Available online at www.sciencedirect.com

ScienceDirect

Mendeleev Commun., 2024, **34**, 776–779

Mendeleev Communications

Towards accurate machine learning predictions of properties of the P–O bond cleaving in ATP upon enzymatic hydrolysis

Igor V. Polyakov,[a] Kirill D. Miroshnichenko,[a] Tatiana I. Mulashkina,[a] Alexander A. Moskovsky,[a] Ekaterina I. Marchenko[b] and Maria G. Khrenova*[a,c]

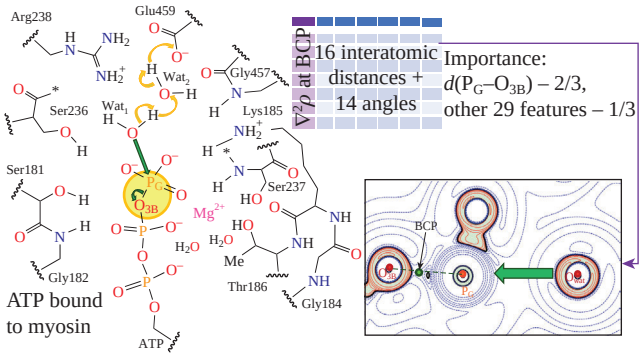
[a] *Department of Chemistry, M. V. Lomonosov Moscow State University, 119991 Moscow, Russian Federation. E-mail: khrenovamg@my.msu.ru*

[b] *Department of Materials Science, M. V. Lomonosov Moscow State University, 119991 Moscow, Russian Federation*

[c] *Bach Institute of Biochemistry, Federal Research Centre 'Fundamentals of Biotechnology' of the Russian Academy of Sciences, 119071 Moscow, Russian Federation*

DOI: 10.1016/j.mencom.2024.10.003

Molecular dynamic simulations using QM/MM potentials are performed for the enzyme–substrate complex of adenosine triphosphate (ATP) with the motor protein myosin. Machine learning methods are applied to a dataset consisting of the geometry parameters of the active site in the enzyme–substrate complex to predict the Laplacian of electron density at the bond critical point of the P_G–O_{3B} bond being broken in ATP. Using a gradient boosting machine learning model, a mean absolute error of 0.01 a.u. and an R^2 score of 0.99 are achieved, and it is found that the P_G–O_{3B} bond length is the most important feature, contributing 2/3, while other geometry features contribute 1/3.



Keywords: machine learning, myosin, ATP hydrolysis, QM/MM molecular dynamics, Laplacian of electron density.

Myosin is an ATPase that acts as a motor in a variety of activities, including muscle contraction.[1–3] It hydrolyzes adenosine triphosphate (ATP) to adenosine diphosphate (ADP) and inorganic phosphate. This enzyme has been extensively studied by both experimental and theoretical methods, including the effects of disease-associated amino acid substitutions.[4,5] The chemical reaction most likely occurs *via* a dissociative mechanism, *i.e.*, the P_G–O_{3B} bond is broken before the covalent bond is formed between the P_G atom and the oxygen of the catalytic water molecule (Figure 1). This reaction proceeds *via* a two-water mechanism.[4] Nucleophilic attack of P_G by a water molecule initiates the reaction that involves cleavage of the P_G–O_{3B} bond (Figure 1, green arrows) and proton transfer from the catalytic water Wat1 to Glu459 along a hydrogen bond network comprising the auxiliary water molecule Wat2 (Figure 1, yellow arrows). ATP forms coordination bonds with the Mg^{2+} cation of the active site and hydrogen bonds with neighboring amino acid residues. These interactions promote proper binding of ATP to the active site and likely activate the ATP molecule for the chemical reaction.

GTPases share the same dissociative mechanism, and it has been shown that upon binding the enzyme active site activates a substrate, namely weakening the P_G–O_{3B} bond strength during the formation of the enzyme–substrate (ES) complex.[6] This can be easily visualized by plotting the calculated Laplacian of electron density ($\nabla^2\rho$) in the plane formed by the P_G–O_{3B} bond and the nucleophilic oxygen atom of the catalytic water molecule (Figure 2).[6] The Laplacian of electron density discriminates spatial regions of local electron density concentration with $\nabla^2\rho(\boldsymbol{r}) < 0$ and electron density depletion with $\nabla^2\rho(\boldsymbol{r}) > 0$. In the non-activated states, electron density concentration is observed in the P_G–O_{3B} bond region. After activation, this bond is characterized by electron density depletion that means its weakening. This observation seems reasonable since the P_G–O_{3B} bond should be broken in the very beginning of a chemical reaction and should therefore be prepared for this. Information about the electron structure of a molecule can be obtained not only from the spatial distribution

