Supplementary Information (SI) for Physical Chemistry Chemical Physics.
This journal is © the Owner Societies 2025

Supplementary information for Examining proton conductivity of metal-organic frameworks by means of machine learning

Ivan V. Dudakov
MSU Institute for Artificial Intelligence
Lomonosov Moscow State University
Moscow 119192, Russia

Sergei A. Savelev
Department of Chemistry
Lomonosov Moscow State University
Moscow 119991, Russia

Iurii M. Nevolin
Frumkin Institute of Physical Chemistry
and Electrochemistry
Russian Academy of Sciences
Moscow 119071, Russia

Artem A. Mitrofanov
Department of Chemistry
Lomonosov Moscow State University
Moscow 119991, Russia

Vadim V. Korolev*
MSU Institute for Artificial Intelligence
Lomonosov Moscow State University
Moscow 119192, Russia

Yulia G. Gorbunova
Frumkin Institute of Physical Chemistry
and Electrochemistry
Russian Academy of Sciences
Moscow 119071, Russia

* *Email address*: V.Korolev@iai.msu.ru

1 Methods

1.1 The curated database of proton-conducting metal–organic frameworks (MOFs)

Information on proton conductivity of MOFs was collected from scientific publications. The Google Scholar search engine was utilized with the following query: `("metal organic framework" OR "MOF" OR "coordination polymer") AND ("proton conductivity" OR "proton conductor")`. Topical reviews were examined as well. We manually collected proton conductivity values from tables and plots containing temperature dependences of the quantity; the `WebPlotDigitizer`[1] tool was used to extract numerical data. To ensure the reliability of the structural data, we chose only MOFs that were deposited in the Cambridge Structural Database[2] (available as of June 2023). The downloaded crystal structures were curated by elimination of solvent molecules/ions and via resolving of disorder. Connected atomic components in a crystal graph were determined using a simple distance-based criterion (the sum of the corresponding covalent radii[3] and a tolerance distance of 0.5 Å): components with less than 10 atoms were removed from the crystal structure. Then, we reassigned site occupancy values as follows: `OCC = 1 IF OCC_INIT > 0.5 ELSE OCC = 0`. Hydrogen atoms missing in organic linkers were added by means of the HSite module integrated into the `ToposPro` package[4]. In total, 219 distinct MOFs were included in the database. Each entry in the database contains a curated crystal structure and physicochemical properties, including temperature, relative humidity, and proton conductivity. Besides, solvent molecules were taken into account by an analysis of the acid dissociation constant (in the logarithmic form: pK_a) and of the degree of protonation. Activation energies of proton transfer were computed from the temperature dependencies of proton conductivity, under the assumption of the Arrhenius behavior.

1.2 Neural networks for predicting proton-conducting properties

MOFTransformer[5], a multimodal Transformer encoder pretrained on 1 million hypothetical MOFs, was chosen as a basic model for predicting proton conductivity of MOFs. The original model, which included two modalities (a crystal graph and energy grids), was modified to incorporate task-specific global-state features. Temperature, relative humidity, the acid dissociation constant, and the degree of protonation of a guest protic carrier were encoded via special tokens (as was done to normalized cell volume `[VOL]` in the original architecture) and were added in this form at the final position of input embedding, which consisted of the aforementioned local and global-state features.

For consistent estimation of model performance, we applied the stratified k-fold iterator with nonoverlapping groups[6] corresponding to distinct chemical compounds. In other words, the data points that share a common MOF structure were assigned together to either the training or test subset. Fivefold cross-validation was conducted through training of 10-model ensembles. In our experiments, we evaluated two pretraining strategies. In the first one, the models trained on structural features (MOF topology prediction, void fraction prediction, and metal cluster–organic linker classification) as provided by authors were used for further (pre)training. In the second strategy, we trained models on band gap values taken from the Quantum-MOF database[7]. In particular, a set of 20,374 values calculated using the Perdew–Burke–Ernzerhof exchange-correlation functional[8] was split into training (80%), validation (10%), and test (10%) subsets. The obtained metrics for band gap prediction—MAE of 0.30 eV and RMSE of 0.45 eV—were lower in comparison with the previous results; we explain this outcome by a modification of model architecture: global-state embeddings were not trained at this stage and were added to make the model compatible with the subsequent fine-tuning procedure.

