# Supplementary materials

Table S1. Descriptor packages used to develop machine learning models.

| N | Abbreviation of descriptor package | Number of descriptors after filtering | refs. |
| --- | --- | --- | --- |
| 1 | ALogPS, OESTATE | 263 | [1–4](https://www.zotero.org/google-docs/?tgTIG6) |
| 2 | CDDD | 507 | [5](https://www.zotero.org/google-docs/?5TSs5l) |
| 3 | EPA | 375 | [6](https://www.zotero.org/google-docs/?AtsOKb) |
| 4 | FCFP4 (RDKit) | 534 | [7,8](https://www.zotero.org/google-docs/?Ay8Lzy) |
| 5 | GSFrag | 379 | [9](https://www.zotero.org/google-docs/?Y0i9cg) |
| 6 | JPlogP | 76 | [10](https://www.zotero.org/google-docs/?elVcsE) |
| 7 | MOLD2 | 416 | [11](https://www.zotero.org/google-docs/?fbkXZ5) |
| 8 | QNPR | 603 | [12](https://www.zotero.org/google-docs/?JPBrWG) |
| 9 | EFG/ToxAlerts | 223 | [13](https://www.zotero.org/google-docs/?x6g8WF) |

Table S2. Performances of models developed with different approaches

| Machine learning models based on descriptors | | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Descriptors | DNN | | | | CatBoost | | | |
| Training set, 5CV | | Test sets, QWK score | | Training set, 5CV | | Test sets, QWK score | |
| RMSE | R2 | Leaderboard | private | RMSE | R2 | Leaderboard | private |
| ALogPS, OESTATE | 0.378 | 0.024 | 0.113 | 0.096 | 0.38 | 0.016 | 0.107 | 0.092 |
| CDDD | 0.38 | 0 | 0.109 | 0.080 | 0.381 | 0 | 0.103 | 0.094 |
| EPA | 0.38 | 0.016 | 0.114 | 0.080 | 0.381 | 0.015 | 0.107 | 0.076 |
| FCFP4 | 0.38 | 0 | 0.057 | 0.049 | 0.382 | 0 | 0.063 | 0.050 |
| GSFrag | 0.379 | 0.016 | 0.090 | 0.061 | 0.382 | 0.011 | 0.095 | 0.076 |
| JPlogP | 0.38 | 0.012 | 0.105 | 0.096 | 0.382 | 0.013 | 0.114 | 0.077 |
| MOLD2 | 0.381 | 0 | 0.127 | 0.090 | 0.381 | 0.013 | 0.107 | 0.085 |
| QNPR | 0.382 | 0 | 0.102 | 0.084 | 0.384 | 0 | 0.115 | 0.088 |
| EFG | 0.38 | 0.011 | 0.112 | 0.091 | 0.3821 | 0.0121 | 0.111 | 0.074 |
| Machine learning models based on the representation learning | | | | | | | | |
| Transformer CNN | 0.38 | 0.015 | 0.117 | 0.096 |  | | | |
| AttFP | 0.386 | 0 | 0.042 | 0.038 |
| ChemProp | 0.385 | 0 | 0.067 | 0.037 |
| DimeNetPP\* | 0.383 | 0 | 0 | 0 |
| GINE | 0.384 | 0 | 0.086 | 0.063 |
| GIN | 0.384 | 0 | 0.088 | 0.061 |
| Schnet | 0.385 | 0 | 0.074 | 0.055 |
| Gatv2 | 0.384 | 0 | 0.096 | 0.076 |
| GAT | 0.385 | 0 | 0.121 | 0.090 |
| HamNet | 0.385 | 0 | 0.030 | 0.033 |
| GraphSAGE\* | 0.384 | 0 | 0 | 0 |

1CatBoots model based on EFG descriptors was not included in the final submission by mistake. DimeNetPP and GraphSAGE provided the same predicted values for the test set compounds and could be excluded without affecting consensus model results.