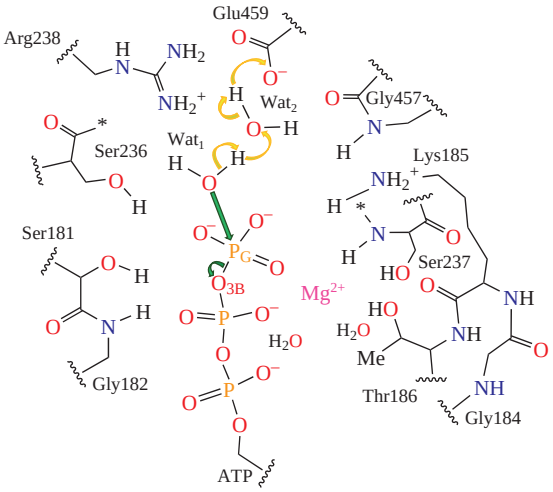


Figure 1 The active site of myosin in complex with ATP included in the QM part of the QM/MM simulations. Yellow arrows show the proton transfer pathway during the reaction. Green arrows correspond to the nucleophilic attack and cleavage of the P_G–O_{3B} bond. The asterisk indicates a fictitious disconnection of the Ser236–Ser237 peptide bond, introduced for clarity of the figure.

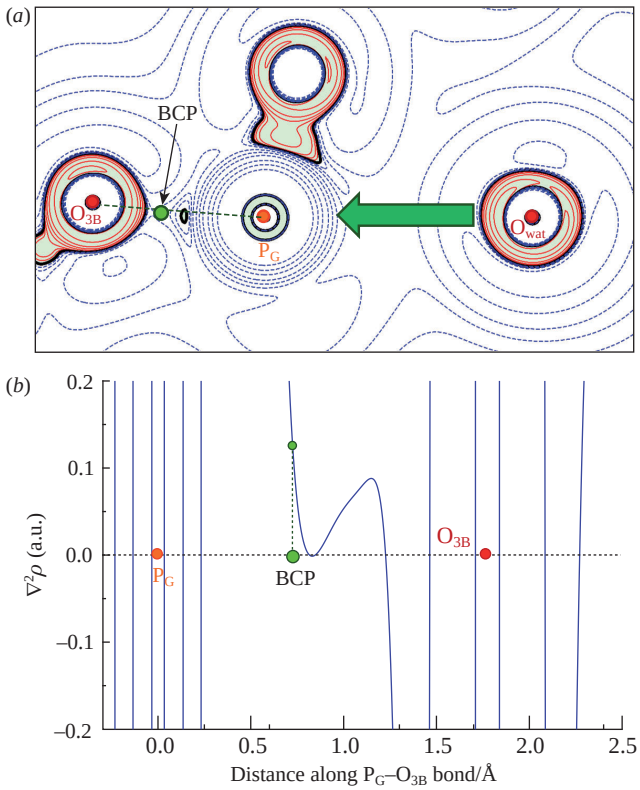


Figure 2 (*a*) 2D map of the Laplacian of electron density in the plane of the P_G–O_{3B} bond being broken in ATP and the nucleophilic oxygen atom O_{wat} of the water molecule, calculated on the QM/MM MD frame of the myosin–ATP ES complex. Positive and negative isovalues are shown in blue dashed and red solid lines, respectively. The arrow shows the direction of nucleophilic attack. (*b*) Laplacian of electron density along the P_G–O_{3B} bond. The green dot is BCP of the P_G–O_{3B} bond. Contour lines correspond to ±(2 or 4 or 8) × 10^n a.u., $-2 \leqslant n \leqslant 1$, blue dashed contour lines indicate regions of electron density depletion [$\nabla^2\rho(\boldsymbol{r}) > 0$], red solid lines show electron density concentration [$\nabla^2\rho(\boldsymbol{r}) < 0$], and black solid lines correspond to $\nabla^2\rho(\boldsymbol{r}) = 0$. The area with $\nabla^2\rho(\boldsymbol{r}) < 0$ is colored light green.

of the corresponding descriptors but also from their values at the critical points of the electron density. These ideas are developed within the quantum theory of atoms in molecules (QTAIM).[7] Of particular interest are the bond critical points (BCPs), which are the minima of the electron density in one spatial direction and the maxima in the other two. The BCP is located on the bond path that connects the two interacting atoms, and the electron density descriptors characterize and classify this interaction. The

