CompBat Deliverable

D1.2 Report on methodology for high-throughput screening of RFB compounds

Grant Agreement number	875565
Action Acronym	CompBat
Action Title	Computer aided desing for next generation flow batteries
Funding Scheme	H2020-LC-BAT-2019
Duration of the project	36 months, 1 February 2020 – 31 January 2023
Work package	WP1 - High-throughput screening and machine learning
Due date of the deliverable	31 July 2021
Actual date of submission	30 July 2021
Lead beneficiary for the deliverable	AALTO
Dissemination level of the deliverable	Public

Project coordinator's scientific representative

Dr. Pekka Peljo

University of Turku

Department of Mechanical and Materials Engineering

pekka.peljo@utu.fi

CompBat project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 875565. This document has been produced by the CompBat project. The content in this document represents the views of the authors, and the European Commission has no liability in respect of the content.

compbat

Authors		
Name	**Beneficiary**	**E-mail**
Kari Laasonen	AALTO	kari.laasonen@aalto.fi
Rasmus Kronberg	AALTO	rasmus.kronberg@aalto.fi
Imre Pápai	TTK	papai.imre@ttk.hu
Andrea Hamza	TTK	hamza.andrea@ttk.hu
Ádám Madarász	TTK	madarasz.adam@ttk.hu
Németh Flóra Boróka	TTK	nemeth.flora@ttk.hu

Internal QA			
Reviewer	**Date of review**	**Comments**	**Date of revision**
Pekka Peljo	30.7.2021	OK	

Abstract/Executive summary (of the deliverable)

This report presents the methodology and the results of our machine learning (ML) analysis carried out within Task 1.2 of the CompBat project. The molecular database DB-I reported in deliverable D1.1 has been updated in terms of the applied computational protocol, as well as the number of pyridoxal derivatives. The updated database DB-II has been obtained with an improved protocol and it includes 6712 molecules. Various machine learning approaches, including the commonly used random forest algorithm and several deep-learning techniques, were applied to our molecular libraries and their performance was tested for reduction potentials and aqueous solubilities. We used different molecular representations (strings, fingerprints and graphs) in the ML analysis, which always involved a training process followed by a test procedure. We demonstrate that all the applied ML models perform well for both investigated properties; the overall accuracy of the ML predictions is higher than that of the applied quantum chemical computational protocol. Feature attribution analysis of pyridoxal derivatives has also been carried out to assign reliable and consistent importance to different molecular substructures. The impact of electron-withdrawing and donating groups on redox potentials, as well as charged and polar groups on solubilities is confirmed. The presented ML analysis represents an efficient methodology for high-throughput screening of RFB compounds, which will be further exploited within the CompBat project.

CompBat project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 875565. This document has been produced by the CompBat project. The content in this document represents the views of the authors, and the European Commission has no liability in respect of the content.

compbat

Contents

1. Introduction
2. Updated Molecular Database
3. Machine Learning Methodology
4. Results and Discussion
5. Conclusions
6. References

Appendix 1: Consortium
Appendix 2: Database in tabulated form
Appendix 3: Nested cross-validation
Appendix 4: Hyperparameters of DeepChem methods
Appendix 5: Hyperparameters of the 3DGCN model

1 Introduction

The CompBat project aims at developing various tools for discovery of new prospective candidates for next generation of redox flow batteries (RFB). High-throughput screening (HTS) that enables the identification of promising candidates of water-soluble redox-active compounds for experimental synthesis and electrochemical characterization is one target tool, and these developments are carried out within work package WP1.

In the first phase of this project, we developed an efficient computational protocol and built a molecular library, which includes the structures, reduction potentials and aqueous solvation free energies of over 6200 pyridoxal-based neutral and charged species. This molecular database was reported in deliverable D1.1 of this project. In a follow-up work, the molecular library has been expanded by additional pyridoxal derivatives, and herein we present an updated version of the molecular database, which was obtained by an improved version of the computational protocol.

As a next step towards high-throughput screening solutions for discovery of new candidates for the next generation of redox flow batteries, we applied various machine learning (ML) approaches using the developed pyridoxal databases. Our primary goal was to test the predicting power of different ML methods for the redox potential and solubility data, but in addition, we intended to perform feature-property analysis as well, which provides insight into the importance of different molecular substructures involved in the pyridoxal derivatives. In this report, we describe the applied ML methodology and present the results of machine learning analysis.

CompBat project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 875565. This document has been produced by the CompBat project. The content in this document represents the views of the authors, and the European Commission has no liability in respect of the content.

2 Updated Molecular Database

The molecular library generated in the first period of this project and reported in our D1.1 deliverable involves the 3D structures, reduction potentials and aqueous solubility data for about 6200 molecules classified into four different molecular sets depending on the substituents of their common pyridoxal framework (Figure 1). This molecular library will be referred to as Database I (**DB-I**) in the present report.

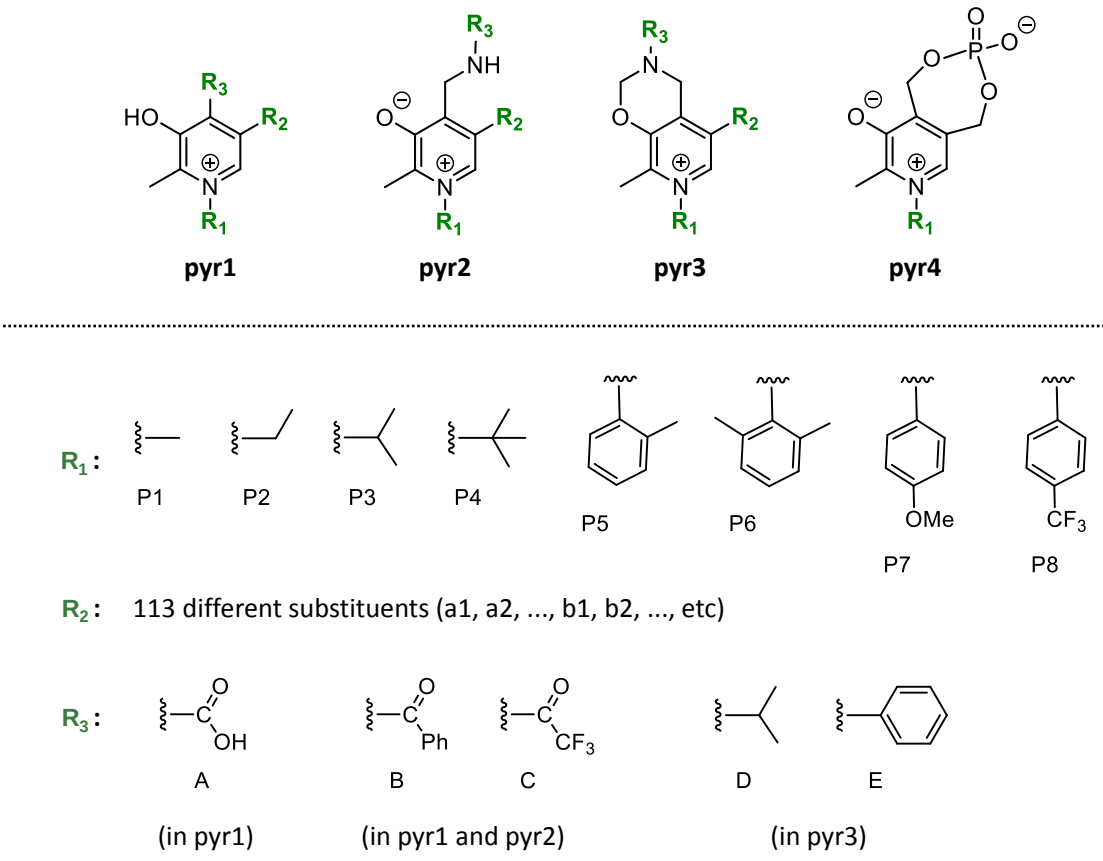

Figure 1: Molecular sets derived from the pyridoxal framework. The full set of R_2 substituents was provided in deliverable D1.1.

The combination of R_1, R_2 and R_3 substituents gives rise to 6336 molecules (2712 for **pyr1**, 1808 for **pyr2** and **pyr3**, 8 for **pyr4** sets); however, our original computational protocol gave rise to unexpected chemical rearrangments (cyclization, for instance) for some molecules, which were omitted from the database. This computational artifact could be eliminated by implementing an updated version of the *crest* module [1] used in our computational protocol.

CompBat project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 875565. This document has been produced by the CompBat project. The content in this document represents the views of the authors, and the European Commission has no liability in respect of the content.

In our original protocol, the final structures of all the species were sorted based on the electronic energies computed for the conformers. However, as all calculations are performed in aqueous phase, it is more consistent to perform the selection of conformers according to their solvent phase Gibbs-free energies. We have, therefore, improved our protocol with two major changes: a) the updated version of the *crest* module is applied for the conformational search, and b) the conformers are sorted based on their solvent phase Gibbs-free energies.

The molecular library has also been expanded by 376 additional molecules proposed by the JYU partners. Namely, the **pyr1** set has been expanded by new R_2 substituents (Figure 2), the $–CH_2OH$ group for R_3, and a new set **pyr5** was also introduced (Figure 3).

