SLAS Discovery 29 (2024) 100144

Contents lists available at ScienceDirect

SLAS Discovery

journal homepage: www.elsevier.com/locate/slasd

Full Length Article

The openOCHEM consensus model is the best-performing open-source predictive model in the First EUOS/SLAS joint compound solubility challenge

Andrea Hunklinger [a], Peter Hartog [a], Martin Šícho [b,c], Guillaume Godin [d], Igor V. Tetko [a,e,*]

[a] Institute of Structural Biology, Molecular Targets and Therapeutics Center, Helmholtz Munich-Deutsches Forschungszentrum für Gesundheit und Umwelt (GmbH), DE-85764 Neuherberg, Germany
[b] Leiden Academic Centre for Drug Research, Leiden University, 55 Einsteinweg, 2333 CC Leiden, the Netherlands
[c] CZ-OPENSCREEN: National Infrastructure for Chemical Biology, Department of Informatics and Chemistry, Faculty of Chemical Technology, University of Chemistry and Technology Prague, Technická 5, 166 28, Prague, Czech Republic
[d] dsm-firmenich SA, Rue de la Bergère 7, CH-1242 Satigny, Switzerland
[e] BIGCHEM GmbH, Valerystr. 49, DE-85716 Unterschleißheim, Germany

ARTICLE INFO

Keywords:
Solubility prediction
Kaggle challenge
OCHEM
Consensus
Descriptor based models
Representation learning
Transformer CNN
Graph neural networks

ABSTRACT

The EUOS/SLAS challenge aimed to facilitate the development of reliable algorithms to predict the aqueous solubility of small molecules using experimental data from 100 K compounds. In total, hundred teams took part in the challenge to predict low, medium and highly soluble compounds as measured by the nephelometry assay. This article describes the winning model, which was developed using the publicly available Online CHEmical database and Modeling environment (OCHEM) available on the website https://ochem.eu/article/27. We describe in detail the assumptions and steps used to select methods, descriptors and strategy which contributed to the winning solution. In particular we show that consensus based on 28 models calculated using descriptor-based and representation learning methods allowed us to obtain the best score, which was higher than those based on individual approaches or consensus models developed using each individual approach. A combination of diverse models allowed us to decrease both bias and variance of individual models and to calculate the highest score. The model based on Transformer CNN contributed the best individual score thus highlighting the power of Natural Language Processing (NLP) methods. The inclusion of information about aleatoric uncertainty would be important to better understand and use the challenge data by the contestants.

1. Introduction

The solubility of chemical compounds is a critical parameter in drug discovery, as it directly affects the bioavailability of the compound. Low solubility can result in inaccurate high-throughput screening (HTS) outcomes and can mask toxicity effects during early-stage drug development [1,2]. The intrinsic solubility of a compound refers to its solubility in a neutral state at thermodynamic equilibrium between the solid and dissolved state [3]. Although several modeling approaches have been developed to predict intrinsic solubility, usually small sets of compounds were used since acquiring such experimental data is time-consuming and costly, requiring the achievement of thermodynamic equilibrium and titration to obtain the pH at which the compound is neutral [4]. A cheaper approach involves the determination of kinetic solubility, whereby compounds are first dissolved in DMSO and then added to water to determine precipitation concentration [5]. The kinetic solubility of compounds is frequently higher than intrinsic solubility due to supersaturation and strongly assay dependent [6]. The parameters of assay such as temperature, time required for stirring, etc., could strongly affect the results. An even faster approach, as used for this challenge [7], is to determine threshold solubility, which is crucial in assessing whether a given compound is soluble enough for an HTS screening assay [8].

Although experimental methods can determine solubility, they are often time-consuming and require a considerable amount of resources. Furthermore, the number of compounds that need to be tested in the

* Corresponding author at: Institute of Structural Biology, Molecular Targets and Therapeutics Center, Helmholtz Munich-Deutsches Forschungszentrum für Gesundheit und Umwelt (GmbH), DE-85764 Neuherberg.
E-mail address: igor.tetko@helmholtz-munich.de (I.V. Tetko).

https://doi.org/10.1016/j.slasd.2024.01.005
Received 18 May 2023; Received in revised form 6 January 2024; Accepted 22 January 2024
Available online 3 February 2024

initial phase is substantial, making it challenging to determine the solubility experimentally. Therefore, approximations allow researchers to make a preliminary selection of promising compounds. Developing computational methods to predict solubility can increase the speed of approximations even further than the experimental approaches, and machine learning has emerged as a powerful tool to achieve this. While various computational methods for predicting solubility of compounds were developed [3,9], the conclusion from the recent challenge organized by JCIM was that prediction of this property remains a difficult and challenging task and no significant improvements in the field have been observed over the last ten years [10].

The development of approaches to predict kinetic solubility is difficult due to the limited number of high-quality data as well as dependency on the used assay. Studies using large solubility datasets are frequently performed at pharma companies with *in house* data. For example, back in 2010 Solubility Forecast Index (SFI) was proposed by GlaxoSmithKline scientists [11] based on analysis of 100k kinetic solubility data. The index (SFI = $clogD_{pH7.4}$ + #Ar) provided a simple interpretable relation between predicted octanol-water distribution coefficient calculated at pH7.4 (pH of the assay) and number of aromatic rings in a molecule. There is a public set of kinetic solubility data of about 60k compounds deposited in PubChem [12]. These data have a limited range of solubility values and most of them are insoluble or moderate soluble with only 1.8 % compounds highly soluble according to PubChem (>60 μg/mL) definition. The SFI index reliably predicted highly and medium soluble compounds (>83 %) but was able to correctly identify only 56 % of insoluble (<10 μg/mL) compounds [13], thus indicating challenges to predict low soluble compounds. Recently a multi-task neural network model to predict pH-dependent solubility based on 300k compounds was published by Bayer [14].