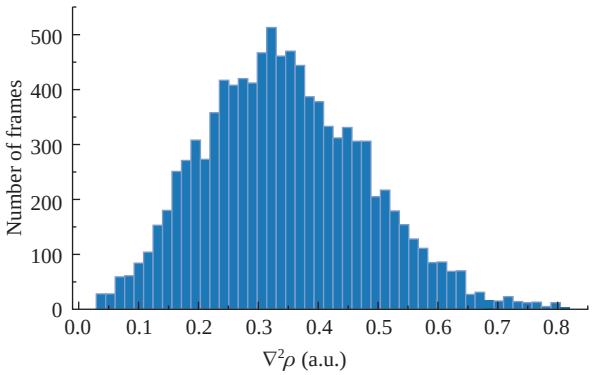


Figure 3 Distribution of the Laplacian of electron density at the BCP of the P_G–O_{3B} bond being broken, calculated on 10 000 QM/MM MD frames of the myosin–ATP ES complex trajectory.

Laplacian of electron density is also an informative descriptor when calculated at the BCP (Figure 2).[8,9]

The aim of this work is to use machine learning (ML) methods to establish a relationship between the features of the geometry of the myosin active site in the ES complex and the Laplacian of electron density calculated at the BCP of the P_G–O_{3B} bond being broken.

The data set was obtained from the molecular dynamics (MD) trajectory of the ES complex calculated using combined quantum mechanics/molecular mechanics (QM/MM) potentials.† The hydrolysis reaction mechanism and the role of amino acid residues in the active site were considered to select a set of features for ML (Table S1, see Online Supplementary Materials). These are the nucleophilic attack and breaking bond distances, proton transfer path distances, lengths of other covalent bonds with the phosphorus atom and interatomic distances between ATP and neighboring groups, a total of 18 features [Figure S1(*a*), see Online Supplementary Materials]. In addition, 15 angles in the active site were selected [Figure S1(*b*)]. This set of 33 features was extracted from each QM/MM MD frame.

The target values are described by a single normal distribution (Figure 3). The geometry features are distributed differently (Figures S2 and S3). Some of these plots represent a single normal distribution, such as Dist6 (P_G–O_{3B} bond), while many others are composite and mixed distributions, indicating the presence of populations with different states.

The feature covariance matrix was calculated to depict the feature dependencies (Figure S4). Strong correlations were observed for

† *Computational protocol*. A full-atom model of the ES complex of myosin and ATP was constructed based on the crystal structure of the myosin complex with ADP and vanadate anion (PDB ID: 1VOM).[10] Hydrogen atoms were added using the Reduce program[11] to reproduce neutral pH. The ES complex was solvated in a rectangular water box and neutralized. Classical MD simulation for 2 ns was performed with fixed coordinates of the protein and ATP to relax the solvation shell. MD simulations were performed using the NAMD software package.[12] Then, a 5 ns MD trajectory was calculated without additional restraints to relax the protein. The CHARMM[13] force field was used to describe the enzyme and ATP molecule, and TIP3P[14] was used for water molecules. The preparation of the full-atom model as well as visualization and analysis of the structures were carried out using the VMD program.[15]

The system preparation was followed by a 10 ps production QM/MM MD run. The QM subsystem (see Figure 1) included the side chains of Ser181, Gly182, Lys185, Thr186, Ser236, Ser237, Arg238, Ser456, Gly457 and Glu459, the phosphate groups of ATP, the Mg^{2+} cation and four water molecules, for a total of 140 atoms with a net charge of –1. The QM part was solved using the Kohn–Sham DFT approach with the PBE0 hybrid functional[16] empirically corrected for D3 dispersion interactions[17] and the 6-31G** basis set. The energies and forces in the QM subsystem were calculated in the TeraChem program;[18] the MM part and the MD step were calculated in the NAMD program using a custom interface.[19] The QM method was chosen in accordance with the reference QM/MM studies of enzymatic reaction mechanisms, which were summarized in a recent review.[20] All MD calculations with classical and combined potentials were performed in the NPT ensemble at p = 1 atm and T = 300 K with an integration step of 1 fs using a Nosé–Hoover barostat[21] and a Langevin thermostat.[22]

QTAIM theory[7] was utilized to determine the BCPs using the Multiwfn program.[23] BCPs are saddle points of the electron density and characterize the interacting atoms. The Laplacian of electron density $\nabla^2\rho$ was calculated on different MD frames for the BCP of the P_G–O_{3B} bond being broken. The dataset included 10^4 samples, each consisting of $\nabla^2\rho$ and 33 features. These data were combined into the Pandas DataFrame.

