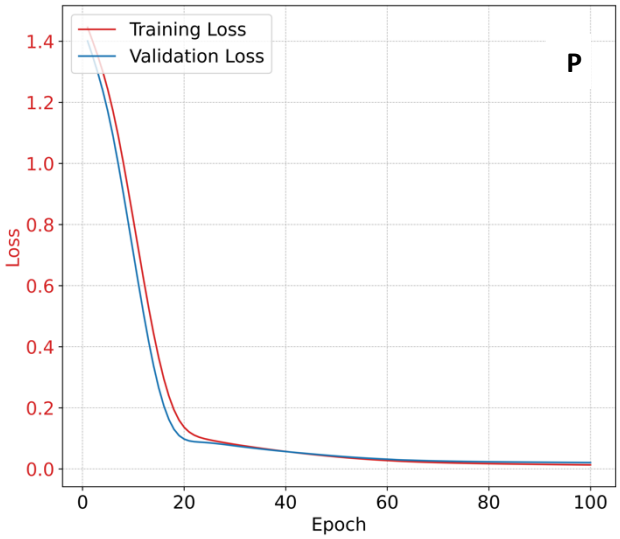
The dataset was split randomly into training (70%) and test (30%) subsets by cluster count, maintaining this proportion resulted in 4496 structures (65%) in the training set and 2381 structures (35%) in the test set. The RF was trained using Scikit-learn;[S9] the MLP was trained in PyTorch[S10] using stochastic gradient descent with the Adam optimizer and no regularization. The number of training epochs was 100 in order to achieve possible overfitting. For all models, the loss function was the mean square error (MSE) with respect to RESP charges. The learning curves of the MLP models for each chemical element are shown in Figure S2 for a single fold. The results for the remaining folds are identical and are not shown.