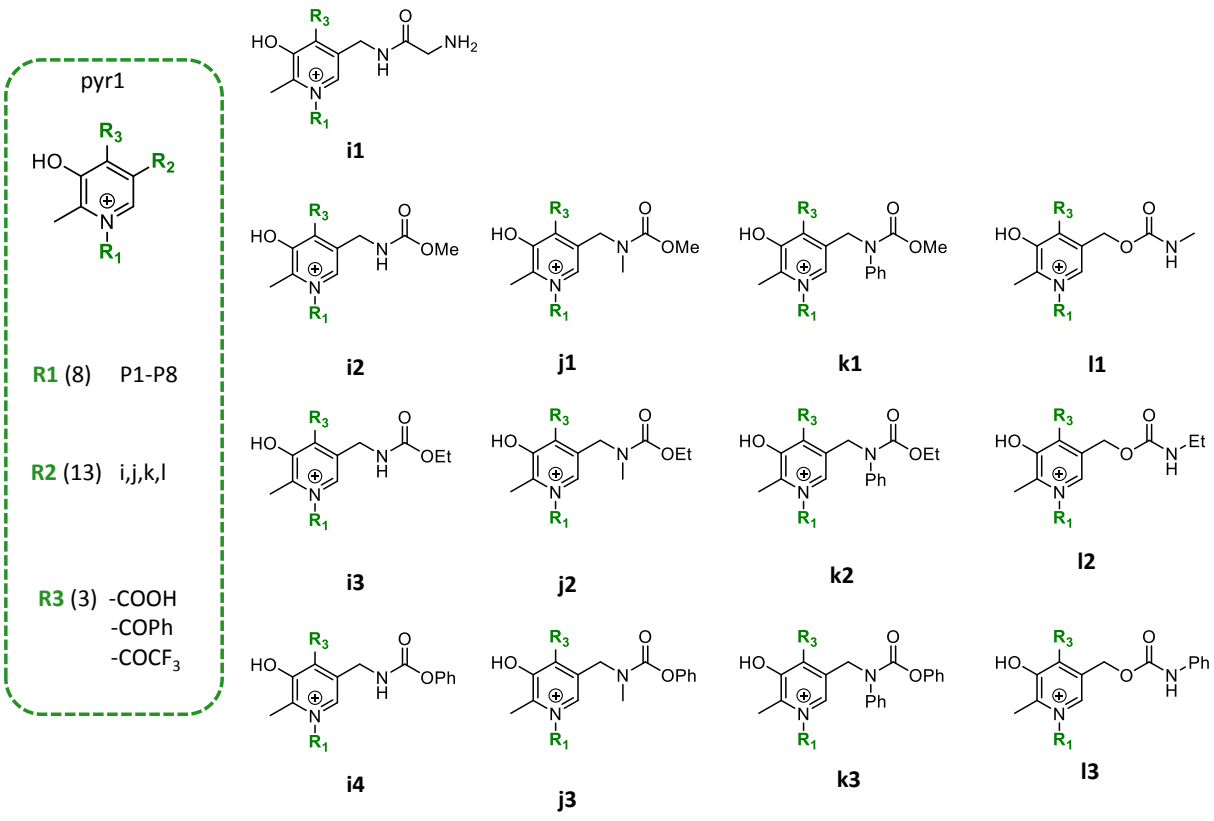

Figure 2. Proposed R_2 substituents for the **pyr1** molecular set.

CompBat project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 875565. This document has been produced by the CompBat project. The content in this document represents the views of the authors, and the European Commission has no liability in respect of the content.

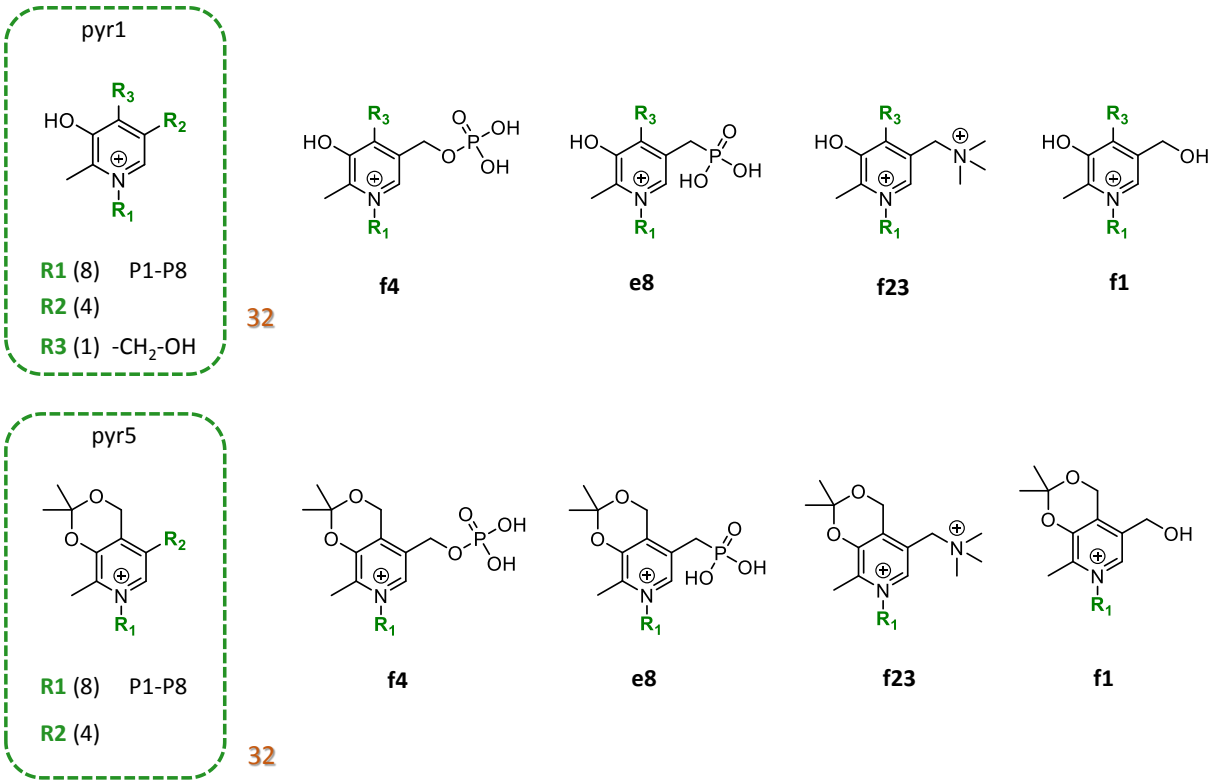

Figure 3. Proposed R_2 substituents for the **pyr1** (R_3=-CH_2-OH) and **pyr5** molecular set.

The new database thus consists of 6712 molecules, for which the reduction potentials and the Gibbs free energies of solvation were recomputed using our updated computational protocol. Connectivity check has been performed for the reduced and oxidized forms of all molecules. The changed connectivity indicates that bond formation or breaking took place during the reduction process. This information is also included in the generated database. The new version of the molecular library is referred to as Database II (**DB-II**). The computed data are provided in a tabulated form as a supplementary information of this report (separate Excel documents; see Appendix 2).

We note that the overall distribution of the reduction potential values (Figure 4) is very similar to the results obtained by the original protocol. The potential values for the new set **pyr5** are in the range of -1.6 /-1.2 V vs. standard hydrogen electrode (SHE). In the potential region of electrochemical relevance (above -1.0 V) the predominant set is **pyr1**.

CompBat project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 875565. This document has been produced by the CompBat project. The content in this document represents the views of the authors, and the European Commission has no liability in respect of the content.

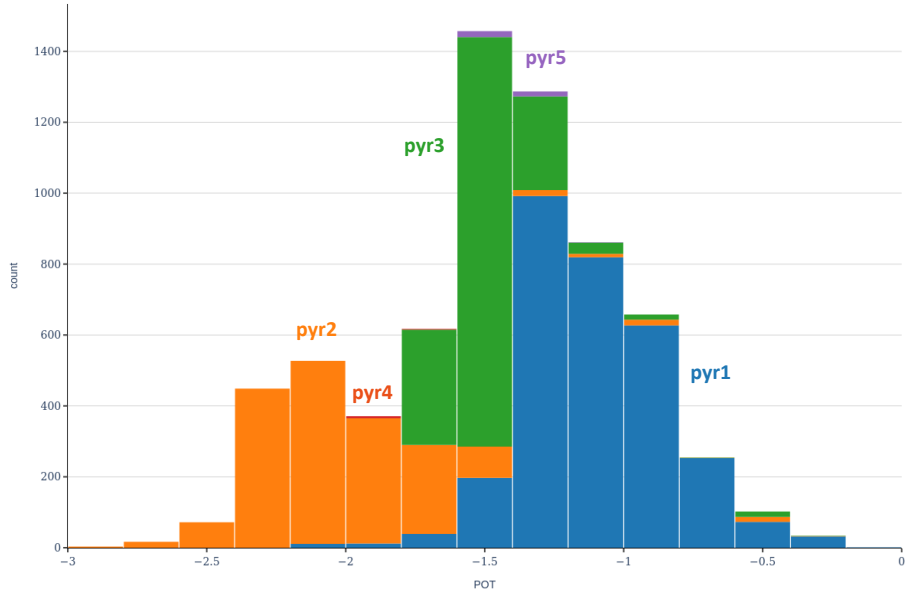

Figure 4. Distribution of computed potentials (vs. SHE) coloured according to the molecular sets.

3 Machine Learning Methodology

In this section, we describe the general concepts and the methodology of the machine learning approach used in our studies.

Data processing

The molecules of our datasets are stored in SDF (Structure-Data File) format, which is a Molfile representation. It contains the molecules' atomic compositions, the bonds between them, and also some other properties. We used the *Rdkit* open-source chemoinformatics package [2] in Python to convert the Molfile data to other molecular representations (SMILES, fingerprints and graphs). Two different approaches were used to evaluate the performance of the ML methods. The dataset was either split randomly into training, validation and test sets (in 80%, 10% and 10% ratio), or a nested cross-validation procedure was applied (80%-20% train-test split). Both approaches allow to fine-tune the model hyperparameters to obtain optimal results. The performance of the ML models was quantified in terms of metrics, such as mean absolute error (MAE), root mean square error (RMSE), and the coefficient of determination (R^2) parameters, obtained from regression analysis.

CompBat project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 875565. This document has been produced by the CompBat project. The content in this document represents the views of the authors, and the European Commission has no liability in respect of the content.

Molecular representations

We used the following molecular representations in our present work.

SMILES strings

The Simplified Molecular-Input Line-Entry System (SMILES) [3] is simple line notation for encoding molecular structures. It is a linear series of ASCII characters bearing information on the atoms and bonds of the molecules, and on their connectivity as well.