The scikit-learn library,[24] LightGBM library[25] and TabNetRegressor from pytorch-tabnet[26] were used to train the ML models and make predictions. The full dataset was split into training–validation and test subsets in a 4 : 1 ratio. Four-fold cross-validation and GridSearchCV were utilized for hyperparameter optimization. Linear models were employed using a pipeline with StandardScaler to standardize features before implementing regression. Mean absolute error (MAE) was used as the loss metric for validation.

The ML protocol code and dataset are deposited on Zenodo.[27]

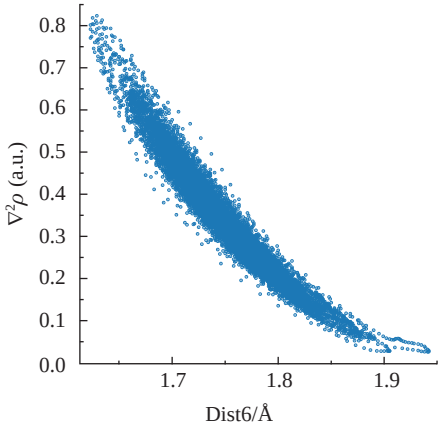


Figure 4 Laplacian of electron density at the BCP of the P_G–O_{3B} bond being broken as a function of its length, calculated on 10 000 QM/MM MD frames of the ES complex trajectory.

Ang1 with Dist2/11/12 and Ang9/10, Dist12 with Dist2 and Ang1/9/10, Ang8 with Dist11 and Ang11 with Ang13. Features Ang1 and Dist12 were excluded from the dataset to avoid multicollinearity for proper construction of linear models (Figure S5).

Figure 4 demonstrates the dependence of the Laplacian of electron density at the BCP on Dist6 in the form of a banana-shaped plot, which in principle can be approximated by a linear function. Other features or their linear combination cannot be easily interpreted, since the plot is a scattered blob. This means that there is a multidimensional banana-shaped plot (Figure S6) of the target value depending on the selected features.

Simple linear regression yielded MAE = 0.0256 a.u. and R^2 = 0.939, which is already an almost perfect prediction. The regression coefficients are presented in Table S2. The major contribution (~68%) comes from the breakable P_G–O_{3B} bond (Dist6). Other important contributions are made by Ang13 (~4%), Ang12 (~4%), Dist3 (~3%), Ang4 (~2%), Dist9 (~2%), Dist13 (~2%), Dist15 (~2%), Ang11 (~2%) and Ang15 (~2%). All other 19 features together contribute about 9%. Both linear regression models with regularization, Lasso (L1) and Ridge (L2), did not reduce the MAE loss metric. It is expected that the Laplacian value at BCP is mainly affected by the P_G–O_{3B} bond length (Dist6) in the linear model, but the contributions of other features still seem to be significant. We could point out that the position of the magnesium cation relative to the phosphate groups of ATP (Ang13, Ang12, Dist9, Dist15 and Ang11) and the direction of water attack (Dist3 and Ang15) contribute ~14% and ~5%, respectively. However, if we drop all features except Dist6, the MAE of the resulting linear regression is 0.0263 a.u. (R^2 = 0.934). Basically,

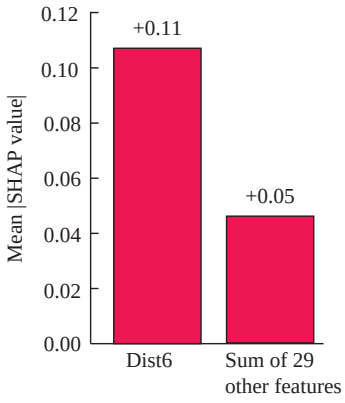


Figure 5 Shapley values calculated for the LGBM model. Dist6 corresponds to the P_G–O_{3B} bond being broken.

Dist6 is the only meaningful feature for the linear models to extract the underlying relationship, while the others can be attributed to predicting noise.