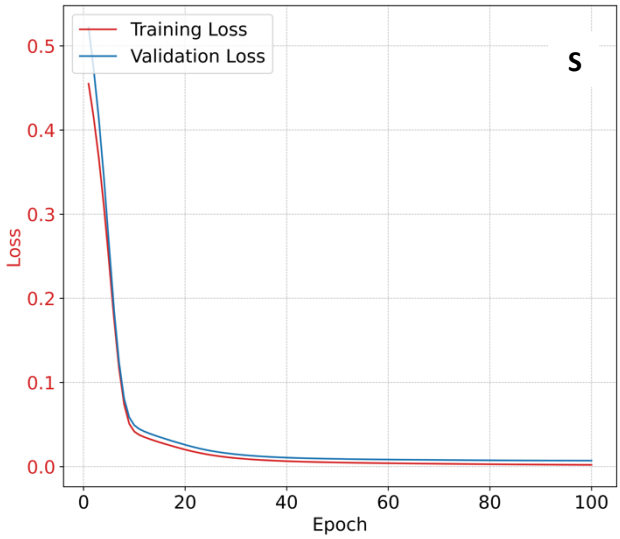

Training Loss
Validation Loss
H
Loss
Epoch


Training Loss
Validation Loss
C
Loss
Epoch


Training Loss
Validation Loss
N
Loss
Epoch


Training Loss
Validation Loss
O
Loss
Epoch


Training Loss
Validation Loss
P
Loss
Epoch


Training Loss
Validation Loss
S
Loss
Epoch

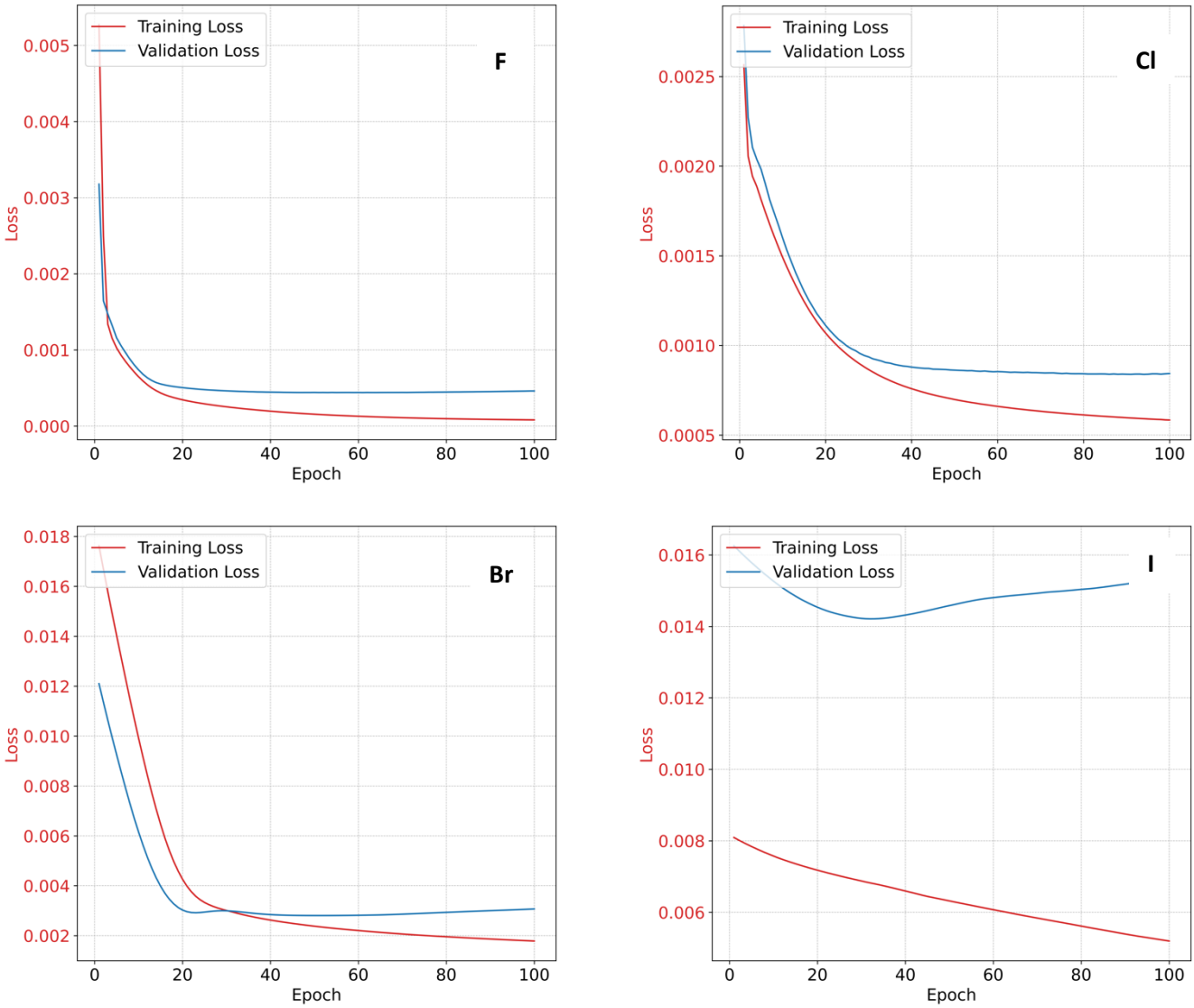


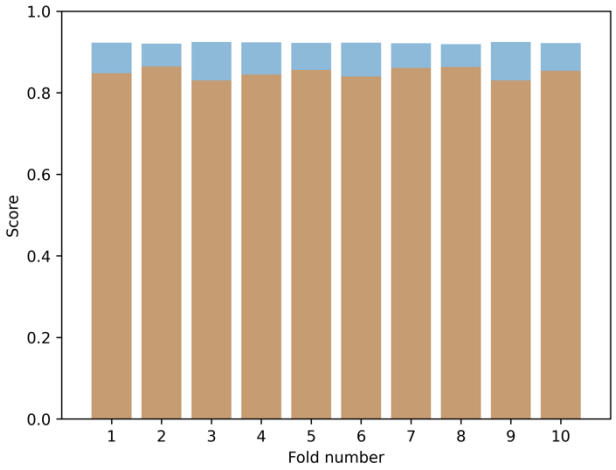
Figure S2 Training (red) and test (blue) loss dynamics (MSE) on a single fold for the MLP models, calculated for atoms of each chemical element.

S2. Cross-validation

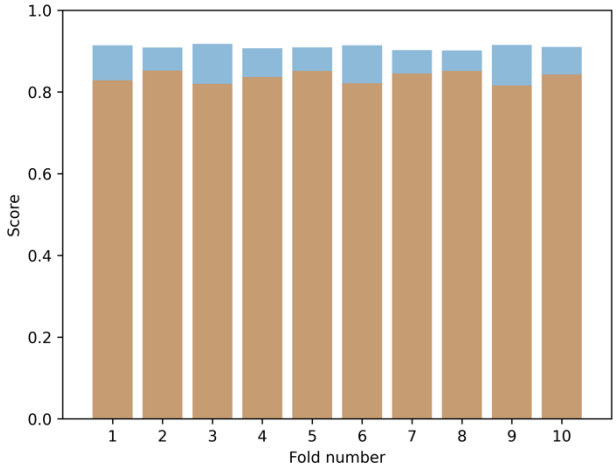
Table S1 10-fold cross-validation average values of the R^2 correlation coefficient and root mean square error (RMSE) for the RF and MLP models on the training and test sets, calculated for halogen atoms.