Molecular fingerprints

Molecular fingerprints are representations of chemical structures originally designed to assist in chemical database substructure searching, but later used for analysis tasks and machine learning analysis as well. The most common type of fingerprint is a series of binary digits. i.e. bits either on (1) or off (0), that represent the presence or absence of particular substructures in the molecule. Several fingerprint types and formats are available, of which the Extended-Connectivity FingerPrints (ECFP) [4] were utilized in our studies. ECFP fingerprints involved either 1024 or 2048 bit vectors.

Graph representation

A molecule can be represented as a graph comprising atoms as nodes and bonds as edges. For each node, there is a feature vector corresponding to the atom features (e.g. atom type, number of other atoms attached, etc.) and other chemical informations, such as aromaticity, for instance. The adjacency and feature matrices are the basic inputs in this representation. The adjacency matrix represents the connectivity between the atoms (either 0 or 1); the feature matrix is constructed from a set of feature vectors representing the atom properties such as atom type, formal charge, and hybridization.

Applied machine learning methods

Several machine learning methods have been applied in the present work that range from classic machine learning solutions to more complex, deep learning techniques, which are based on convolutional neural networks (CNN) and graph convolutional networks (GCN).

Random forest regression analysis

One of the ML workflows employed in this work was implemented using the tools provided by the *scikit-learn* Python package [5]. Random forest (RF) [6] ensemble learning was used as the ML algorithm using 500 decision trees for a reasonable balance between accuracy and computational cost. 5x5-fold nested cross-validation (CV) was used for model validation and testing. Notably, the nested CV approach facilitates an exhaustive analysis of the full dataset composed of more than 6200 pyridoxal derivatives (DB-I). Biased performance metrics and feature attributions were avoided as different disjunct subsets of the data were iteratively used for model hyperparameter optimization and validation, as well as for model training and testing. A schematic of the nested CV procedure is presented in the Appendix 3. For each outer CV fold the number of features considered when splitting

CompBat project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 875565. This document has been produced by the CompBat project. The content in this document represents the views of the authors, and the European Commission has no liability in respect of the content.

a node was tuned. Considering approximately 40–60% of all features was in most cases observed to yield the best performance. In the present RF analysis, we used ECFP fingerprint molecular representation as implemented in the *Rdkit* package.

Deep-learning models

Deep-learning ML methods have also been applied to our databases **DB-I** and **DB-II**. These ML approaches are based on artificial neural networks, specifically on convolutional neural networks (CNN) [7], and they use multiple layers to progressively extract higher-level features from the raw input. CNNs have three main types of layers. The convolutional and pooling layers are important in feature extraction, whereas the fully connected layer serves decision making purposes and executes the regression.

In some of our deep-learning studies we used the *DeepChem* [8] machine learning library via the *Google Colab* service [9]. The deep-learning analysis was carried out in collaboration between TTK and the Department of Automation and Applied Informatics of the Budapest University of Technology and Economics.

Various types of custom *DeepChem* models have been tested, but herein, we present results only for the following models:

TextCNNModel – This CNN model applies multiple 1D convolutional filters to the padded strings, then max-over-time pooling is applied on all filters, extracting one feature per filter. All features are concatenated and transformed through several hidden layers to form predictions. The model was initially developed for sentence-level classification tasks, with words represented as vectors [10]. In this implementation, SMILES strings are dissected into characters and transformed to one-hot vectors in a similar way.

RobustMultitaskRegressor – This model implements a neural network for robust multitasking. The key idea of this model is to have bypass layers that feed directly from features to task output [11]. This might provide some flexibility to route around challenges in multitasking with destructive interference. Multi-task learning (MTL) aims to improve the performance of multiple related tasks by exploiting the intrinsic relationships among them.

AttentiveFPModel – This model was developed for graph property prediction based on graph convolution networks (GCN). It is more than a simple GCN because it introduces an attention mechanism for extracting nonlocal effects at the intramolecular level [12]. The model is atom-centric, so every atom has a node representation, which is a mix of the atom's properties and its neighborhood's features, and the bonds between them. Figuring out which nodes have bigger impact is the main task of this model.

In addition to these *DeepChem* models, we tested a three-dimensional graph convolutional network (***3DGCN***), which is based on 3D molecular graphs [13] so it utilizes the spatial information from molecular topology. In this model, the adjacency and feature matrices are extended by the relative position matrix to account for the position of the vertices. This latter matrix is designed to involve the

CompBat project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 875565. This document has been produced by the CompBat project. The content in this document represents the views of the authors, and the European Commission has no liability in respect of the content.

inter-atomic positions, rather than individual atomic positions, which ensures translational and rotational invariance.

The ***3DGCN*** model was implemented on local computers using the *PyCharm* [14] integrated development environment and *Conda* [15] open source package management system. Some of the Python codes and libraries were obtained from the 3DGCN GitHub project [16].

4 Results and Discussion

4.1 Results of the RF analysis

The distribution of the calculated redox potentials and solvation free energies in the **DB-I** dataset is visualized in Figures 5a and 5b. Whereas the redox potentials are rather evenly distributed with a mean and standard deviation of -1.45 ± 0.41 eV, the solubilities are more heavily concentrated to the low-solubility region (above -100 kcal/mol).

Analyzing the **DB-I** pyridoxal dataset using the outlined ML workflow yields RF predictions for the sample molecule redox potentials and solubilities, which are compared against the computational values in Figures 5c and 5d. Predictions on both training (trainval) and test set samples are illustrated for one example outer CV fold to contrast the performance on the two data subsets. We emphasize, however, that the performance on the test data is nonetheless what determines the true generalization performance of the models. We note that training an RF model with one set of hyperparameters and the present dataset (80%-20% train-test split) is relatively fast, taking roughly 1–2 min depending on the hyperparameters. Conversely, formation of predictions with a trained model takes only a few seconds. If opted, the nested CV approach adds, however, to the total computational cost depending on the number of inner and outer folds and most importantly the specified number of randomized hyperparameter search iterations.

Clearly, the RF algorithm is able to predict rather well pyridoxal redox potentials on the higher end of the distribution (E_{red} > -1.5 V), while more negative values show a more pronounced scattering with respect to the true computational values. Overall, the performance is nevertheless relatively good as shown by the unbiased generalization performance metrics averaged over all outer CV folds in Table I. Indeed, the MAE of the test set reaches almost the level of chemical accuracy, roughly 0.04 eV (1 kcal/mol). It is furthermore noteworthy, that this error is smaller in magnitude than the intrinsic error both in the tight-binding calculations benchmarked against DFT (MAE 0.1 V) as well as the DFT values benchmarked against experimental results (MAE 0.24 V) in deliverable D1.1.

The solubilities are illustrated in Figure 5d to be very well predicted by the employed RF algorithm, also reflecting a MAE close to the limit of chemical accuracy. Importantly, even though the majority of the samples are concentrated to the higher end of the solvation free energy distribution, the model is capable of very accurate prediction of highly soluble molecular samples. This indicates that the employed ECFP molecular representation is especially suited for the prediction of solvation properties.

CompBat project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 875565. This document has been produced by the CompBat project. The content in this document represents the views of the authors, and the European Commission has no liability in respect of the content.

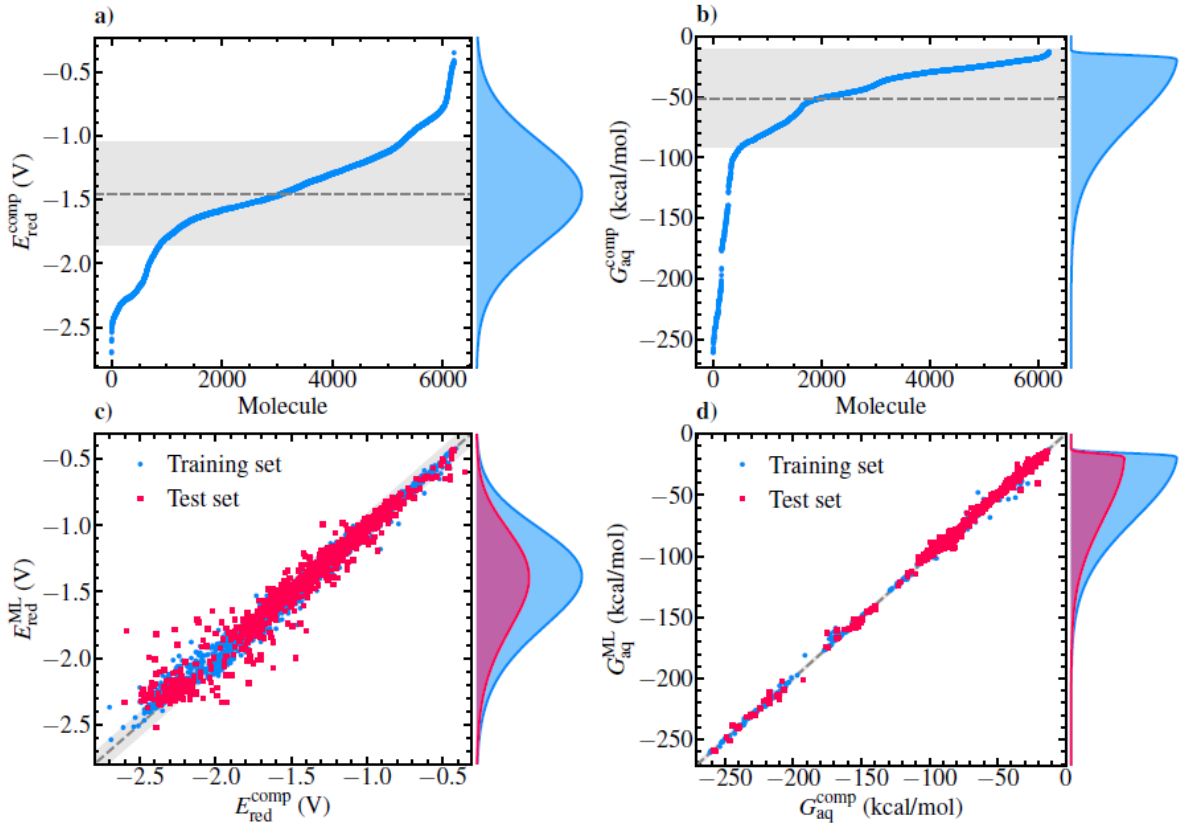

Figure 5. Sorted a) redox potentials and b) solubilities in the pyridoxal dataset. The mean and standard deviation -1.45 ± 0.41 V and -51.1 ± 40.2 kcal/mol, respectively, are marked by the dashed line and shaded area in both panels. Parity plots of the c) redox potentials and d) solubilities predicted by the RF model versus the true computational values of the training and test set samples of an example outer CV split. Ideal correlation is marked by the dashed line and the shaded area corresponds to ± the test set RMSE averaged over all CV splits. The skew normal distributions right of the panels illustrate the distribution of samples within the full data and stratified example training and test sets. The test set distribution (red) has been arbitrarily scaled for visual purposes.