To further improve the prediction quality and determine the importance of different features, another class of models, gradient boosted decision trees, was considered. The best MAE of 0.01 a.u. in four-fold cross-validation was obtained using LightGBM with a depth of six and a learning rate of 0.34. To test whether there is a significant difference in feature importance between the linear regression and LGBM models, Shapley values were calculated using the SHAP (SHapley Additive exPlanations) library.[28] The mean SHAP values imply that the contribution of Dist6 is dominant, as in the case of the linear regression model, while the individual contributions of all other features are miniscule, but their combined importance accounts for approximately a third of the model prediction (Figure 5). The SHAP beeswarm plot hints that lower Dist6 values correspond to higher SHAP values and are thus more important in the model predictions (Figure S7).

The LGBM model trained with only the Dist6 feature yields MAE = 0.022, which is much worse than MAE = 0.01 for the full set of features. Unlike linear models, gradient boosting is able to extract valuable information from additional features.

Next, the TabNet artificial neural network model[26] was tested. The computed mean cross-validation MAE loss for the TabNet regression model was ~0.02 a.u., which is better than that for the linear regression but worse than that for the tuned LGBM model, so the LGBM model was chosen for testing.

The LGBM regressor MAE loss on the test dataset is 0.009 a.u. (R^2 = 0.992), which is even lower than the validation result. In comparison, the dummy mean prediction yielded a test MAE loss of 0.106 a.u. and R^2 = 0.

In conclusion, it was demonstrated that the electron density property of the cleaving P_G–O_{3B} bond, determined by the Laplacian of electron density at the corresponding BCP, depends on a set of geometry parameters. Different ML models were applied to determine the contributions of individual geometry features to $\nabla^2\rho$. The lengths of the P_G–O_{3B} bond being broken contribute about 2/3, and other features are important in combination. Notable features are the nucleophilic attack distance and features determining the exact position of the Mg^{2+} cation coordinating the ATP phosphates.

This work was supported by the Interdisciplinary Scientific and Educational School of Moscow State University ‘Brain, cognitive systems, artificial intelligence” (project no. 23-Sh03-04). The research was carried out using equipment of the shared research facilities of the HPC computing resources at Lomonosov Moscow State University.

Online Supplementary Materials

Supplementary data associated with this article can be found in the online version at doi: 10.1016/j.mencom.2024.10.003.

References

1 M. A. Hartman and J. A. Spudich, *J. Cell Sci.*, 2012, **125**, 1627; https://doi.org/10.1242/jcs.094300.
2 A. B. Kolomeisky, *J. Phys.: Condens. Matter*, 2013, **25**, 463101; https://doi.org/10.1088/0953-8984/25/46/463101.
3 H. L. Sweeney and E. L. F. Holzbaur, *Cold Spring Harbor Perspect. Biol.*, 2018, **10**, a021931; https://doi.org/10.1101/cshperspect.a021931.
4 M. G. Khrenova, T. I. Mulashkina, R. A. Stepanyuk and A. V. Nemukhin, *Mendeleev Commun.*, 2024, **34**, 1; https://doi.org/10.1016/j.mencom.2024.01.001.
5 A. Chakraborti, J. C. Tardiff and S. D. Schwartz, *J. Phys. Chem. B*, 2024, **128**, 4716; https://doi.org/10.1021/acs.jpcb.4c01601.
6 M. G. Khrenova, B. L. Grigorenko and A. V. Nemukhin, *ACS Catal.*, 2021, **11**, 8985; https://doi.org/10.1021/acscatal.1c00582.

7 R. F. W. Bader, *Atoms in Molecules: A Quantum Theory*, Clarendon Press, Oxford, 1994; https://global.oup.com/academic/product/atoms-in-molecules-9780198558651?cc=us&lang=en&.

8 R. V. Rumyantcev, G. Yu. Zhigulin, G. S. Zabrodina, M. A. Katkova, S. Yu. Ketkov and G. K. Fukin, *Mendeleev Commun.*, 2023, **33**, 41; https://doi.org/10.1016/j.mencom.2023.01.012.

9 T. N. Gribanova, R. M. Minyaev and V. I. Minkin, *Mendeleev Commun.*, 2023, **33**, 302; https://doi.org/10.1016/j.mencom.2023.04.002.

10 C. A. Smith and I. Rayment, *Biochemistry*, 1996, **35**, 5404; https://doi.org/10.1021/bi952633+.

11 J. M. Word, S. C. Lovell, J. S. Richardson and D. C. Richardson, *J. Mol. Biol.*, 1999, **285**, 1735; https://doi.org/10.1006/jmbi.1998.2401.