The conventional mean square error served as a loss function for band gap predictors, whereas the following uncertainty-aware expression[9] was utilized to model proton conductivity as a variable, μ, with associated heteroscedastic variance σ^2:

$$\mathcal{L} = \frac{1}{N}\sum_{i}^{N}\frac{1}{2}\exp\left(-s_i\right)\left\|y_i-\mu_i\right\|^2+\frac{s_i}{2} \tag{1}$$

where N is the number of data points, y_i and μ_i are ground-truth and predicted values of the ith example, and s_i is log-variance of *aleatoric* uncertainty (i.e., $s_i = \log\sigma_i^2$). Total uncertainty was approximated as a sum of the aforementioned data-related uncertainty and *epistemic* uncertainty calculated as a variance of ensemble predictions. The combination of the sigmoid layer and cross entropy served as a loss function in the case of classifiers trained on binarized activation energy values (`BCEWithLogitsLoss` as implemented in the `PyTorch` library[10]). Models for predicting the band gap were trained at a learning rate of 10^{-4} and a batch size of 16. Models for predicting proton conductivity were trained at a learning rate of 3×10^{-4} and

a batch size of 32. Models for classifying a proton conduction mechanism were trained at a learning rate of 3×10^{-4} and a batch size of 8. Optimal values of the above hyperparameters were identified by the grid search method. All models were trained for 20 epochs.

1.3 High-throughput screening of the Cambridge Structural Database

We assessed the density functional theory–ready dataset (prepared in accordance with a procedure described in ref.[7]) as a starting point to perform high-throughput screening for the discovery of superprotonic conductors. In short, the "non-disordered" CSD MOF subset was curated by elimination of crystal structures with the following drawbacks: CSD-flagged errors, the absence of hydrogen and carbon atoms, any disorder, unbounded atoms, ions in a crystal graph, atoms with missing 3D coordinates, or interatomic distances less than 0.75 Å. The QMOF-42349 dataset was then preprocessed via removal of unbound solvent molecules included in the CSD list of solvents[2]; atomic connectivity was calculated by the `MOFfragmentor` routine[11]. Then, duplicate crystal structures were detected with the help of `StructureMatcher` (as implemented in the `pymatgen` package[12]) and were excluded from the consideration. The final curated set containing 32,929 structures was utilized as input data for predicting proton conductivity at 80 °C and 98% relative humidity. We processed outputs taken from five 10-model ensembles that led to total uncertainty $\sqrt{\sigma^2_{aleatoric} + \sigma^2_{epistemic}} < 0.2$ lg(S cm^{-1}), resulting in a subset containing 7,739 structures. Two-dimensional projection of chemical space was generated by the uniform manifold approximation and projection (`UMAP`) algorithm[13], which was applied to materials embeddings produced by the MOFTransformer model.

2 Metrics

Classification metrics can be natively defined within confusion matrix $\mathcal{C}$ for $\mathcal{K}$ classes. Its element $\mathcal{C}_{ij}$ contains the number of entities of class i that were assigned to class j by a predictive model.

$$\text{overall accuracy} = \frac{c}{s} \tag{2}$$

$$\text{recall}_k = \frac{\mathcal{C}_{kk}}{t_k} \tag{3}$$

$$\text{balanced accuracy} = \frac{1}{\mathcal{K}} \sum_k^{\mathcal{K}} \text{recall}_k \tag{4}$$

where $c = \sum_k^{\mathcal{K}} \mathcal{C}_{kk}$, $s = \sum_i^{\mathcal{K}} \sum_j^{\mathcal{K}} \mathcal{C}_{ij}$, and $t_k = \sum_i^{\mathcal{K}} \mathcal{C}_{ki}$.

The receiver-operating characteristic curve is a two-dimensional curve in which the *true positive rate* is plotted against the *false positive rate*. The area under the receiver-operating characteristic curve (ROC AUC) can serve as a measure of (binary) classifier performance[14].

If a set of $\mathcal{N}$ test samples is used, then the following regression metrics can be computed:

$$\text{mean absolute error (MAE)} = \frac{1}{\mathcal{N}} \sum_i^{\mathcal{N}} |y_i - \hat{y}_i| \tag{5}$$

$$\text{root mean square error (RMSE)} = \frac{1}{\mathcal{N}} \sum_i^{\mathcal{N}} (y_i - \hat{y}_i)^2 \tag{6}$$

where y_i and $\hat{y}_i$ are respectively the true value and predicted value of the ith sample.

The Spearman correlation coefficient can be computed as follows, assuming that all ranks of the observations are distinct integers:

$$r_s = 1 - \frac{6 \sum_i^{\mathcal{N}} d_i^2}{\mathcal{N}(\mathcal{N}^2 - 1)} \tag{7}$$

where $d_i = R[y_i] - R[\hat{y}_i]$ is the difference between ranks of the true value and predicted value of the ith sample.