Table I: Redox potential and solubility test set scoring (units in V and kcal/mol, respectively).

	5×5-fold stratified CV		Leave-one-group-out CV	
	E_{red}^{ML}	G_{aq}^{ML}	E_{red}^{ML}	G_{aq}^{ML}
MAE	0.05 ± 0.01	1.3 ± 0.1	0.10 ± 0.05	3.8 ± 2.5
RMSE	0.09 ± 0.01	2.3 ± 0.2	0.14 ± 0.05	6.0 ± 4.2
R^2	0.95 ± 0.01	0.99 ± 0.01	0.86 ± 0.10	0.97 ± 0.05

CompBat project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 875565. This document has been produced by the CompBat project. The content in this document represents the views of the authors, and the European Commission has no liability in respect of the content.

As an aggressive test of the model generalization performance, we have also performed an alternative cross-validation strategy in the nested CV outer loop. Here, the data is not split with stratification, but instead each member of the R_1 substituent group is left out in turn and the trainval set is formed using all remaining molecules. Consequently, we test whether the RF algorithm together with the applied molecular representation is able to predict redox potentials and solubilities of samples with completely unseen substituents. The performance metrics of this leave-one-group-out (LOGO) experiment are also shown in Table I and indicate that the model performance notably worsens. Specifically, compared to the stratified nested CV, the errors are seen to increase by a factor of 2-3 for both the redox potentials and the solvation free energies. To improve this performance, either more training data is needed or the molecular representation should be tuned to better reflect fundamental properties of the molecules affecting redox potentials and solubilities.

4.2 Performance of DeepChem models

Redox potentials and solvation free energies for both, **DB-I** and **DB-II** pyridoxal datasets were analyzed by using the models incorporated in the *DeepChem* framework: *TextCNNModel*, *RobustMultiRegressor* and *AttentiveFPModel*. Herein we present the results obtained for the pyridoxal dataset **DB-II**. The database was split into training-validation-test parts in the ratio of 80%-10%-10%. For each dataset, 5-5 trainings were performed, by keeping the same 80% amount of the training set albeit randomly reordering the data of the full dataset before split. A number of 100 *epoch* were used. The Early Stopping monitored the MAE of the validation set at each 602 (**DB-I**) or 671 (**DB-II**) training steps and decided of training continuation. Default settings of the hyperparameters were used in the training and prediction procedures (Appendix 4).

Analyzing the 5 different trainings by the *TextCNNModel* for database **DB-II** (Table II), it is apparent that the MAE for the three sets is almost invariable to the reordering of the database and the *TextCNNModel* is able to predict the redox potential at fairly high accuracy. The expected value for MAE of the test set is 0.062 ± 0.005 V, while the R^2 for the whole database is 0.984 ± 0.001.

Table II: MAE metrics of redox potentials predicted by *TextCNNModel* for 5 separate trainings (units in V).

MAE	Training	Validation	Test	RMSE	R^2
train 1	0.036	0.068	0.066		
train 2	0.030	0.063	0.061	0.062	0.984
train 3	0.034	0.064	0.060	±	±
train 4	0.033	0.063	0.064	0.005	0.001
train 5	0.044	0.065	0.076		

CompBat project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 875565. This document has been produced by the CompBat project. The content in this document represents the views of the authors, and the European Commission has no liability in respect of the content.

Predictions on all training, validation and test sets of the pyridoxal dataset are illustrated in Figure 6a. The correlation between the true computational values and predicted redox potentials is rather good. A higher scattering can be observed at more negative values of the potential, in the range below -2.0 V. The MAE of the test set for the training presented in Figure 6 is 0.060 V.

The distribution of the absolute error for the whole database is presented in Figure 6b. It is apparent, that the MAE is less than 0.19 V for 98.88% percent of the molecules, and for the 93.5% of the molecules is less than 0.10 V.

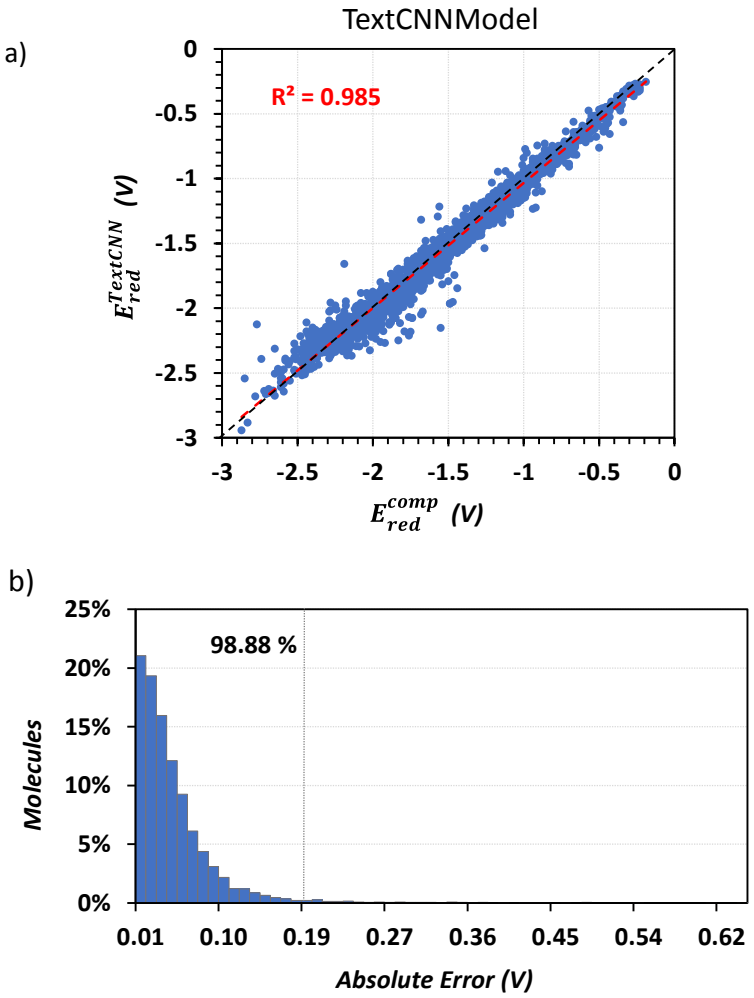

Figure 6. Redox potentials as obtained by TextCNN modelling. a) Scatter plot between the DFT-computed versus *TextCNNModel*-predicted redox potentials. b) Distribution over the absolute error of predicted redox potentials. The amount of molecules represented by % are shown in the vertical axis.

CompBat project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 875565. This document has been produced by the CompBat project. The content in this document represents the views of the authors, and the European Commission has no liability in respect of the content.

The performances for predicting the redox potentials of *RobustMultitaskRegressor* and *AttentiveFPModel* are presented in Figure 7.

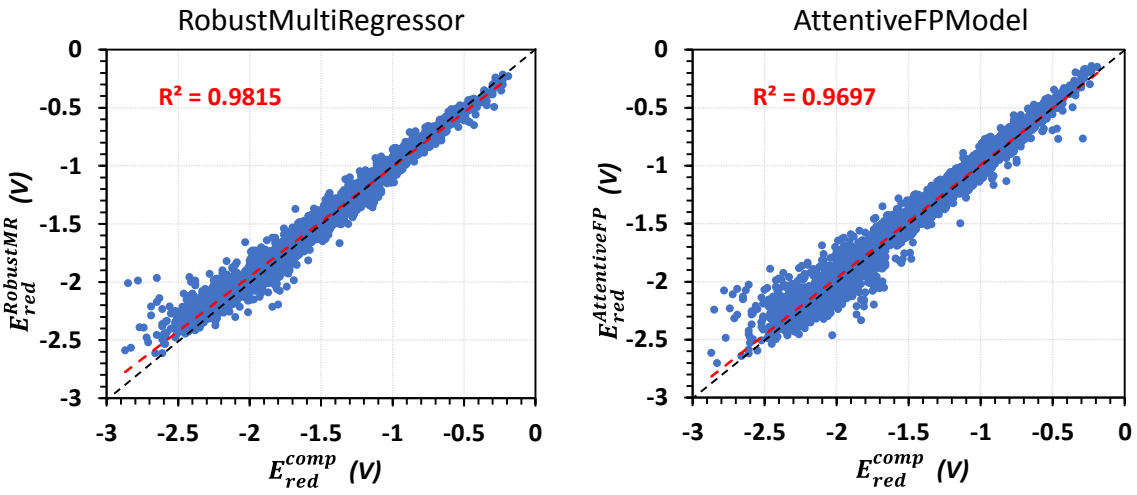

Figure 7. Scatter plot between the DFT-computed and the predicted redox potentials by the *RobustMultiRegressor* and *AttentiveFPModel*, respectively.

Predictions performed for the redox potentials have similar accuracy to the training with *TextCNNModel*. Interestingly, the potentials in the lower range (below -2 V) are more scattered by the *AttentiveFPModel*. However, the MAE for the test set, which is the measure of the accuracy of the predictions is almost invariable: 0.061 ± 0.002 V and 0.065 ± 0.006 V for *RobustMultiRegressor* and *AttentiveFPModel* trainings, respectively.

The distribution of the absolute error (see Figure 8) is slightly higher for *AttentiveFPModel* as only 96.6% of the molecules are in the error range of 0-0.19 V. Nevertheless the MAE for more than 98.0% of the molecules is less than 0.24 V.