This study outlines the solution we presented to the EU-OPENSCREEN - a not-for-profit European Research Infrastructure Consortium (ERIC) [15,16]. The challenge was established to identify the state-of-the-art computational methods for reliable predictions of the threshold solubility of compounds [17]. Here, we present the consensus model, which was selected as the winning open-source solution and address some difficulties encountered during the participation in the challenge. Finally, we place our solution into a greater context and review both our results as well as the data provided by the challenge.

2. Material and methods

2.1. Data

To create a broad data set for this challenge to enable the applicability of resulting models to a wide range of chemical space, EU-OPENSCREEN selected 101,111 small compounds. These chemicals were experimentally tested by the BioFarma Research group from the University of Santiago de Compostela [16]. The exact experimental procedure using nephelometry was described by Brea et al. [7].

The average values of the control compounds Amiodarone and Phenytoin were used to convert the nephelometer values of the compounds into categorical values. The three solubility classes low, medium and high were represented with the numerical values of 0, 1 and 2 respectively throughout the challenge. With a nephelometer result lower than 50,000, which is the average value of Phenytoin, the molecules are classified as high soluble. Between 50,000 and 100,000 they end up in the medium class and with a value higher than the nephelometer average of Amiodarone (100,000) they were sorted in the class with low solubility.

2.2. Challenge evaluation

For the challenge, 101,017 small molecules were made available to participants as SMILES strings [16]. The training set contained 70,710 compounds with activity data and the test set had 30,307 compounds without activity data. The challenge participants were expected to predict the solubility classes of the test set compounds using just chemical structure features and open-source resources. The evaluation of prediction was done using the Quadratic Weighted Kappa (QWK) score [17–19] in two phases: a public performance score was generated using one part of the test compounds (leaderboard set) and was made public on the Kaggle leaderboard immediately after every submission. In the final phase the contributed (usually best-performing) prediction of each team was evaluated on the remaining test set data (private set) after the submission deadline. The performance on the private set was used to identify the winning contribution. Randomly stratified sampling was used to ensure that the same fractions of low, medium, and high soluble compounds were present in the training, leaderboard, and private sets. A more detailed description on the challenge setup and evaluation was described by Wang et al. [17]. The challenge participants were limited to one submission per day and per team, and each participant was required to join only one team during the challenge, which ran from September 19th to December 31st 2022. As a result, each team had a maximum of 102 submissions at their disposal. Our team submitted a total of 18 predictions during the competition and the submission with the best-performing prediction for the leaderboard set was selected for the final evaluation. After the challenge ended, participants were allowed to submit new predictions, and the scores for both the leaderboard and private test sets were made available immediately. These uncounted scores were used for the majority of the analyses reported below.

2.3. General modeling framework

Open Chemical database and Modelling environment (OCHEM) [20] and its recent open source release - openOCHEM (https://github.com/openochem) were used to develop and disseminate models. OCHEM contains numerous descriptors and machine learning methods, which include both commercial and publicly available tools. openOCHEM contains only open source code (GPL, LGPL, AGPL) or binary codes, which were released under permissive licenses and can be used without restrictions for academic, educational, recreational or evaluation purposes. All models were validated using 5-fold cross-validation (5CV). The statistical coefficients calculated using 5CV protocol were also used to evaluate model performances and select models for the challenge as described below.

2.4. Molecule standardization

All molecules were standardized and cleaned using built-in OCHEM workflow. Molecules were stripped from salts (only the largest component was kept for analysis), neutralized then returned as canonical SMILES [21] for processing by various descriptor packages and machine learning methods. OCHEM stores molecules in Kekule representation. A conversion from aromatic to Kekule representation could sometimes result in error. We did not observe any errors with molecules downloaded from the challenge website. An analysis of the InChi detected one duplicated molecule (EOS17062/EOS102135) which was present in both training and test sets.

2.5. Molecular descriptors

In order to ensure the reproducibility and sharing of the developed model, the selection of descriptor packages and machine learning approaches was limited to those available in open-source openOCHEM. To eliminate redundancy, unsupervised filtering was employed by removing descriptors with less than two different values per training dataset and those with strong correlation (Pearson correlation > 0.95). The number of descriptors for each package after filtering is presented in Supplementary Table S1. Based on previous studies indicating that the use of 3D descriptors may not significantly improve solubility

prediction, we focused on 2D descriptors [3]. Following preliminary analysis, we selected nine descriptor packages to build our models. These packages are briefly described below.

ALogPS, OEstate: These 2D descriptors include predicted lipophilicity [22] and intrinsic solubility [23] of molecules which were developed using extended E-state indices [24,25]. The predicted properties as well as E-state indices themselves were used for model development.

Continuous and Data-Driven Descriptors (**CDDD**) [26] were generated from low-level molecular latent representation of a pre-trained deep neural network learning model.

GSFrag [27] counts occurrences of certain special fragments in chemical structures, which include from $k = 2$ to $k = 10$ vertices.

JPlogP [28] is another program to predict lipophilicity of compounds which was made open-source by the authors together with descriptors used to develop it.

EPA included 2D descriptors developed by the US Environmental Protection Agency as part of their Toxicity Estimation Software Tool [29].

Mold2 [30] was developed by the National Center for Toxicological Research and the U.S. Food and Drug Administration and includes both 1D and 2D descriptors for chemoinformatics and toxico-informatics problems.

Quantitative Name Property Relationship (**QNPR**) [31] descriptors are based on 1D molecular representation (SMILES string or IUPAC name), which were split into fragments of a specified length (range).

Functional class fingerprints of diameter 4 (**FCFP4**) [32] were calculated using RDKit [33], which is a cheminformatics and molecular modeling toolkit developed for Python. It provides extensive functionality for working with molecules. For our study we used circular fingerprints of length 1024.

Extended Functional Groups (**EFG**) [34] were based on SMART patterns to identify typical functional groups representing across different chemical classes used in medicinal chemistry.