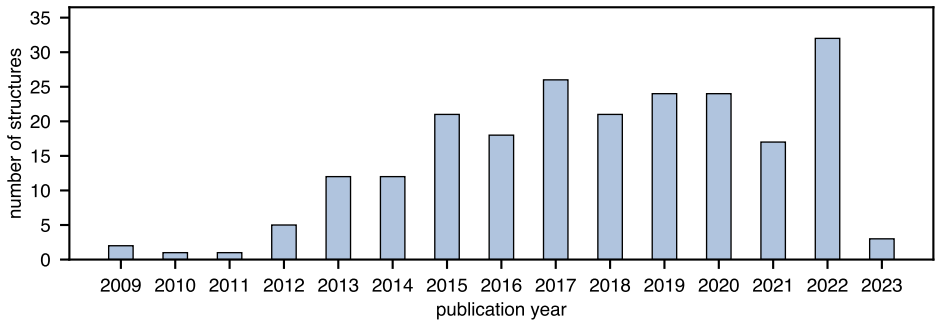

Figure S1. The expansion of the dataset with time. The distribution (by year) of CSD structures having associated proton conduction data.

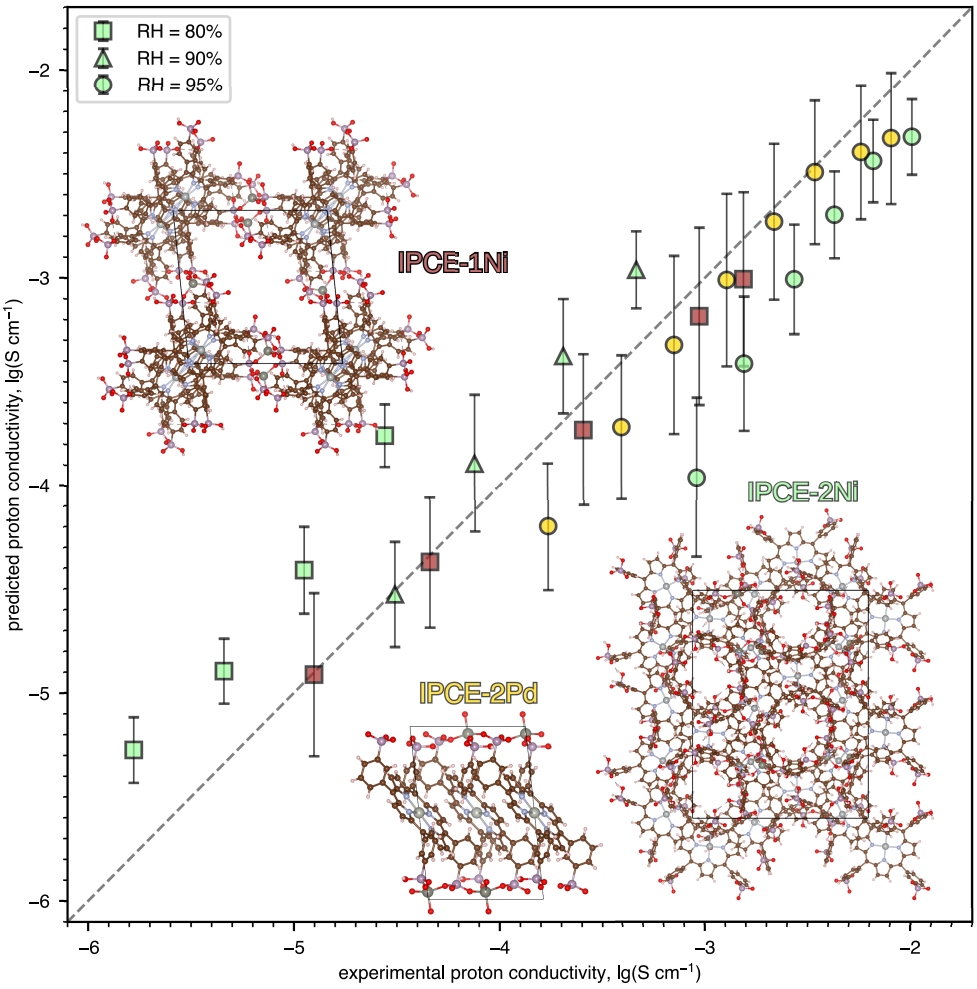

Figure S2. Predictive performance in the case of a specific class of MOFs. A scatter plot of experimental and predicted proton conductivity values; three porphyrinylphosphonate-based MOFs are examined here.