Element	R^2 RF Train	R^2 MLP Train	R^2 RF Test	R^2 MLP Test	RMSE RF Train	RMSE MLP Train	RMSE RF Test	RMSE MLP Test
F[a]	0.785	0.870	0.491	0.510	0.016	0.012	0.024	0.023
Cl[b]	0.551	0.719	0.324	0.491	0.031	0.024	0.038	0.024
Br[b]	0.447	0.572	-0.007	0.050	0.048	0.042	0.054	0.054
I[b]	0.232	0.209	-0.080	-0.084	0.079	0.080	0.069	0.097

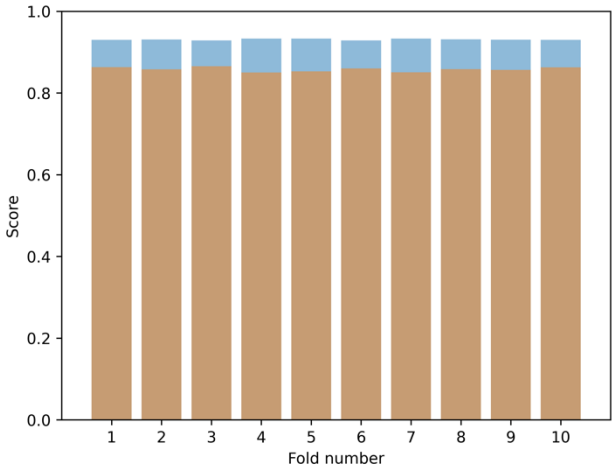
[a] The low metrics for fluorine atom models are presumably related to the wide distribution of RESP charges of spatially shielded fluorine atoms in similar atomic environments.
[b] The low metrics for heavy halogen models are presumably related to the high anisotropy of electrostatic potential in aryl halides and the small amount of training data.



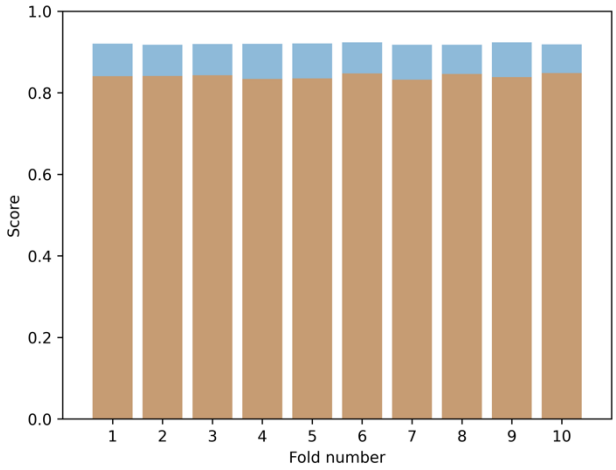
H (RF)



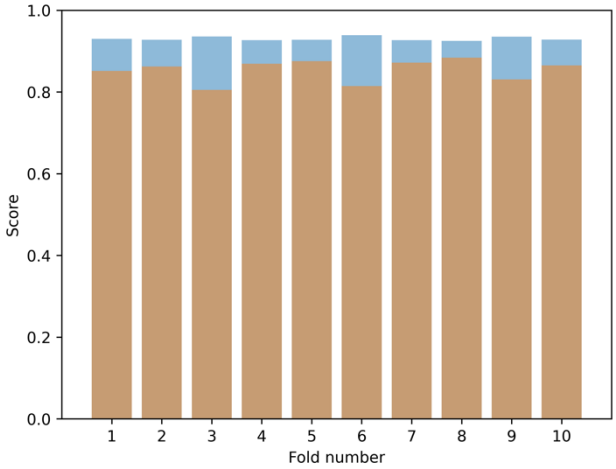
H (MLP)



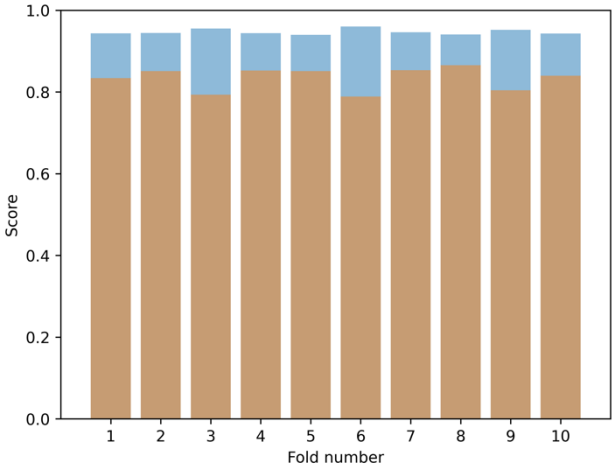
C (RF)



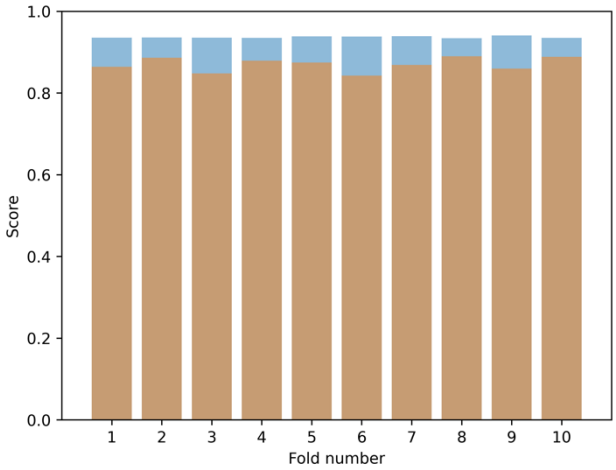
C (MLP)