2.6. Machine learning methods

The final submission contained models based on molecular descriptors and molecular representations based on smiles and graphs (Fig. 1). The machine learning methods based on descriptors included Deep Neural Network (DNN) [35] and CatBoost [36]. DNN [35] was a high-dense neural network composed of seven layers. The calculated descriptors were used as input to the network and the output of the network was the target activity class used as a regression task. CatBoost [36] is an open-source gradient boosting algorithm on decision trees. It uses ordered boosting which allows it to avoid the problem of overfitting, in particular for small training sets. In addition, we also analyzed Random Forest (RF) [37], Associative Neural Networks (ASNN) [38], Support Vector Machines (SVM)[39], Partial Least Squares (PLS) [40], XGBoost [41], as well as traditional k-Nearest Neighbors (kNN) and Multiple Linear Regression (MLR).

Additionally, we also used a number of representation learning methods implemented within KGCNN [42], which were adapted to be used in the OCHEM (Table 1). These methods do not need descriptors but learn internal features based on representation of molecules as SMILES or graphs. The differences in methods appear due to various attributes of nodes and vertices of graphs (i.e., atoms and bonds for molecules), different training algorithms, model pretraining and/or use of 3D information, which was generated internally using RDKit [33] from SMILES [21]. We also used the Transformer CNN [43], which is based on analysis of internal latent representation learned during SMILES canonisation of 1.7 M molecules from the ChEMBL database.

All methods were used with default parameters as specified on the OCHEM web site (https://ochem.eu), which also provides more detailed information about methods and descriptors.

2.7. Statistical parameters

The (Cohen) kappa metric used to rank models in the challenge was first described in 1960 by Jacob Cohen [19]. Kappa was introduced as a coefficient of agreement for the voting of multiple judges for nominal,

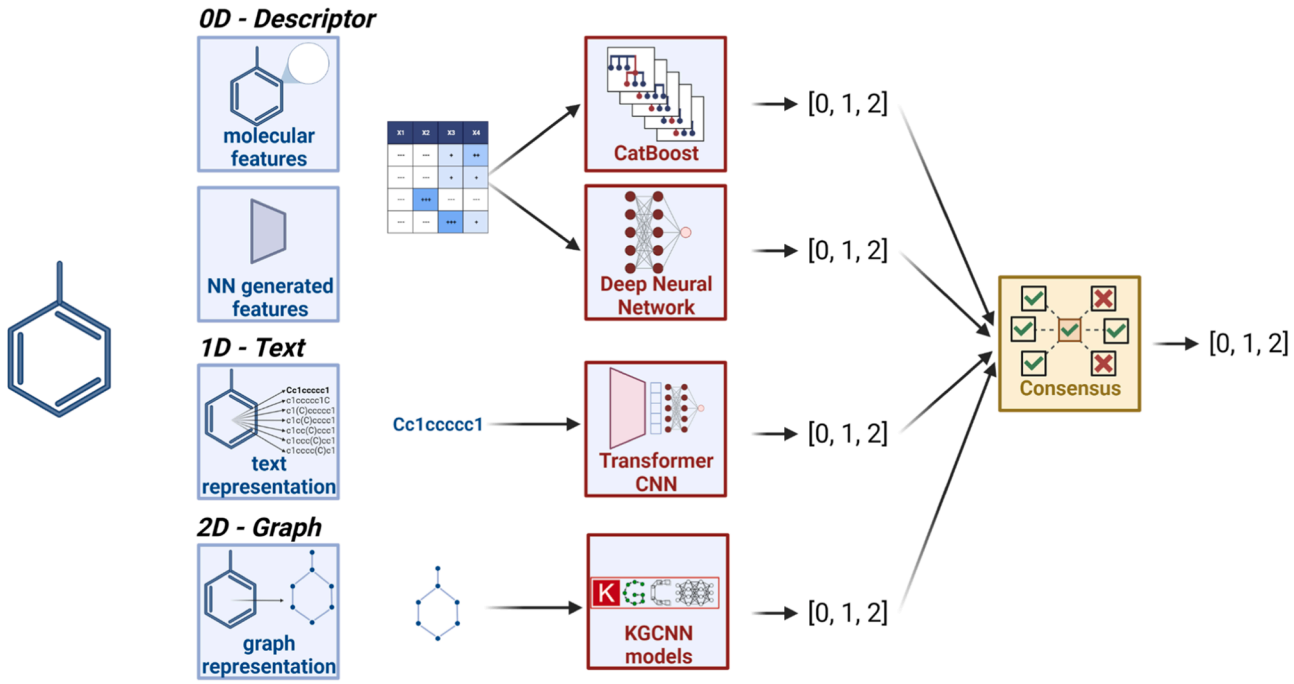

Fig. 1. Overview of molecular representations and models. Molecules were represented by rule-based feature extractions and model-based feature generators, SMILES text, and graph representations. Subsequently, descriptor-based models including CatBoost and DNNs, Transformer CNN and KGCNN models were generated and tested. For the final submission, these models were then combined through a consensus scoring.