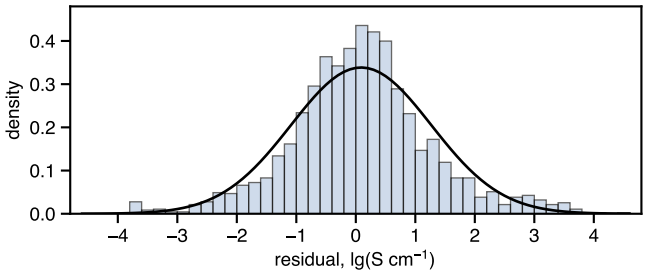

Figure S3. The residual distribution for the test dataset.

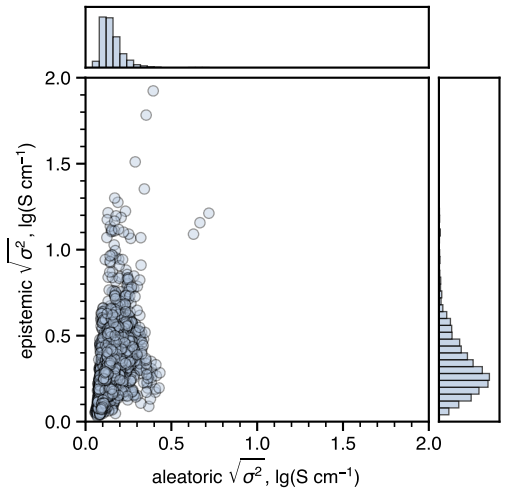

Figure S4. Aleatoric and epistemic components of expected error for the test dataset.

References

[1] Ankit Rohatgi. Webplotdigitizer.

[2] Peyman Z Moghadam, Aurelia Li, Seth B Wiggin, Andi Tao, Andrew GP Maloney, Peter A Wood, Suzanna C Ward, and David Fairen-Jimenez. Development of a cambridge structural database subset: a collection of metal–organic frameworks for past, present, and future. *Chemistry of Materials*, 29(7):2618–2625, 2017.

[3] Beatriz Cordero, Verónica Gómez, Ana E Platero-Prats, Marc Revés, Jorge Echeverría, Eduard Cremades, Flavia Barragán, and Santiago Alvarez. Covalent radii revisited. *Dalton Transactions*, (21):2832–2838, 2008.

[4] Vladislav A Blatov, Alexander P Shevchenko, and Davide M Proserpio. Applied topological analysis of crystal structures with the program package topospro. *Crystal Growth & Design*, 14(7):3576–3586, 2014.

[5] Yeonghun Kang, Hyunsoo Park, Berend Smit, and Jihan Kim. A multi-modal pre-training transformer for universal transfer learning in metal–organic frameworks. *Nature Machine Intelligence*, 5(3):309–318, 2023.

[6] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830, 2011.

[7] Andrew S Rosen, Shaelyn M Iyer, Debmalya Ray, Zhenpeng Yao, Alan Aspuru-Guzik, Laura Gagliardi, Justin M Notestein, and Randall Q Snurr. Machine learning the quantum-chemical properties of metal–organic frameworks for accelerated materials discovery. *Matter*, 4(5):1578–1597, 2021.

[8] John P Perdew, Kieron Burke, and Matthias Ernzerhof. Generalized gradient approximation made simple. *Physical review letters*, 77(18):3865, 1996.

[9] David A Nix and Andreas S Weigend. Estimating the mean and variance of the target probability distribution. In *Proceedings of 1994 ieee international conference on neural networks (ICNN'94)*, volume 1, pages 55–60. IEEE, 1994.

[10] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

[11] Kevin Maik Jablonka, Andrew S Rosen, Aditi S Krishnapriyan, and Berend Smit. An ecosystem for digital reticular chemistry. *ACS Central Science*, 9(4):2374–7951, 2023.

[12] Shyue Ping Ong, William Davidson Richards, Anubhav Jain, Geoffroy Hautier, Michael Kocher, Shreyas Cholia, Dan Gunter, Vincent L Chevrier, Kristin A Persson, and Gerbrand Ceder. Python materials genomics (pymatgen): A robust, open-source python library for materials analysis. *Computational Materials Science*, 68:314–319, 2013.

[13] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

[14] Tom Fawcett. An introduction to roc analysis. *Pattern recognition letters*, 27(8):861–874, 2006.