The overall performance of the three models can be considered very similar (see Table III). The MAE for the test set shows very little variance, the averaged RMSE calculated for the *AttentiveFPModel* over the full data is relatively larger. In terms of computational costs for the predictions there are no significant differences for the three methods. The featurizations for over 6000 molecules are in the range of seconds to less than 2 minutes, while the whole training and predictions are of the magnitude of few hours. The *RobustMultiRegressor* model, which has the lowest MAE is the fastest of all, the training takes only 2-5 minutes. With proper optimization of the model hyperparameters, the predictions are expected to reach or even outperform the chemical accuracy.

CompBat project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 875565. This document has been produced by the CompBat project. The content in this document represents the views of the authors, and the European Commission has no liability in respect of the content.

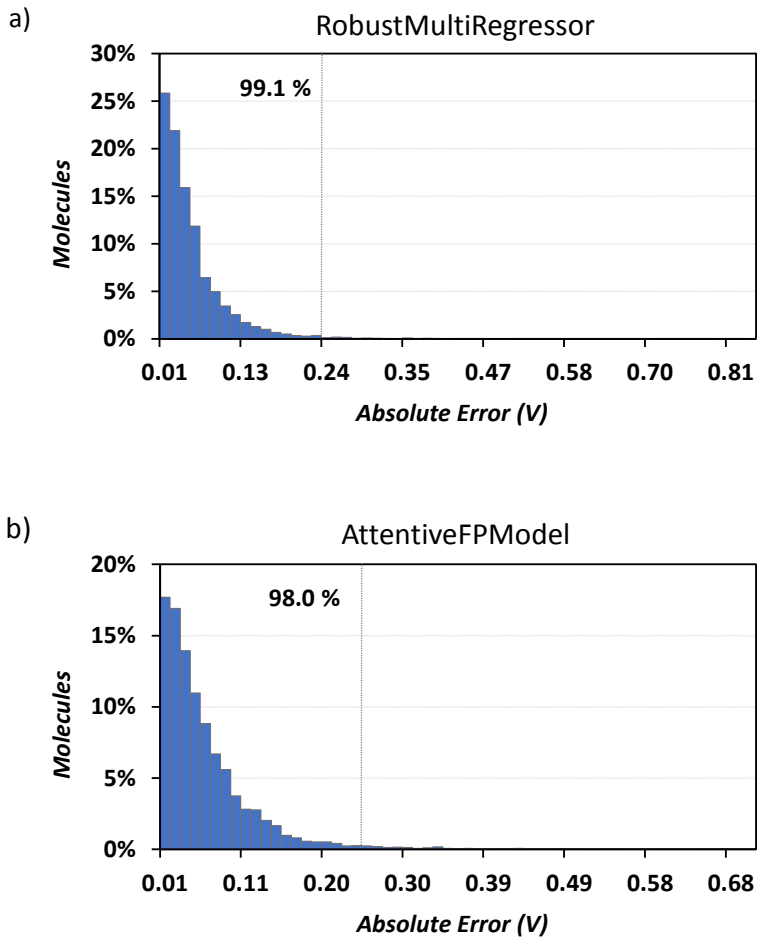

Figure 8. Distribution over the absolute error of redox potentials as obtained by a) *RobustMultiRegressor* and b) *AttentiveFPModel* predictions. The amount of molecules represented by % are shown in the vertical axis.

Table III: Performance of ML trainings for redox potential (units in V). The notations refer to the considered sets: Training (Tr), Validation (V) and Test (Te). The averages for each model are computed over the 5 individual trainings.

Model	MAE Te	RMSE Tr + V + Te	R^2 Tr + V + Te
TextCNNModel	0.065 ± 0.005	0.062 ± 0.005	0.984 ± 0.001
RobustMultiRegressor	0.061 ± 0.002	0.063 ± 0.006	0.981 ± 0.0003
AttentiveFPModel	0.065 ± 0.006	0.08 ± 0.006	0.972 ± 0.004

All three *DeepChem* models were applied for predicting the solubilities. The scatter plot of the solvation free energies predicted by *TextCNNModel* versus the true computed values is shown in Figure

CompBat project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 875565. This document has been produced by the CompBat project. The content in this document represents the views of the authors, and the European Commission has no liability in respect of the content.

9. The correlation is notably high and in the high solubility range (-200/-300 kcal/mol) the model is able to predict in very high accuracy, with very low scattered values. More than 99% of molecules do not reach the absolute error value 10 kcal/mol.

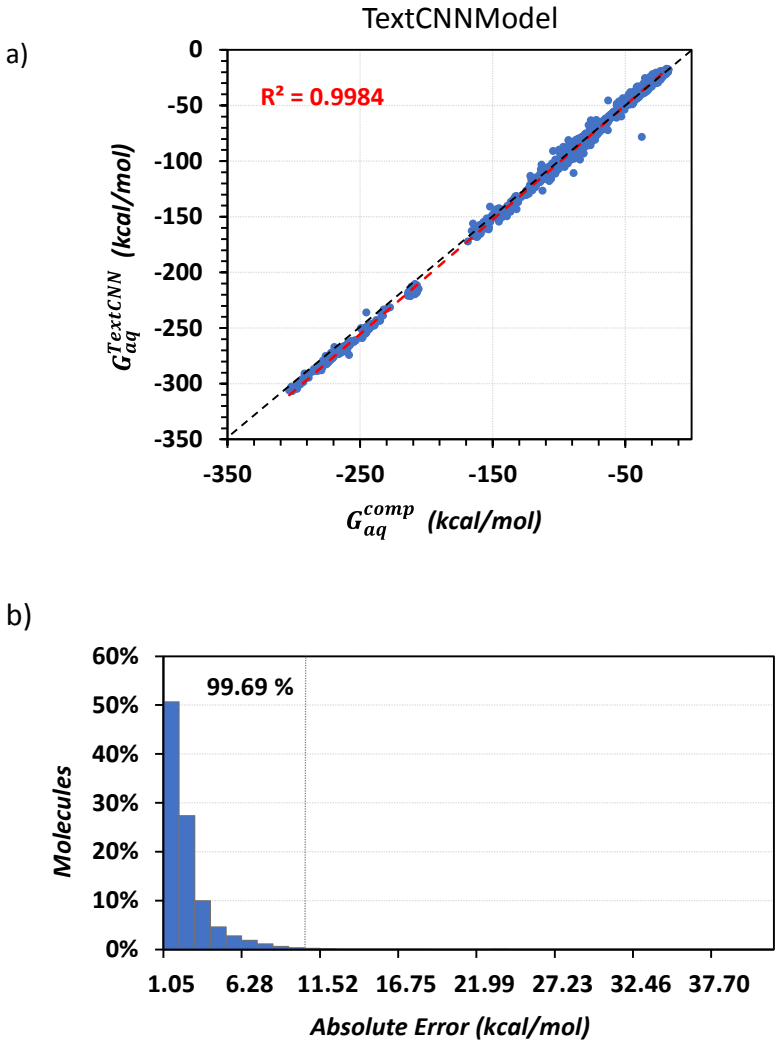

Figure 9. Solubilities as obtained by TextCNN modelling. a) Scatter plot between the DFT-computed versus *TextCNNModel*-predicted solvent corrections. b) Distribution over the absolute error of predicted solubilities. The amount of molecules represented by % are shown in the vertical axis.

Comparison of performance for predicting the free solvation energies is presented in Table IV. The coefficient R^2 is above 0.99 in all three cases. The MAEs for the test sets are rather small, they are not larger than 3 kcal/mol. Given that the values of the solvation free energies of the molecules can reach hundreds of kcal/mol this small MAE confirms that all three models are capable of very high predictive accuracy.

CompBat project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 875565. This document has been produced by the CompBat project. The content in this document represents the views of the authors, and the European Commission has no liability in respect of the content.

Table IV: Performance of ML trainings for solubilities (units in kcal/mol). The notations refer to the considered sets: Training (Train), Validation (Val) and Test (Test). The averages for each model are computed over the 5 individual trainings.

Model	MAE Test	RMSE Train + Val + Test	R^2 Train + Val + Test
TextCNNModel	1.9 ± 0.3	2.5 ± 0.2	0.998 ±0.0004
RobustMultiRegressor	1.8 ± 0.1	2.3 ± 0.1	0.998 ± 0.0002
AttentiveFPModel	2.5 ± 0.2	3.6 ± 0.4	0.994 ± 0.0012

4.3 Results of the 3DGCN analysis

The *3DGCN* model was applied to database **DB-II**. 5 independent trainings were performed splitting the dataset into training, validation and test sets in 80-10-10% ratio. The trainings were performed using the hyperparameters optimized in the reference work [13]. The batch size was set to 16, and the size of the convolutional filter was 128 (further details are given in Appendix 5).

The mean error metrics of separate runs shows only a slight variation (Table V). The MAE for the test set is 0.080 ± 0.001 V, which is slightly higher than those obtained with the previous deep-learning models, but still smaller than the DFT values benchmarked against experimental results (MAE = 0.24 V). The RMSE value for the test set is 0.110 ± 0.003 V.

Table V: Error metrics of redox potentials predicted by *3DGCN* for 5 separate trainings (units in V).

	MAE Train	MAE Val	MAE Test	RMSE Train	RMSE Val	RMSE Test
train 1	0.062	0.075	0.079	0.083	0.103	0.108
train 2	0.058	0.073	0.081	0.080	0.101	0.114
train 3	0.052	0.081	0.080	0.070	0.112	0.108
train 4	0.048	0.072	0.078	0.065	0.101	0.108
train 5	0.055	0.080	0.079	0.074	0.114	0.109
average	0.055±0.005	0.077±0.004	0.080±0.001	0.075±0.007	0.107±0.005	0.110±0.003

The scatter plot between the DFT-computed versus 3DGCN-predicted redox potentials and the distribution over the absolute error of predicted redox potentials for an arbitrarily chosen training set is presented in Figure 10. The correlation of determination R^2 is 0.94, the scatter of the predicted versus computed values is higher in the more negative region of the redox potential, similarly to those found with the other ML models.