Table 1
Overview of graph neural network representation learning used in the challenge. *

N	Abbreviation	Paper full name or Acronym	ref
1	AttFP	Attentive Fingerprints	[44]
2	ChemProp	Directed Message Passing Neural Network	[45]
3	DimeNetPP	Directional Message Passing for Molecular Graphs	[46]
4	GIN	Graph Isomorphism Network	[47]
5	GINE	Graph neural network efficiently pre-training to improve accuracy of GIN	[48]
6	Schnet	One of the first network designed to model atomistic systems by making use of continuous-filter convolutional layers	[49]
7	GAT	Graph ATtention networks	[50]
8	Gatv2	Overcomes limitation of GAT by using dynamic attention	[51]
9	HamNet	Conformation-Guided Molecular Representation with Hamiltonian Neural Networks	[52]
10	GraphSAGE	Inductive Representation Learning on Large Graphs	[53]

*See detailed descriptions of KGCNN algorithms at https://github.com/aimat-lab/gcnn_keras and implementation at https://github.com/openochem/ochem-code/tree/main/WfTools/tools/kgcnn.

though non-ordered, categories. It differs from the simple summation of identical votes by respecting the amount of agreement that is expected to occur by chance. In 1968, Jacob Cohen extended the theory and described the weighted kappa coefficient, which takes the magnitude of disagreement into account [18]. The Quadratic Weighted Kappa (QWK) metric used for the challenge works with a quadratic distance function between categories to enhance the sanction of higher disagreements (i.e., low vs higher accuracy). Like all kappa coefficients, it ranges from +1 for an exact agreement throughout all elements and −1, which would indicate a systematic disagreement. A value of zero shows an alignment, which would occur at random considering the categorical frequencies.

$$O_{i,j} = \begin{pmatrix} n_{0,0} & n_{0,1} & n_{0,2} \\ n_{1,0} & n_{1,1} & n_{1,2} \\ n_{2,0} & n_{2,1} & n_{2,2} \end{pmatrix}, \quad (1)$$

$$E = \left(\sum_i n_{i,0} \sum_i n_{i,1} \sum_i n_{i,2}\right) \otimes \left(\sum_j n_{0,j} \sum_j n_{1,j} \sum_j n_{2,j}\right)$$

For the calculation of the quadratic kappa metric the confusion matrix $O_{i,j}$ and the expected matrix E will be needed. The latter is calculated as the outer product of the two histograms, one coming from the actual labels and the other from the predictions. After normalizing the two matrices, the quadratic kappa value can be calculated as follows using a 3×3 weights matrix w with a quadratic distance function.

$$QWK = 1 - \frac{\sum_{i,j} w_{i,j} \cdot O_{i,j}}{\sum_{i,j} w_{i,j} \cdot E_{i,j}}, w = \begin{pmatrix} 0 & 0.25 & 1 \\ 0.25 & 0 & 0.25 \\ 1 & 0.25 & 0 \end{pmatrix} \quad (2)$$

The Kaggle challenge has provided code and examples of usage of this coefficient [54]. Two strategies were mentioned to optimize performance of models for QWK:

1) Consider the modeling as regression model and optimize the root mean squared error (RMSE)
2) Use QWK directly as error (loss) function

Since the QWK loss function was not available in OCHEM we, following suggestion of Kaggle [54], considered the classification problem as a regression model with numerical values 0, 1 and 2 for the respective classes and selected models with minimal RMSE, which is a traditional loss function for regression tasks (see also the ToxCast challenge [55]).

As aforementioned, we assumed that the proportions of molecules with low, medium and high solubilities in the test sets were identical to those in the training set. Therefore, following model development we sorted all molecules according to their predicted solubility scores and assigned respective classes using the same percentages of compounds as observed in the training set.

3. Results

3.1. Initial investigation

The challenge setup allowed only a limited number of submissions (maximum of 102) and we therefore wanted to approximate the submission score with an openOCHEM internal score. As mentioned before, we approximated submissions using a five-fold cross-validated RMSE value. This also contributed to less overfitting to the public submission set, and in theory generalizing to the unseen private submission set. QWK test sets were calculated after the end of the competition to provide more insight into the decision-making strategies.

We used 18 submissions throughout the time of the challenge to explore different strategies (e.g., two-steps classification, using weighted learning, exploration of classification strategies based on k-nearest neighbors). These initial analyses were comprehensively described in the BSc thesis of the first author. All strategies provided results lower or similar to using consensus modeling when fitting RMSE for a regression task. Because of simplicity and less number of fitting parameters (e.g., weighting of individual classes) this strategy was selected to develop the final model.

During some of the initial steps we included 3D descriptors as well as commercial descriptors. Since their use did not significantly increase performance of developed models, we switched only to those available in the openOCHEM in order to make the final model publicly available and fulfill the challenge requirements. Development of the openOCHEM platform itself as well as implementation of the KGCNN methods were another significant part of the challenge efforts.

3.2. Descriptor-based model analysis

Individual machine learning models were analyzed using an initial descriptor set of AlogPS and OEstate. Internal five-fold cross-validation results and external test QWK scores from models available in openOCHEM are shown in Table 2. Post hoc analysis showed that descriptor-based models included some of the best performing models in both leaderboard and private QWK scores. An exception was ASNN, which contributed the highest score for the Leaderboard dataset but had a lower score for the private set. Overall the RMSE correlated with the QWK scores. Based on Table 2 data we calculated R=−0.38 and R=−0.46 for leaderboard and private scores, respectively. Thus, indeed, RMSE could be a proxy to identify models with high QWK score in the absence of a direct loss function for the latter one.

Table 2
Performance of different methods using AlogPS + OEstate descriptor set.

	Training set 5CV results		Test sets QWK	
Method	RMSE	R^2	Leaderboard	private
DNN	**0.378 ± 0.003**	0.024 ± 0.002	0.113	0.096
CatBoost	**0.380 ± 0.003**	0.016 ± 0.002	0.107	0.092
MLR	0.381 ± 0.003	0 ± 0.001	0.107	0.084
ASNN	0.382 ± 0.003	0 ± 0.001	0.116	0.075
KNN	0.382 ± 0.003	0 ± 0.001	0.067	0.050
PLS	0.382 ± 0.003	0 ± 0.001	0.081	0.058
RF	0.385 ± 0.003	0.014 ± 0.001	0.097	0.095
XGBoost	0.410 ± 0.003	0 ± 0.001	0.081	0.059
LibSVM	0.934 ± 0.005	0 ± 0.001	0.077	0.051

3.3. The final model design

Experience from previous competitions in the ToxCast toxicity prediction challenge [55] and Tox21 challenge [56] introduced the idea of consensus modeling as a successful strategy to improve model performance. Descriptor-based machine learning methods with the two smallest RMSE scores, i.e. DNN and CatBoost, were combined into a consensus model. Selecting these two principally different descriptor-based machine learning approaches, neural networks and decision trees, we hoped to eliminate variances and biases of each specific approach.