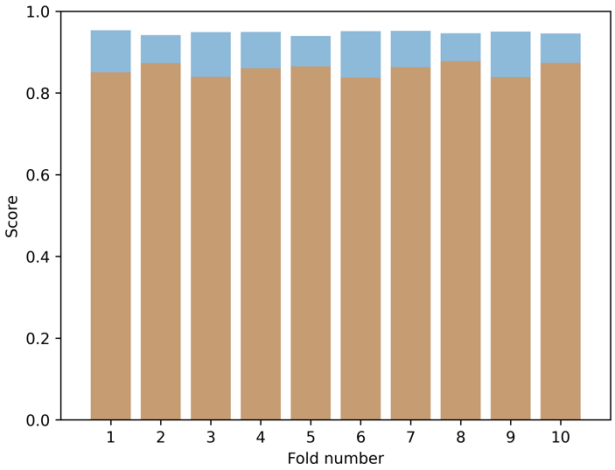
N (RF)



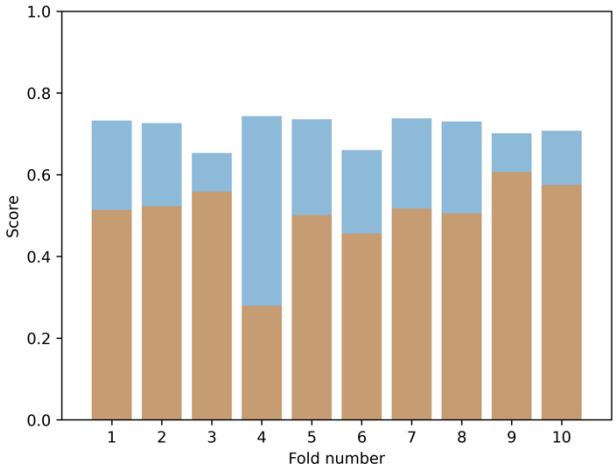
N (MLP)



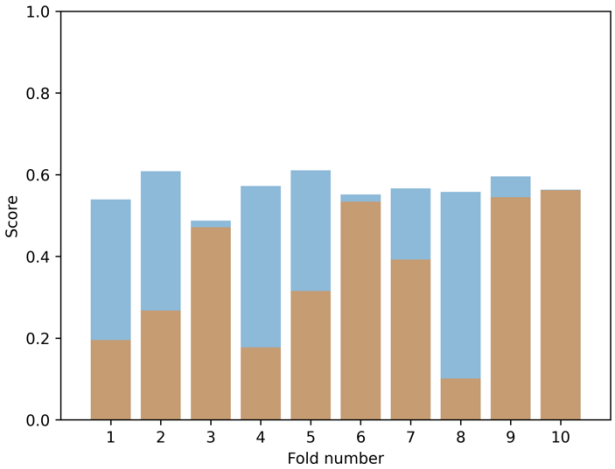
O (RF)



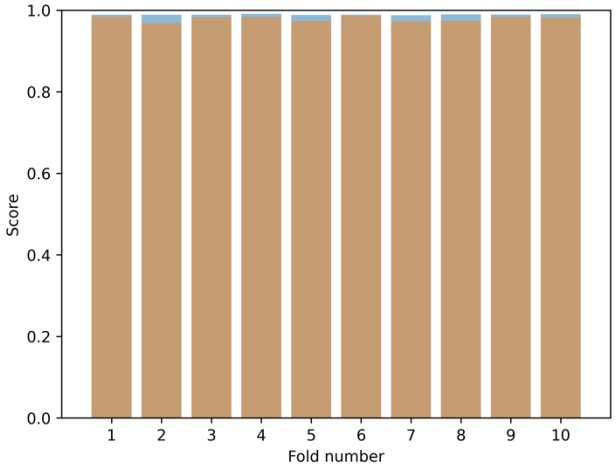
O (MLP)



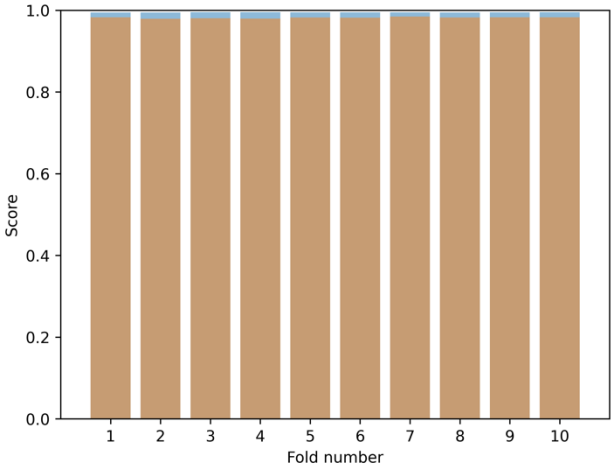
P (RF)