CompBat project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 875565. This document has been produced by the CompBat project. The content in this document represents the views of the authors, and the European Commission has no liability in respect of the content.

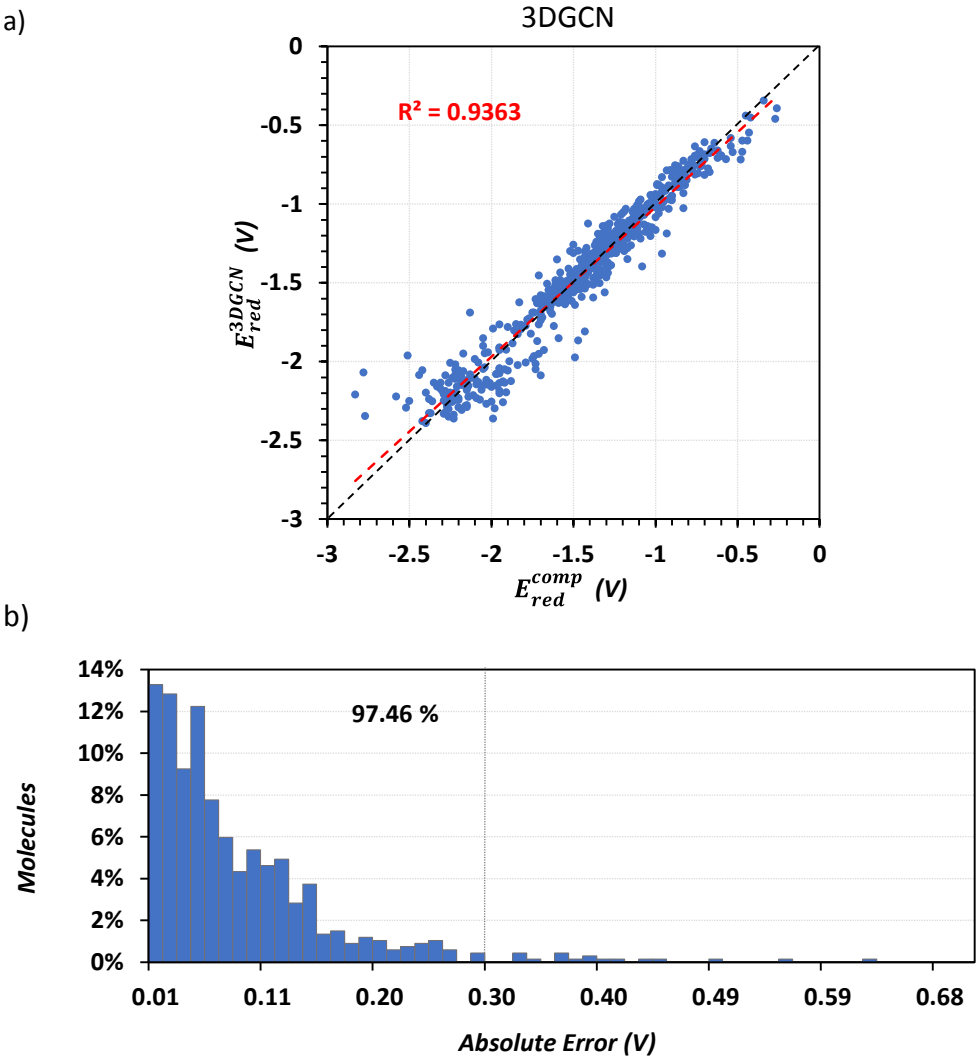

Figure 10. Redox potentials as obtained by *3DGCN* modelling. a) Scatter plot between the DFT-computed versus *3DGCN*-predicted redox potentials. b) Distribution over the absolute error of predicted redox potentials. The amount of molecules represented by % are shown in the vertical axis.

The accuracy of prediction for the solubilities is much better than for the redox potentials as illustrated in Figure 11. R^2 is above 0.99 and the scatter of the predicted values versus full computed ones is very narrow. The trend line for all the data is almost fully overlapping with the ideal correlation show in black dashed line in Figure 11a. The MAE for the majority of molecules is below 11 kcal/mol.

CompBat project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 875565. This document has been produced by the CompBat project. The content in this document represents the views of the authors, and the European Commission has no liability in respect of the content.

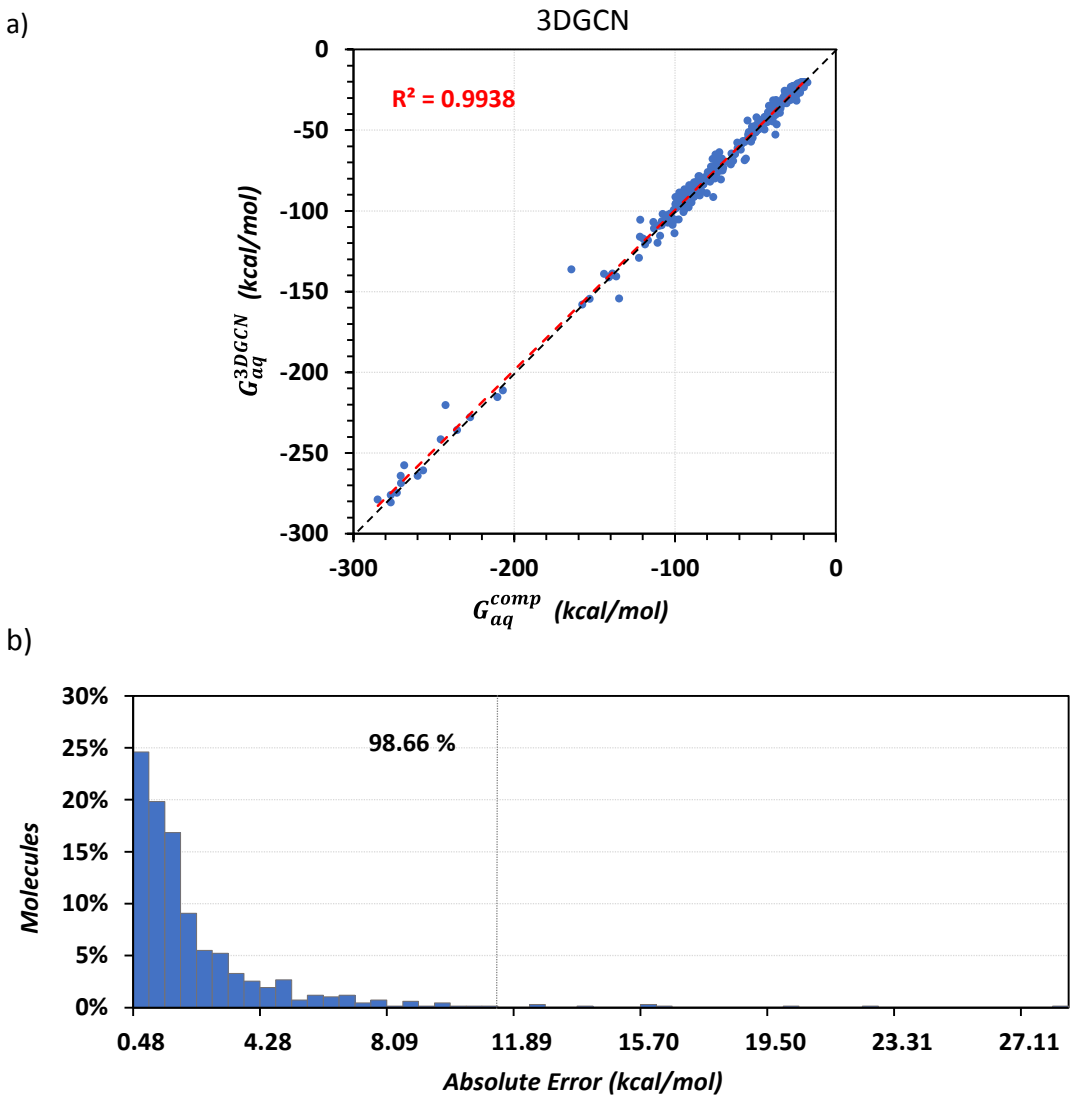

Figure 11. Solubilities as obtained by *3DGCN* modelling. a) Scatter plot between the DFT-computed versus *3DGCN*-predicted solvent corrections. b) Distribution over the absolute error of predicted solubilities. The amount of molecules represented by % are shown in the vertical axis.

Clearly the *3DGCN* is more suitable for predicting the free solvation energies, the overall R^2 is 0.993 ± 0.001, while in the case of redox potentials the expected value for R^2 is smaller and the standard deviation is larger. Even so, the expected value of MAE for the predicted redox potentials for the test sets can be considered good enough as compared to the MAE of DFT values benchmarked against experimental results, or xtb-computed values against the DFT values.

The overall performance of *3DGCN* is summarized in Table VI.

CompBat project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 875565. This document has been produced by the CompBat project. The content in this document represents the views of the authors, and the European Commission has no liability in respect of the content.

Table VI: Performance of *3DGCN* trainings for redox potentials and solubilities (units in V and kcal/mol, respectively). All values are reported for the test sets. The averages for each model are computed over the 5 individual trainings.

3DGCN	MAE	RMSE	R^2
Redox potential	0.080 ± 0.001	0.110 ± 0.003	0.937 ± 0.004
Solubilities	2.5 ± 0.4	3.5 ± 0.5	0.994 ± 0.002

4.4 SHAP feature attribution analysis

Shapley additive explanations (SHAP) introduced by Lundberg and Lee [17] were employed as consistent and robust feature attributions assessing which molecular substructures are most impactful when forming redox potential and solubility predictions. The SHAP values are based on the classical Shapley values [18] known from co-operative game theory, solving the problem of fair credit distribution among players participating in a collaborative game. In the present context, the game is the formation of a ML model output and the co-operating players are the specified features, i.e. the molecular substructures encoded by the ECFP4 fingerprints. In this work, the SHAP feature attribution analysis was implemented using the *shap* Python package [17] employing the efficient tree-specific version of the algorithm (TreeSHAP) [19]. Leveraging the above described nested CV procedure, each pyridoxal redox potential and solubility prediction was explained sequentially in the outer CV loop by attributing SHAP values (local importances) to the respective input features and aggregating the results for each fold.