Additionally, representation learning methods resulted in similar scores compared to descriptor-based methods (see supplementary Table S1). Based on the same consideration to diversify models for consensus, we included these representation methods in the consensus model b. However, since we did not analyze the results of individual models nor their QWK scores during the challenge, we did not notice that DimeNetPP and GraphSAGE models predicted exactly the same values for all compounds from the both test sets (see Supplementary Table S2). Therefore exclusion of these models did not influence prediction results.

In order to fully use the power of different approaches, we built a consensus model based on all models. Since performance of all methods were very similar, we used a simple average of model predictions. While the final challenge model (#1, Table 3) had larger RMSE than the best descriptor based model developed with DNN using AlogPS+OEstate descriptors (#3, Table 3), it had higher QWKs for both Leaderboard and private sets and contributed a winning top I model of the challenge. After the end of the challenge we noticed that by mistake the CatBoost model based on EFG descriptors was not included in the consensus. Its inclusion slightly increased QWK scores for both sets (#2, Table 3).

The best individual model based on representation learning, Transformer CNN, (#4, Table 3) provided a slightly higher QWK score compared to that of the best DNN model for the Leaderboard set, but the same score for the private set.

The consensus model built using the ASNN method had higher QWK scores than the respective consensus models built using DNN and CatBoost. An attempt to improve results by substituting DNN models with ASNN provided lower QWK scores for both the Leaderboard set (0.142) and private sets (0.112) as compared to the challenge model. Thus, it looks like the combination of modeling methods selected by us provided a nearly optimal score within our approach.

We also analyzed whether linear machine learning methods could be used instead of DNN and CatBoost. For this study we developed a consensus model based on 18 descriptor-based models using Partial Least Squares (PLS) and Multiple Linear Regression Analysis (MLR). Despite the 5CV RMSE of this model (#11, Table 3) was similar to that of the winning model, QWK scores for test sets were lower.

After the end of the challenge Dr. Bernhard Rohde described results of his submission and indicated that there was a variability of the proportions of insoluble compounds per plate [57]. He determined the information about plate and position of each sample on the plate from the data labels based on assumptions of the enumeration of compounds within each assay. His submission calculated higher scores QWK=0.198 and 0.217 for Leaderboard and private sets, respectively. The label-specific descriptors were not accepted as eligible information by the challenge organizers and thus this submission was not listed amid the challenge winners. To understand this organizers' decision, it is essential to understand that for a robust prediction of an unseen molecule solubility, we must not include prior or biased knowledge.

Since his code was publicly available, we decided to evaluate the impact of inclusion of the plate-specific information on the performance of our models. Since representation learning methods did not allow the use of external descriptors we limited our study only to descriptor-based methods.

We added the following descriptors

1) Position of the investigated sample on the plate (row and column number)
2) Plate number
3) Geometrical distance of the sample to plate border
4) Probabilities of low and (low+medium) compounds per plate estimated using training set data

The inclusion of these six descriptors significantly decreased the RMSE for the training set and improved the QWK scores for both private and Leaderboard sets but was still lower than results calculated by Dr. Rohde.

The developed model (#1 and #2) on the main OCHEM website as https://ochem.eu/article/27 or can be provided on a request to be executed within the docker container of openOCHEM.

Table 3
Statistical parameters of representative models.

N	Solution	Individual models in the solution	Training set, 5CV RMSE	Test sets, QWK score Leaderboard	private	Comment
1	Winning challenge consensus	28	0.379	0.147	0.116	An average of descriptors and representation learning models
2	Consensus challenge corrected	27	0.379	0.150	0.119	Same as #1 with addition of CatBoost model based on EFG descriptors and exclusion of DimeNetPP and GraphSAGE
3	DNN, AlogPS+OEstate	1	0.378	0.113	0.096	Best descriptor-based model
4	Transformer CNN	1	0.380	0.117	0.096	Best representation learning model
5	Consensus using molecular descriptors	18	0.378	0.140	0.114	An average of descriptors based models from #2
6	Consensus using DNN models	9	0.378	0.132	0.104	An average of DNN models from #2
7	Consensus using CatBoost models	9	0.378	0.129	0.103	An average of CatBoost models from #2
8	Consensus using only ASNN models	9	0.382	0.131	0.115	An average of ASNN models
9	Consensus using ASNN instead of DNN	27	0.382	0.142	0.112	Same as #2 with ASNN models used instead of DNN
10	Consensus using representation learning models	11	0.383	0.132	0.107	An average of KGCNN and Transformer CNN models from #2
11	Consensus using linear approaches	18	0.380	0.117	0.089	Same as #5 but using PLS and MLR model
12	Consensus using plate specific descriptors	27	0.368	0.206	0.207	Same as #5 but including plate-specific descriptors

4. Discussion

The winning model contributed by our team was based on a consensus of 28 models built on open-source descriptors and algorithms. The consensus modeling allows leveraging performance of individual models by decreasing their variances and biases. The same idea is the basis of ensemble methods, such as Random Forest (RF) [37], which uses bagging, or Associative Neural Networks (ASNN) [38], which uses random split of data, in which an average of models developed using the same approach is used. It is also used in boosting approaches, such as CatBoost [36] and XGBoost [41] which focuses on fitting more hard to learn samples. Based on successful experience with using consensus models in ToxCast [55] and Tox21 [56] challenges, our strategy was to use the same approach with the best set of descriptor and machine learning methods available in openOCHEM, which resulted in the winning solution of the first EUOS/SLAS challenge. This result highlights an importance of consensus modeling as well as shows the high potential of the OCHEM platform to contribute best accuracy models.