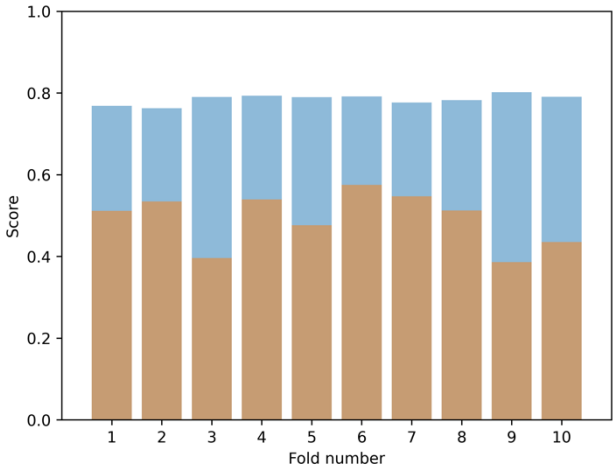
P (MLP)



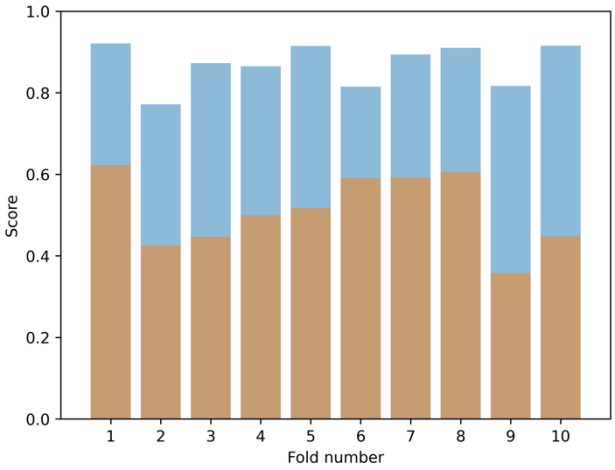
S (RF)



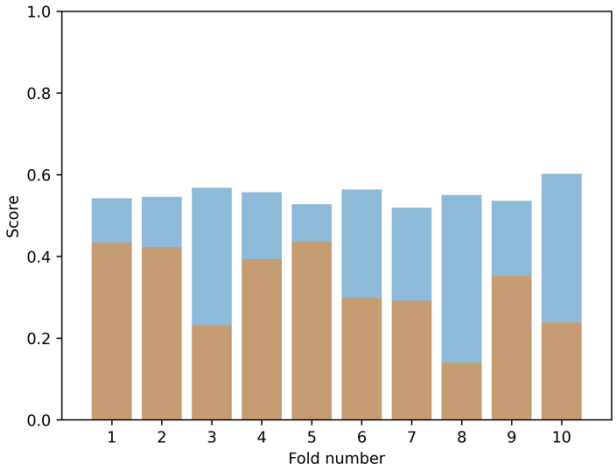
S (MLP)



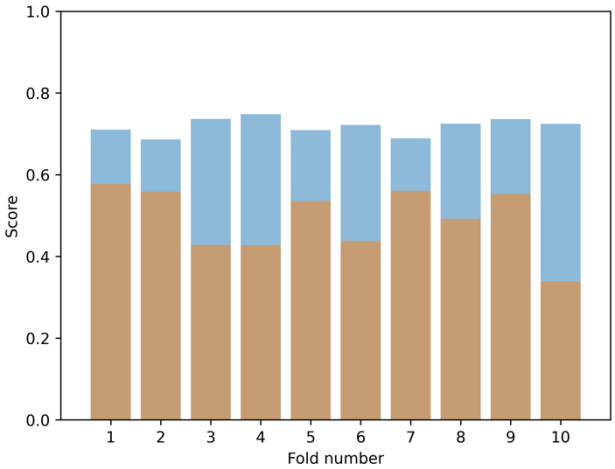
F (RF)



F (MLP)



Cl (RF)



Cl (MLP)