Shapley additive explanations are computed in the outer CV loop. Thus, feature importances are attributed to each test set sample in a sequential manner, looping through the full dataset over the course of the nested CV. Considering first the redox potentials, a global summary of the feature importances is given in Figure 12a, where the mean magnitude of the SHAP values over the whole dataset are illustrated for the ten most impactful features. Figure 12a demonstrates clearly that one feature stands out as particularly important for the formation of ML redox potential predictions, namely the molecular substructure encoded by the bit number 1010. Specifically, the absence of this feature increases the redox potential, while its presence has a decreasing impact on the prediction. Thus the correlation between the SHAP value and the feature value is negative. The other features having a pronounced impact on the redox potential predictions have a significantly smaller global importance and the differences fall in many cases within the error bars.

CompBat project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 875565. This document has been produced by the CompBat project. The content in this document represents the views of the authors, and the European Commission has no liability in respect of the content.

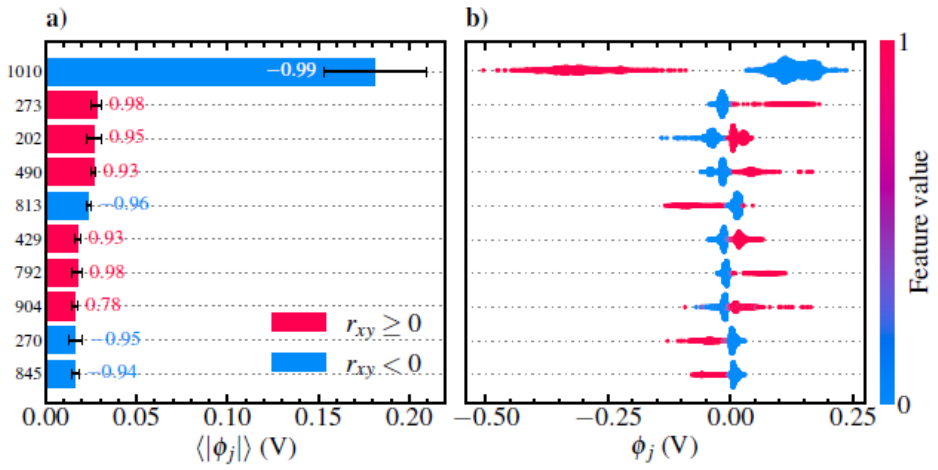

Figure 12. a) Mean magnitude of SHAP values for the ten most impactful molecular substructures as a measure of global importance in forming redox potential predictions. The average is taken for each feature over all test set samples in each outer CV fold and the error bars denote the standard deviation of the global measure over the outer CV folds. The red and blue color coding delineates whether the correlation between the marginal contribution (SHAP values) and the feature value is positive or negative, i.e. does the presence (absence) of a molecular substructure increase (decrease) the redox potential or vice versa. b) Violin-like scatter plots of SHAP values for the ten (globally) most important features and each sample molecule in the full pyridoxal redox potential dataset. The feature values are color coded as illustrated by the color bar and the vertical dispersion of SHAP values indicates the sample density, i.e. number of molecules with similar SHAP values for a given feature.

The global feature importance as gauged by the mean magnitude of the SHAP values gives a simple overview of the average importance of each feature. It is, however, informative to utilize the property of SHAP as a local feature attribution method and visualize the SHAP values separately for each sample molecule and feature in Figure 12b. The violin-like scatter plots show the distribution of the attributed importances for each molecule, and for feature 1010 the distribution shows a distinct binary separation where the presence of the molecular substructure consistently negatively shift the redox potential predictions, by even as much as -0.5 eV. On the other hand, the absence of the feature slightly more moderately, but still considerably, increases the model outputs. The local accuracy of the SHAP values demonstrates also nicely that for individual samples the attributed importances may attain rather large values also for those features reflecting a relatively small global importance, such as feature 904.

The molecular substructures encoded by the ECFP fingerprints can be explained using the bit information stored when mapping the fragments into the binary representation. Using example molecules, the most important substructures are highlighted in Figure 13. Interestingly, the highly impactful feature 1010 is found to correspond to an acyclic substituent Ar –CH2R with non-hydrogen neighboring substituents in both ortho positions. Notably, such moieties are found only in the pyr2 molecular set, where the amine side chain attached to the aromatic pyridoxal core satisfies this condition. The other globally important features affecting the redox potential predictions correspond to perfluorinated groups (273, 490, 429), alkoxides (202, 792), and cyclic aryl ether moieties (270). Also, arylamine and -amide groups (813, 845) are found within the ten note that the positive correlation

CompBat project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 875565. This document has been produced by the CompBat project. The content in this document represents the views of the authors, and the European Commission has no liability in respect of the content.

between the redox potential and the presence of perfluorinated groups can be explained by their character as electron-withdrawing groups (EWG), increasing the tendency of the molecule to be reduced. The same applies for nitro groups that can be found just below the top ten most important substructures. Conversely, the amine/amide features of the pyr2 set are characterized as electron-donating groups (EDG) with the opposite effect on the redox potential. The same applies for the cyclic aryl ether, while the positive correlation of the alkoxide groups is most likely an interaction effect with the perfluorinated substructures, as these are often present together as seen in Figure x. Otherwise, alcohols and alkoxides are generally considered EDGs. The most important features with the same correlation as the feature 1010, which is understandable considering that feature 1010 is connected to the amine side chains of the pyr2 molecular set. We note that the positive correlation between the redox potential and the presence of perfluorinated groups can be explained by their character as electron-withdrawing groups (EWG), increasing the tendency of the molecule to be reduced. The same applies for nitro groups that can be found just below the top ten most important substructures. Conversely, the amine/amide features of the pyr2 set are characterized as electron-donating groups (EDG) with the opposite effect on the redox potential [20]. The same applies for the cyclic aryl ether, while the positive correlation of the alkoxide groups is most likely an interaction effect with the perfluorinated substructures, as these are often present together. Otherwise, alcohols and alkoxides are generally considered EDGs.

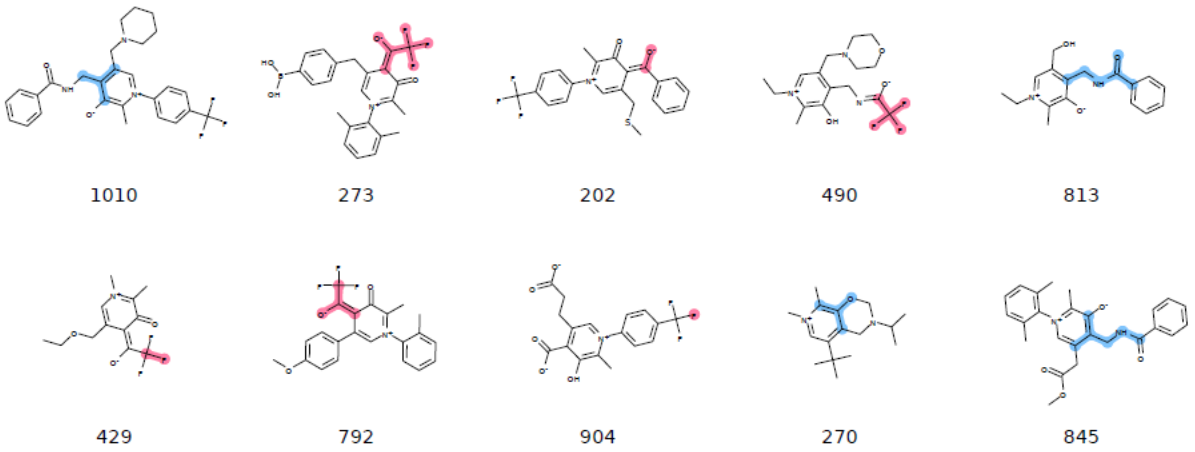

Figure 13. Ten globally most important molecular substructures for forming redox potential predictions. The illustrated molecules are examples and the bit encoded features are highlighted either in red or blue whether the presence of the features has an increasing (positive correlation) or decreasing (negative correlation) influence on the model output.

The same analysis as above is performed on the solubility predictions, revealing in Figure 14a a slightly more even global importance distribution than observed for the redox potential data. Still, one feature stands out as particularly important, namely feature 192. Considering the local feature attributions in Figure 14b, the presence of this feature may have a substantially negative impact on the solvation free energies, increasing the solubility by as much as -120 kcal/mol for some sample molecules. We note that only one feature (814) of the ten most impactful is seen to exhibit a positive

CompBat project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 875565. This document has been produced by the CompBat project. The content in this document represents the views of the authors, and the European Commission has no liability in respect of the content.

correlation, i.e. the presence of the feature increases the solvation free energy. Illustration of the most important substructures in Figure 15 unravels that the molecular solubility is mainly increased by the presence of phosphorus, sulfoxide, phosphoxide, carboxylate, aryl sulfide, oxygen and ammonium groups. In contrast, a protonation of the phosphoxide group yielding an uncharged phosphonic acid correlates with an increased solvation free energy, i.e. decreased solubility. These findings align well with fundamental chemical knowledge that charged and polar groups increase solubilities and gives support to the employed SHAP analysis as a reliable feature attribution method yielding ML prediction explanations consistent with human chemical intuition.