The consensus modeling (or any ensemble) is most successful when it combines models that are diverse. Indeed, the use of any number of identical models in an ensemble will not change the results. The bias-variance decomposition analysis [58] indicates that model error can be split into variance and bias. The variance can be decreased by ensembling models that have uncorrelated errors, i.e. by developing models based on different split of data (such as bagging [37]) or, e.g., using different initialisation of weights in neural networks, etc. The bias of a model depends on the used machine learning method and/or data representation and could be decreased by using a consensus of models developed with different approaches or/and different representation of molecules, i.e., different descriptors, SMILES and/or graphs. However, at the same time averaging of models based on different approaches also decreases the variance since errors contributed by different machine learning methods are less unlikely to be correlated. That is why the consensus model based on averaging different models contributed the highest accuracy. We cannot exclude that use of a large consensus of models, or using bagging to decrease variance of individual models could further increase the accuracy of the consensus model. Such analysis, however, would require much more computational resources and it is not clear whether the achieved improvement would be significant.

The 1st EUOS/SLAS solubility Challenge [17] was overall a productive experience. However, future challenges should adhere to more strict guidelines of data randomization, submission procedures and competition rules to avoid unintended information leakage or, vise versa, explicitly include some additional information which could be available for all participants. All compounds should have hashed or randomized ids, without additional information. Compounds should be also ordered randomly to avoid information leakage. This was particularly relevant in the last section of the test set compounds [59], which included a significantly higher amount of low-soluble compounds. Therefore, using id highly skewed the accuracy of participants who used this information based on informed assumption (such as Dr. Rohde - leaderboard second place, who was able to determine plate number as well as position of compounds on the plate based on the compound id) and those using it by chance (such as the leaderboard first place).

A leak of information due to data ordering could be easily prevented by data shuffling mentioned above which could invalidate results of the first leaderboard place. This will not, however, invalidate results based on the plate-specific information. Dr. Rhode models were improved by using both plate number and, to a lesser degree, position of a compound on the plate. The closeness of a compound to the border of the plate is logical, since compounds near to the edge of the plate could be subjected to different technical issues, such as dispersion of a less amount of solvent or/and measurement artifact. However, Brea et al. [7] indicated that these edge effects were mitigated by sealing plates before incubation. This is in line with our finding that inclusion of positions of the investigated sample on the plate (row and column number) and geometrical distance of the sample to plate border did not change the score for the private test set for our descriptor-based methods. This could be due to the fact that we used large numbers of descriptors and thus these additional plate-based descriptors were "lost" amid those derived from the chemical structure of molecules. However, it is more likely that edge effects were not a major indicator for solubility. Future nephelometry protocols could make the methods more robust by randomizing plate location for duplicates and triplicates.

The influence of the plate number is unexpected. As it was shown by Dr. Rhode different plates had very different proportions of insoluble compounds, which could not be explained thus by a chance effect. In our model, inclusion of plate number and probabilities of soluble and medium soluble compounds per plate (based on the plate numbers and observed ratios of different types of compounds per plate) did significantly increase the scores (Table 3, #12) confirming observations by Dr. Rohde [57]. The improvement of scores could reflect either bias in data preparation per plate, i.e., ordering of compounds by providers which could have different solubilities due to, e.g., degradation or change in concentration due to solvent evaporation depending on transportation/storage condition, duration of storage, or different purity, presence of co-solvents, etc., depending on manufacturing process of each provider. Alternatively, measured compounds could be ordered by the time of acquisition and thus again, degradation or evaporation could play a role. All these facts could lead to different enrichments of insoluble compounds per plate. Another problem could be some technical issues with the plates themselves or issues with determination of solubility of Amiodarone and Phenytoin (we can only speculate that these compounds could be, e.g., degraded or their concentrations could be changed due to evaporation of solvent during the course of experiments), which were used to determine thresholds for classification of insoluble and medium soluble compounds. The fact that machine learning methods identified the impact of plate information is important and should be further investigated by scientists performing the experiments. Once reasons are identified, some measures such as renormalization of thresholds due to changes in the solubility of standards could be used. Alternatively, the factors such as time of deposition to collection, provider name, position on the plate, etc. could be provided to all contestants and would allow them to improve accuracy of models and to better predict new molecules.

The participation in a challenge requires significant resources and keeps participants active during its whole duration. However, often a participant can dedicate significant resources only for a short period of time, frequently near to a deadline. In this respect a possibility to submit more than one submission per day whilst limiting the total number of submissions could be important to participants.

A submission of multiple models and selection of the one that provided the highest accuracy for the test set could lead to overfitting. Indeed, we noticed that Leaderboard scores were consistently higher than private scores. We observed this difference for all our submissions as well as for all 100+ individual teams submissions from the challenge website. While models submitted by other teams could have inflated QWK scores for the Leaderboard set due selection of models with higher scores for this set, that was not a case for our analysis, because we compared results without any model selection. Both the Leaderboard and private set scores were highly correlated ($R^2 = 0.978$) based on 100 submissions. The only exception was the submission of Dr. Rohde, which used plate-specific information. An inclusion of the plate-specific descriptors provided a model with a higher score for the private set also for our study. This result may indicate that there was some bias in selection of compounds for the Leaderboard and private sets, which could be due to the use of stratified data sampling for test sets or/and other reasons. The inclusion of plate-specific descriptors compensated for this effect.

Lastly, the competition rules should contain information to avoid any misinterpretation by participants of the competition. This included information about data labels and ordering of compounds which was used by participants for high scoring, but ineligible solutions. Another

example was the discussion on usage of auxiliary data for competition submissions. The remark that "any extra dataset competitors might use, if produced with the same protocol we described, can anytime be used" [60], was interpreted by some of the authors as permission to include additional measurements coming from the same protocol. In this case the extension of experimental data with new measurements could provide significant advantages to some participants, who have an access to such data. A statement that no external data are allowed would be very helpful and would avoid any doubts.