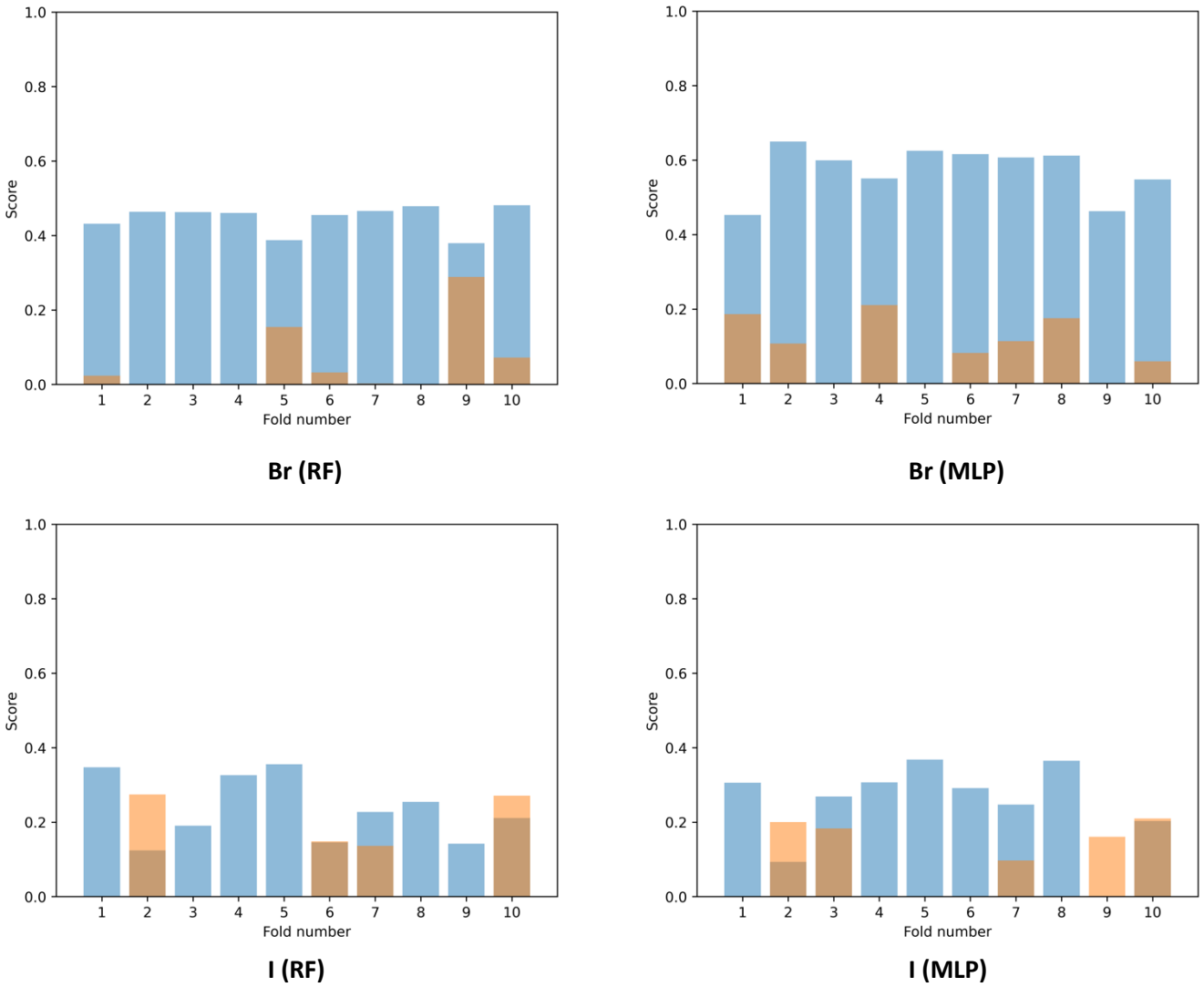
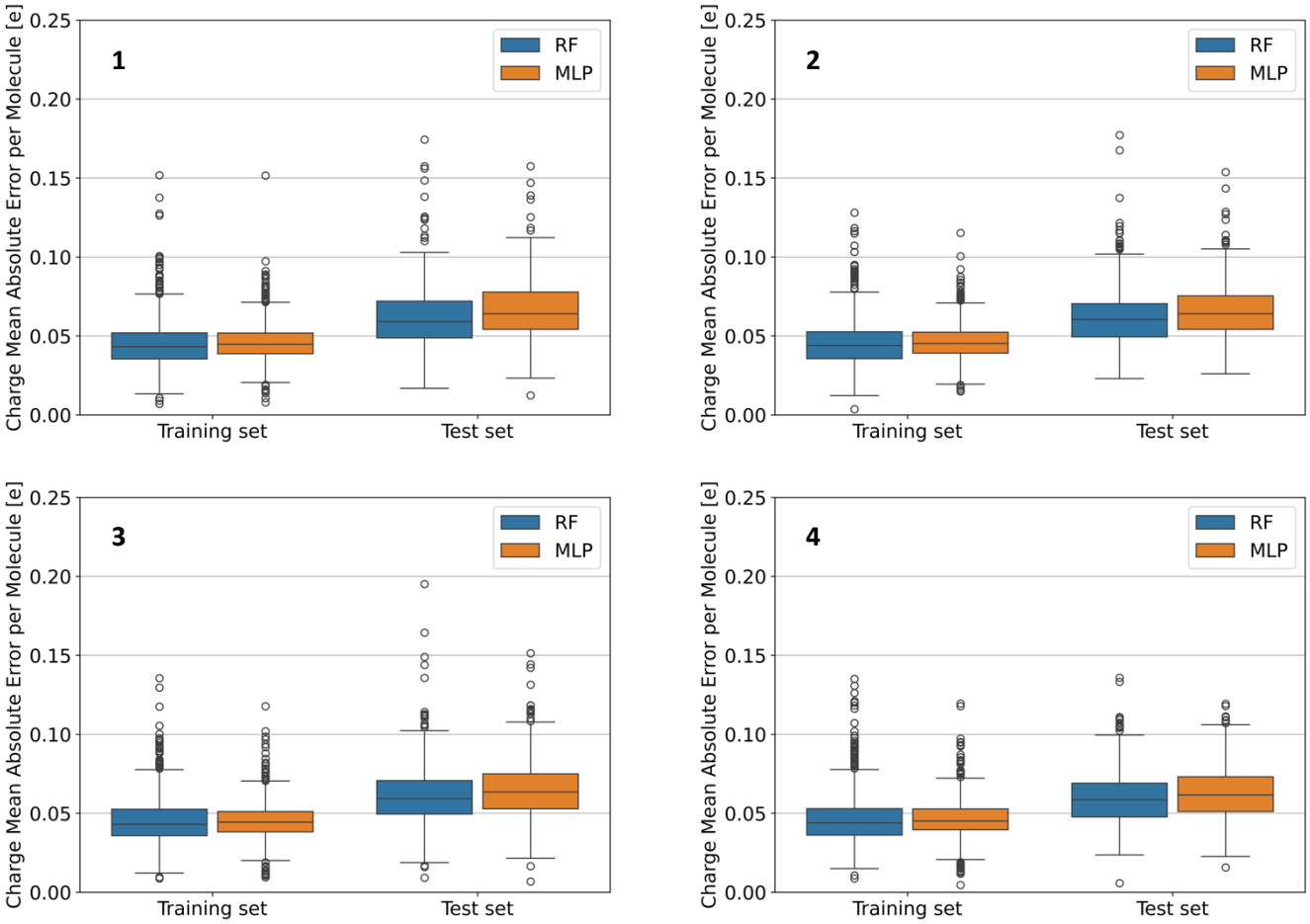