**Instructions to install and use standalone model**

1) Build and start docker as indicated at <https://github.com/openochem/openochem/tree/main/docker>

Check that it is correctly installed and available at

<http://localhost:8080>

<http://localhost:7080/metaserver>

2) Update database using (request the file from the authors)

tar -zxf solubility.tar.gz

docker exec -i ochem-mariadb sh -c 'exec mariadb fragment\_demo' < fragment\_demo.sql

docker exec -i ochem-mariadb sh -c 'exec mariadb --max\_allowed\_packet=1G ochem\_demo' < ochem\_demo.sql

3) Restart docker

docker-compose -f docker/ochem-demo.yml stop

docker-compose -f docker/ochem-demo.yml up

The local version will be available at the same address <http://localhost:8080> as the initial docker. To see models:

a) Select from menu /Database/Articles

c) Click on the Article profile to see all published models

4) To apply models, select the published consensus model and apply it to a single molecule or few compounds. In order to use it with a large number of molecules you will need to attach additional servers (contact us for details).

5) To get a prediction for the challenge set (only possible if you have access to molecular structures available from the Kaggle challenge web site) .

a) Predict the whole set using the consensus model.

b) Export model predictions (in Models/View Pending task clean on green button and export predictions as cvs file with all default options).

c) Run "perl predict.pl modelpredictions\_30307\_compounds.csv > res.csv"

**References**

[1. Tetko, I. V.; Tanchuk, V. Y.; Villa, A. E. Prediction of N-Octanol/Water Partition Coefficients from PHYSPROP Database Using Artificial Neural Networks and E-State Indices. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1407–21.](https://www.zotero.org/google-docs/?aOd9j9)

[2. Tetko, I. V.; Tanchuk, V. Y.; Kasheva, T. N.; et al. Estimation of Aqueous Solubility of Chemical Compounds Using E-State Indices. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1488–93.](https://www.zotero.org/google-docs/?aOd9j9)

[3. Hall, L. H.; Kier, L. B. Electrotopological State Indices for Atom Types: A Novel Combination of Electronic, Topological, and Valence State Information. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 1039–1045.](https://www.zotero.org/google-docs/?aOd9j9)

[4. Huuskonen, J. J.; Villa, A. E. P.; Tetko, I. V. Prediction of Partition Coefficient Based on Atom‐type Electrotopological State Indices. *J. Pharm. Sci.* **1999**, *88*, 229–233.](https://www.zotero.org/google-docs/?aOd9j9)

[5. Winter, R.; Montanari, F.; Noé, F.; et al. Learning Continuous and Data-Driven Molecular Descriptors by Translating Equivalent Chemical Representations. *Chem. Sci.* **2019**, *10*, 1692–1701.](https://www.zotero.org/google-docs/?aOd9j9)

[6. EPA CCTE. Toxicity Estimation Software Tool (TEST), 2022.](https://www.zotero.org/google-docs/?aOd9j9)

[7. Riniker, S.; Landrum, G. A. Open-Source Platform to Benchmark Fingerprints for Ligand-Based Virtual Screening. *J. Cheminformatics* **2013**, *5*, 26.](https://www.zotero.org/google-docs/?aOd9j9)

[8. Landrum, G. RDKit: Open-Source Cheminformatics. **2006**.](https://www.zotero.org/google-docs/?aOd9j9)

[9. Skvortsova, M. I.; Baskin, I. I.; Skvortsov, L. A.; et al. Chemical Graphs and Their Basis Invariants. *J. Mol. Struct. THEOCHEM* **1999**, *466*, 211–217.](https://www.zotero.org/google-docs/?aOd9j9)

[10. Plante, J.; Werner, S. JPlogP: An Improved logP Predictor Trained Using Predicted Data. *J. Cheminformatics* **2018**, *10*, 61.](https://www.zotero.org/google-docs/?aOd9j9)

[11. Hong, H.; Xie, Q.; Ge, W.; et al. Mold2, Molecular Descriptors from 2D Structures for Chemoinformatics and Toxicoinformatics. *J. Chem. Inf. Model.* **2008**, *48*, 1337–1344.](https://www.zotero.org/google-docs/?aOd9j9)

[12. Thormann, M.; Vidal, D.; Almstetter, M.; et al. Nomen Est Omen: Quantitative Prediction of Molecular Properties Directly from IUPAC Names. *Open Appl. Inform. J.* **2007**, *107*, 28–32.](https://www.zotero.org/google-docs/?aOd9j9)

[13. Salmina, E. S.; Haider, N.; Tetko, I. V. Extended Functional Groups (EFG): An Efficient Set for Chemical Characterization and Structure-Activity Relationship Studies of Chemical Compounds. *Mol. Basel Switz.* **2015**, *21*, E1.](https://www.zotero.org/google-docs/?aOd9j9)