The authors note that given the available data, no submission achieved an evaluation that gives significant predictive performance. In order to improve submissions in future competitions, information about data-driven uncertainty (aleatoric uncertainty) needs to be available to the contestants. Future competitions could include more information for participants to model the data-driven uncertainty, including information on replicating experimental values as well as providing numerical values instead of classes. The latter may allow the researchers to use regression instead of classification model and better account for, e.g., compounds that have solubility near to the class threshold. In case of large variability of experimental values, a larger number of repetitive measurements should be done to decrease aleatoric uncertainty. As aforementioned, the position on the plate can be also provided to account for edge artifacts.

In conclusion, we have shown that consensus modeling by averaging predictions contributed by different machine learning models available in OCHEM was a winning strategy. It provided higher QWK scores than individual models or consensus based on the same class of machine learning methods. Amid individual models, Transformer CNN contributed the highest accuracy.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported.

Acknowledgement

We acknowledge the OPENSCREEN consortium and thank them for organizing the 1st EUOS/SLAS Joint Kaggle Challenge: Compound Solubility. Many thanks to the group of Prof. Dr. Eyke Hüllermeier of the Ludwig-Maximilians university in Munich for enabling the challenge participation in the course of a bachelor thesis. This study was partially funded by the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie Actions grant agreement "Advanced machine learning for Innovative Drug Discovery (AIDD)" No. 956832. M.S. was partially supported by the Ministry of Education, Youth and Sports of the Czech Republic (project number LM2023052). We also would like to thank Dr. Bernhard Rohde for useful discussions.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.slasd.2024.01.005.

References

[1] Di L, Kerns HE. Solubility issues in early discovery and HTS. In: Augustijns P, Brewster ME, editors. Solvent systems and their selection in pharmaceutics and biopharmaceutics. New York, NY: Springer New York; 2007. p. 111–36.
[2] Alsenz J, Kansy M. High throughput solubility measurement in drug discovery and development. Adv Drug Deliv Rev 2007;59:546–67.
[3] Balakin KV, Savchuk NP, Tetko IV. In silico approaches to prediction of aqueous and dmso solubility of drug-like compounds: trends, problems and solutions. Curr Med Chem 2006;13:223–41.
[4] Bergström CAS, Avdeef A. Perspectives in solubility measurement and interpretation. ADMET DMPK 2019;7:88–105.
[5] Kerns HE, Di L, Carter TG. In vitro solubility assays in drug discovery. Curr. Drug Metab. 2008;9:879–85.
[6] Sou T, Bergström CAS. Automated assays for thermodynamic (equilibrium) solubility determination. Physicochem. Characterisation Drug Discov. 2018;27:11–9.
[7] Brea J, et al. High-throughput nephelometry methodology for qualitative determination of aqueous solubility of chemical libraries. SLAS Discov 2024. *in press*.
[8] Bevan CD, Lloyd RS. A high-throughput screening method for the determination of aqueous drug solubility using laser nephelometry in microtiter plates. Anal. Chem. 2000;72:1781–7.
[9] Tetko IV, Yan A, Gasteiger J. Prediction of physicochemical properties of compounds. Applied Chemoinformatics 2018:53–81.
[10] Llinas A, Oprisiu I, Avdeef A. Findings of the second challenge to predict aqueous solubility. J. Chem. Inf. Model. 2020;60:4791–803.
[11] Hill AP, Young RJ. Getting physical in drug discovery: a contemporary perspective on solubility and hydrophobicity. Drug Discov Today 2010;15:648–55.
[12] AID 1996 - Aqueous Solubility from MLSMR Stock Solutions - PubChem https://pubchem.ncbi.nlm.nih.gov/bioassay/1996 (accessed Jul 20, 2023).
[13] Guha R, Dexheimer TS, Kestranek AN, et al. Exploratory analysis of kinetic solubility measurements of a small molecule library. Bioorg Med Chem 2011;19:4127–34.
[14] Bonin A, Montanari F, Niederführ S, et al. pH-dependent solubility prediction for optimized drug absorption and compound uptake by plants. J Comput Aided Mol Des 2023;37:129–45.
[15] Harmel RK, et al. Empowering Research in Chemical Biology and Early Drug Discovery – an Update from the European Research Infrastructure EU-OPENSCREEN. SLAS Discov 2024. *in press*.
[16] Andrea Zaliani, Jing Tang, Julio Martin, Robert Harmel, Wenyu Wang. (2022). 1st EUOS/SLAS joint challenge: compound solubility https://kaggle.com/competitions/euos-slas (accessed Mar 29, 2023).
[17] Wang W, et al. Outline and background for the EU-OS solubility prediction competition. SLAS Discov 2024. *in press*.
[18] Cohen J. Weighted Kappa: nominal scale agreement provision for scaled disagreement or partial credit. Psychol Bull 1968;70:213–20.
[19] Cohen J. A Coefficient of Agreement for Nominal Scales. Educ Psychol Meas 1960;20:37–46.
[20] Sushko I, Novotarskyi S, Körner R, et al. Online Chemical Modeling Environment (OCHEM): web Platform for Data Storage, Model Development and Publishing of Chemical Information. J Comput Aided Mol Des 2011;25:533–54.
[21] Weininger D. SMILES, a chemical language and information system. 1. introduction to methodology and encoding rules. J Chem Inf Comput Sci 1988;28:31–6.
[22] Tetko IV, Tanchuk VY, Villa AE. 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.
[23] Tetko IV, Tanchuk VY, Kasheva TN, et al. Estimation of aqueous solubility of chemical compounds using E-state indices. J Chem Inf Comput Sci 2001;41:1488–93.
[24] Hall LH, Kier LB. Electrotopological state indices for atom types: a novel combination of electronic, topological, and valence state information. J Chem Inf Comput Sci 1995;35:1039–45.
[25] Huuskonen JJ, Villa AEP, Tetko IV. Prediction of partition coefficient based on atom-type electrotopological state indices. J Pharm Sci 1999;88:229–33.
[26] Winter R, Montanari F, Noé F, et al. Learning continuous and data-driven molecular descriptors by translating equivalent chemical representations. Chem Sci 2019;10:1692–701.
[27] Skvortsova MI, Baskin II, Skvortsov LA, et al. Chemical graphs and their basis invariants. J Mol Struct THEOCHEM 1999;466:211–7.
[28] Plante J, Werner S. JPlog P: an improved logP predictor trained using predicted data. J Cheminformatics 2018;10:61.
[29] EPA C.C.T.E. Toxicity estimation software tool (TEST), 2022.
[30] Hong H, Xie Q, Ge W, et al. Mold2, molecular descriptors from 2D structures for chemoinformatics and toxicoinformatics. J Chem Inf Model 2008;48:1337–44.
[31] 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.
[32] Riniker S, Landrum GA. Open-source platform to benchmark fingerprints for ligand-based virtual screening. J Cheminformatics 2013;5:26.
[33] Landrum, G.RDKit: Open-source cheminformatics. 2006.
[34] Salmina ES, Haider N, Tetko IV. Extended functional groups (EFG): an efficient set for chemical characterization and structure-activity relationship studies of chemical compounds. Mol Basel Switz 2015;21:E1.
[35] Sosnin S, Karlov D, Tetko IV, et al. Comparative study of multitask toxicity modeling on a broad chemical space. J Chem Inf Model 2019;59:1062–72.
[36] Prokhorenkova L, Gusev G, Vorobev A, et al. CatBoost: unbiased boosting with categorical features. In: Proceedings of the 32nd International Conference on Neural Information Processing Systems. Montréal, Canada: Curran Associates Inc.; 2018. p. 6639–49.
[37] Breiman L. Random forests. Mach Learn 2001;45:5–32.
[38] Tetko IV. Associative neural network. Methods Mol Biol Clifton NJ 2008;458:185–202.
[39] Cortes C, Vapnik V. Support-vector networks. Mach Learn 1995;20:273–97.
[40] Wold S, Sjöström M, Eriksson L. PLS-regression: a basic tool of chemometrics. PLS Methods 2001;58:109–30.