Figure S3 10-fold cross-validation score (R^2) of the RF and MLP models on the training (blue) and test (orange) sets, calculated for atoms of each chemical element.

S3. Cluster-averaged MAE

Table S2 Cluster-averaged MAE (median) for predicted partial atomic charges across all atoms and outlier molecules within each cluster for 10 folds.

Fold	MAE RF Train	MAE MLP Train	MAE RF Test	MAE MLP Test	Outliers RF Train	Outliers MLP Train	Outliers RF Test	Outliers MLP Test
1	0.043	0.045	0.059	0.064	26	9	71	102
2	0.044	0.045	0.061	0.064	28	12	68	85
3	0.043	0.044	0.059	0.064	27	13	54	79
4	0.044	0.045	0.059	0.062	31	9	71	85
5	0.043	0.045	0.058	0.062	28	14	56	73
6	0.043	0.044	0.060	0.065	30	7	61	61
7	0.043	0.046	0.059	0.063	24	12	69	70
8	0.044	0.045	0.058	0.063	23	10	65	65
9	0.044	0.045	0.059	0.063	25	9	79	80
10	0.044	0.045	0.058	0.061	29	9	64	71



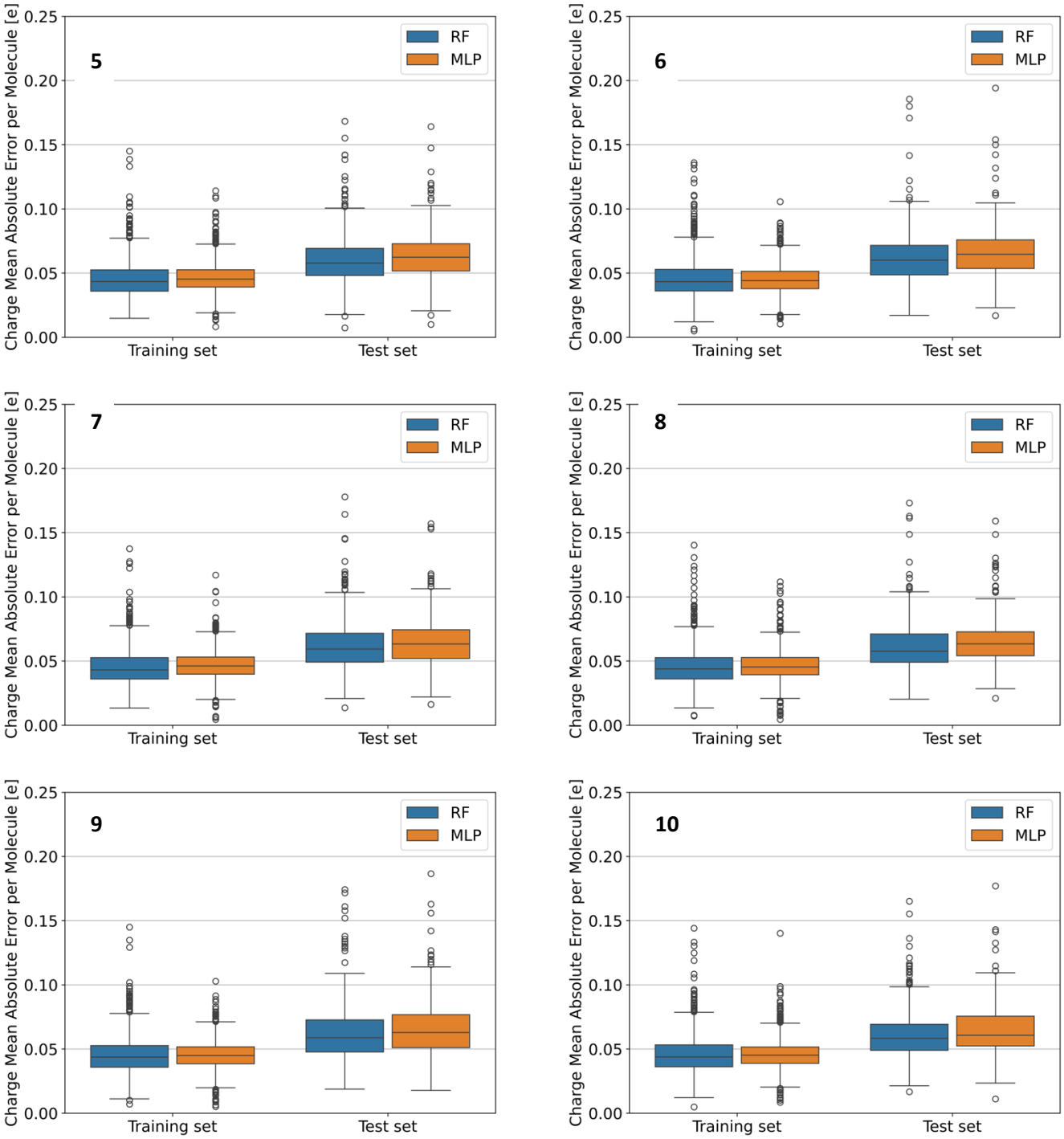


Figure S4 Distributions of cluster-averaged MAE for predicted partial atomic charges across all atoms (single fold) within each cluster for 10 folds. Blue boxes: RF. Orange boxes: MLP.

Data Availability Statement. Users can access both input data and the trained models free of charge at the repository from https://github.com/AIAidedDrugDesign/ChargeBenchmark, accessed on 24 December 2025.

References

S1 PDBbind Database, version 2020; https://www.pdbbind.org.cn.

S2 A. A. Granovsky. Firefly, version 8.2.0, 2016; http://classic.chem.msu.su/gran/firefly/index.html.