12 J. C. Phillips, D. J. Hardy, J. D. C. Maia, J. E. Stone, J. V. Ribeiro, R. C. Bernardi, R. Buch, G. Fiorin, J. Hénin, W. Jiang, R. McGreevy, M. C. R. Melo, B. K. Radak, R. D. Skeel, A. Singharoy, Y. Wang, B. Roux, A. Aksimentiev, Z. Luthey-Schulten, L. V. Kalé, K. Schulten, C. Chipot and E. Tajkhorshid, *J. Chem. Phys.*, 2020, **153**, 044130; https://doi.org/10.1063/5.0014475.

13 R. B. Best, X. Zhu, J. Shim, P. E. M. Lopes, J. Mittal, M. Feig and A. D. MacKerell, Jr., *J. Chem. Theory Comput.*, 2012, **8**, 3257; https://doi.org/10.1021/ct300400x.

14 W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey and M. L. Klein, *J. Chem. Phys.*, 1983, **79**, 926; https://doi.org/10.1063/1.445869.

15 W. Humphrey, A. Dalke and K. Schulten, *J. Mol. Graphics*, 1996, **14**, 33; https://doi.org/10.1016/0263-7855(96)00018-5.

16 C. Adamo and V. Barone, *J. Chem. Phys.*, 1999, **110**, 6158; https://doi.org/10.1063/1.478522.

17 S. Grimme, J. Antony, S. Ehrlich and H. Krieg, *J. Chem. Phys.*, 2010, **132**, 154104; https://doi.org/10.1063/1.3382344.

18 S. Seritan, C. Bannwarth, B. S. Fales, E. G. Hohenstein, C. M. Isborn, S. I. L. Kokkila-Schumacher, X. Li, F. Liu, N. Luehr, J. W. Snyder, Jr., C. Song, A. V. Titov, I. S. Ufimtsev, L.-P. Wang and T. J. Martínez, *Wiley Interdiscip. Rev.: Comput. Mol. Sci.*, 2021, **11**, e1494; https://doi.org/10.1002/wcms.1494.

19 M. C. R. Melo, R. C. Bernardi, T. Rudack, M. Scheurer, C. Riplinger, J. C. Phillips, J. D. C. Maia, G. B. Rocha, J. V Ribeiro, J. E. Stone, F. Neese, K. Schulten and Z. Luthey-Schulten, *Nat. Methods*, 2018, **15**, 351; https://doi.org/10.1038/nmeth.4638.

20 M. G. Khrenova, T. I. Mulashkina, A. M. Kulakova, I. V. Polyakov and A. V. Nemukhin, *Moscow Univ. Chem. Bull.*, 2024, **79**, 86; https://doi.org/10.3103/S0027131424700093.

21 G. J. Martyna, M. L. Klein and M. Tuckerman, *J. Chem. Phys.*, 1992, **97**, 2635; https://doi.org/10.1063/1.463940.

22 K. Singer and W. Smith, *Mol. Phys.*, 1988, **64**, 1215; https://doi.org/10.1080/00268978800100823.

23 T. Lu and F. Chen, *J. Comput. Chem.*, 2012, **33**, 580; https://doi.org/10.1002/jcc.22885.

24 F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot and É. Duchesnay, *Journal of Machine Learning Research*, 2011, **12**, 2825; https://jmlr.org/papers/v12/pedregosa11a.html.

25 G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye and T.-Y. Liu, in *Advances in Neural Information Processing Systems 30: 31st Annual Conference on Neural Information Processing Systems (NIPS 2017)*, eds. U. von Luxburg, I. Guyon, S. Bengio, H. Wallach, R. Fergus, S. V. N. Vishwanathan and R. Garnett, Curran Associates, Red Hook, NY, pp. 3147–3155; https://papers.nips.cc/paper_files/paper/2017/hash/6449f44a102fde848669bdd9eb6b76fa-Abstract.html.

26 [dataset] M. Joseph, *PyTorch Tabular: A Framework for Deep Learning with Tabular Data*, arXiv, v1, 2021; https://doi.org/10.48550/arXiv.2104.13638.

27 [dataset] I. Polyakov, *Machine learning predictions of Laplacian value of bond-critical point in enzymatic phosphate hydrolysis*, Zenodo, v1, 2024; https://doi.org/10.5281/zenodo.11395296.

28 S. M. Lundberg, G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal and S.-I. Lee, *Nature Machine Intelligence*, 2020, **2**, 56; https://doi.org/10.1038/s42256-019-0138-9.

Received: 6th June 2024; Com. 24/7527