[41] Chen T, Guestrin C. XGBoost: a scalable tree boosting system. In: Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Min; 2016. p. 785–94.
[42] Reiser P, Eberhard A, Friederich P. Graph Neural networks in TensorFlow-Keras with RaggedTensor representation (Kgcnn). Softw Impacts 2021;9:100095.
[43] Karpov P, Godin G, Tetko IV. Transformer-CNN: swiss knife for QSAR modeling and interpretation. J Cheminformatics 2020;12:17.
[44] Xiong Z, Wang D, Liu X, et al. Pushing the boundaries of molecular representation for drug discovery with the graph attention mechanism. J Med Chem 2020;63:8749–60.
[45] Yang K, Swanson K, Jin W, et al. Analyzing learned molecular representations for property prediction. J Chem Inf Model 2019;59:3370–88.
[46] Gasteiger J, Groß J, Günnemann S. Directional message passing for molecular graphs. In: International Conference on Learning Representations; 2020.
[47] Xu, K.; Hu, W.; Leskovec, J.; et al. How Powerful Are graph neural networks? *ArXiv181000826 Cs Stat* 2019.
[48] Hu, W.; Liu, B.; Gomes, J.; et al. Strategies for Pre-training graph neural networks. *ArXiv E-Prints* 2019, arXiv:1905.12265.
[49] Schütt KT, Sauceda HE, Kindermans P-J, et al. SchNet – a deep learning architecture for molecules and materials. J Chem Phys 2018;148:241722.
[50] Veličković, P.; Cucurull, G.; Casanova, A.; et al. Graph attention networks. *ArXiv E-Prints* 2017, arXiv:1710.10903.
[51] Brody S, Alon U, Yahav E. How attentive are graph attention networks?. In: International Conference on Learning Representations; 2022.
[52] Li, Z.; Yang, S.; Song, G.; et al. HamNet: conformation-guided molecular representation with hamiltonian neural networks. *ArXiv E-Prints* 2021, arXiv:2105.03688.
[53] Hamilton WL, Ying R, Leskovec J. Inductive representation learning on large graphs. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, California, USA: Curran Associates Inc.; 2017. p. 1025–35.
[54] Understanding the metric: quadratic weighted Kappa https://kaggle.com/code/carlolepelaars/understanding-the-metric-quadratic-weighted-kappa (accessed Mar 11, 2023).
[55] Novotarskyi S, Abdelaziz A, Sushko Y, et al. ToxCast EPA in vitro to in vivo challenge: insight into the Rank-I model. Chem Res Toxicol 2016;29:768–75.
[56] Abdelaziz A, Spahn-Langguth H, Schramm K-W, et al. Consensus Modeling for HTS assays using in silico descriptors calculates the best balanced accuracy in Tox21 challenge. Front Environ Sci 2016:4.
[57] Rohde, Bernhard. 2nd Place Solution (Draft) https://www.kaggle.com/competitions/euos-slas/discussion/376756 (accessed Mar 12, 2023).
[58] Geman S, Bienenstock E, Doursat R. Neural networks and the bias/variance dilemma. Neural Comput 1992;4:1–58.
[59] Andrea Zaliani, Jing Tang, Julio Martin, Robert Harmel, Wenyu Wang. (2022). 1st EUOS/SLAS joint challenge: compound solubility | Kaggle Discussions https://www.kaggle.com/competitions/euos-slas/discussion/377428 (accessed Apr 3, 2023).
[60] Zaliani, A. Can we use other dataset? https://www.kaggle.com/competitions/euos-slas/discussion/365152 (accessed Mar 26, 2023).