S3 D. A. Case, H. M. Aktulga, K. Belfon, D. S. Cerutti, G. A. Cisneros, V. W. D. Cruzeiro, N. Forouzesh, T. J. Giese, A. W. Götz, H. Gohlke, S. Izadi, K. Kasavajhala, M. C. Kaymak, E. King, T. Kurtzman, T.S. Lee, P. Li, J. Liu, T. Luchko, R. Luo, M. Manathunga, M. R. Machado, H. M. Nguyen, K. A. O'Hearn, A. V. Onufriev, F. Pan, S. Pantano, R. Qi, A. Rahnamoun, A. Risheh, S. Schott-Verdugo, A. Shajan, J. Swails, J. Wang, H. Wei, X. Wu, Y. Wu, S. Zhang, S. Zhao, Q. Zhu, T. E. Cheatham 3rd, D. R. Roe, A. Roitberg, C. Simmerling, D. M. York, M. C. Nagan and K. M. Merz Jr., *J. Chem. Inf. Model.*, 2023, **63**, 6183; https://doi.org/10.1021/acs.jcim.3c01153.

S4 R. E. Carhart, D. H. Smith, R. Venkataraghavan, *J. Chem. Inf. Comput. Sci.*, 1985, **25**, 64; https://doi.org/10.1021/ci00046a002.

S5 RDKit: Open-source cheminformatics, version 2024.09.6; https://www.rdkit.org; https://doi.org/10.5281/zenodo.14943932.

S6 Y. Martin, R. Abagyan, G. Ferenczy, V. Gillet, T. Oprea, J. Ulander, D. Winkler and N. Zefirov, *Pure Appl. Chem.*, 2016, **88**, 239; https://doi.org/10.1515/pac-2012-1204.

S7 D. Butina, *J. Chem. Inf. Comput. Sci.*, 1999, **39**, 747; https://doi.org/10.1021/ci9803381.

S8 P. Bleiziffer, K. Schaller and S. Riniker, *J. Chem. Inf. Model.*, 2018, **58**, 579; https://doi.org/10.1021/acs.jcim.7b00663.

S9 Scikit-Learn − Machine Learning in Python; https://scikit-learn.org.

S10 N. Ketkar and J. Moolayil, Introduction to PyTorch. In Deep Learning with Python: Learn Best Practices of Deep Learning Models with PyTorch, Apress, Berkeley, CA, 2021; https://doi.org/10.1007/978-1-4842-5364-9_2.