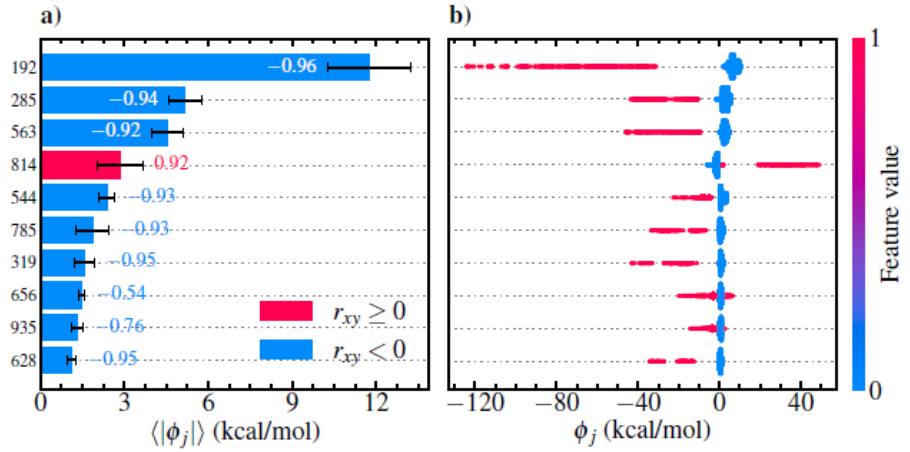

Figure 14. As Figure 11, but now for the solubility (solvation free energy).

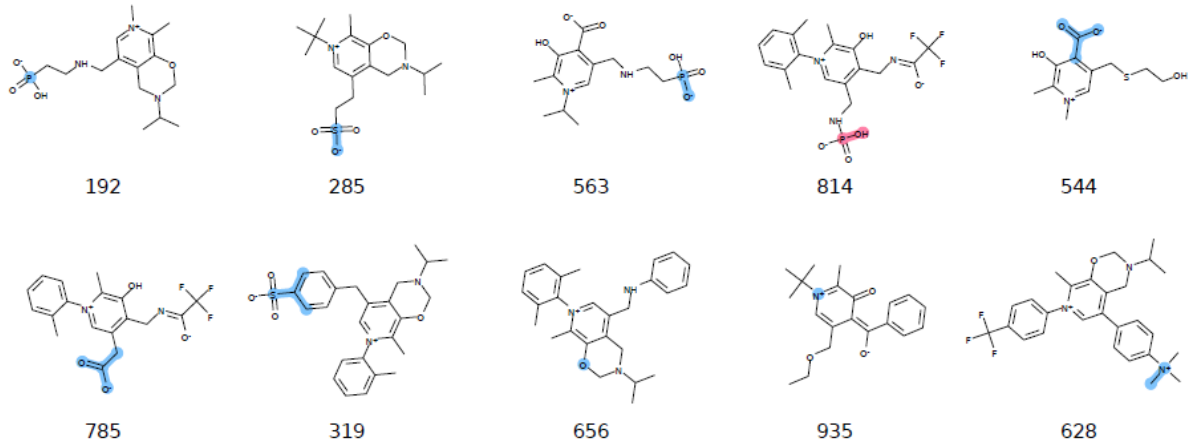

Figure 15. As Figure 12, but now for the solubility (solvation free energy).

The detailed local resolution of the SHAP feature attribution method is highlighted by decoupling feature contributions for individual model outputs. Such a “waterfall” plot is shown in Figure 16a for one example molecule and redox potential prediction from the pyr2 set. Here, the presence of the electron-donating amine side chain results in a net decrease of -0.24 V in the redox potential

CompBat project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 875565. This document has been produced by the CompBat project. The content in this document represents the views of the authors, and the European Commission has no liability in respect of the content.

prediction. Conversely, the electron-withdrawing perfluorinated side chain at the R_2 position increases the model redox potential output by as much as 0.78 V. Including the net contribution of -0.08 V from the remaining 1015 features results in a model output of -0.96 V, roughly 0.5 V more positive than the model base value.

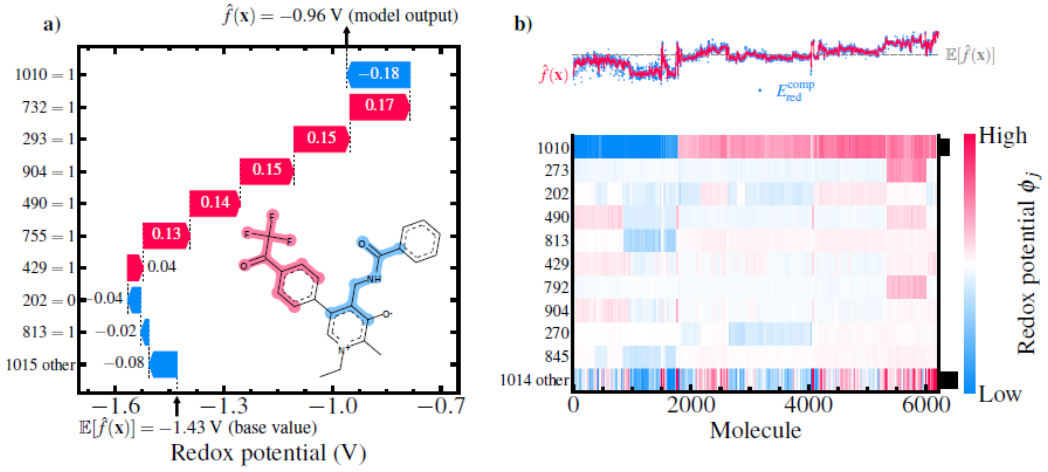

Figure 16. a) Waterfall plot decoupling individual feature contributions for an example pyr2 molecule and model redox potential output. The 9 locally most important features are explicitly shown while the net impact of the remaining 1015 substructures are combined. The model base value is shown below the plot, whereas the resulting output after adding the marginal contributions is shown on the top. b) Heatmap of SHAP values for the ten globally most impactful features and all sample molecules. The remaining 1014 features are shown aggregated on the bottom row. The model output after summing all SHAP values for each molecule, respectively, and including the model base value is plotted in red above the heatmap, while the blue dots indicate the true computational redox potentials. The model base value is indicated by the gray dashed line.

Multiple waterfall plots can be compactly visualized by a heatmap as shown in Figure 16b. Each pixel of the heatmap corresponds to an individual sample molecule and feature combination, color coded by the corresponding SHAP value. The elements are further clustered such that instances with similar explanations (SHAP values) are grouped. Thus, all pyr2 molecules are illustrated to the left in the heatmap, exhibiting more negative redox potentials than the model base value. Molecules with perfluorinated EWGs (273, 490, 792) and lacking feature 1010 are in contrast shown to the right, accordingly reflecting more positive redox potentials as shown by the plotted model predictions and the true computational values above the heatmap.

Analogous visualizations are presented for the solubility data in Figures 17a and 17b, now fora molecule from the pyr3 set. Here, the presence of the charged phosphate group decisively increases the solubility by more than -110 kcal/mol, while the heterocyclic amino ether substructure decreases the solubility by more than 21 kcal/mol. Including the net solvation energy increase of 13.4 kcal/mol caused by the 1015 other features yields the solubility prediction -122.7 kcal/mol. Again, the waterfall plots can be combined as illustrated in Figure 16b. Clearly, the most negative solubilities are clustered top right in the heatmap, where the presence of phosphorus, sulfoxide and phosphoxide groups have a significantly decreasing contribution to the solvation free energy predictions.

CompBat project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 875565. This document has been produced by the CompBat project. The content in this document represents the views of the authors, and the European Commission has no liability in respect of the content.

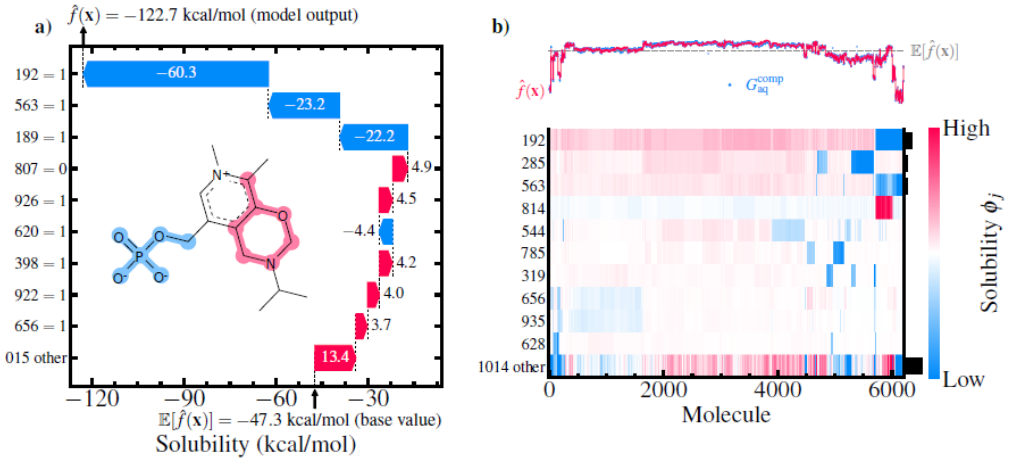

Figure 17. As Figure 15, but now for the solubility (solvation free energy) and a molecule from the pyr3 set.

5 Conclusions

Several machine learning (ML) techniques, including the commonly used random forest algorithm and deep-learning convolutional neural networks, were applied to molecular libraries developed within the CompBat project to assess their performance for predicting reduction potentials and aqueous solubilities of pyridoxal derivatives considered as potential candidates for new generation redox flow batteries. We find that all ML models tested in our present work perform remarkably well exceeding the accuracy of the quantum chemical computational protocol used to generate the pyridoxal databases.

In the spirit of explainable artificial intelligence [21], we applied a robust feature attribution methodology to unravel the contribution of different molecular substructures to the redox potential and solubility predictions. The analysis indicates that electron-withdrawing groups, such as perfluorinated moieties, have a decisively increasing effect on the redox potentials, while the presence of electron-donating amine/amide side chains have a clearly decreasing effect on the predicted redox potentials.

The presented ML analysis demonstrates an efficient methodology for high-throughput screening of RFB compounds, which will be further exploited within the CompBat project.

CompBat project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 875565. This document has been produced by the CompBat project. The content in this document represents the views of the authors, and the European Commission has no liability in respect of the content.

6 References

[1] The new version of the *crest* module released on 1. February 2021. For key references on the semiempirical quantum mechanical methods GFNn-xTB and the CREST method, see: (a) Grimme, S.; Bannwarth, C.; Shushkov, P. A Robust and Accurate Tight-Binding Quantum Chemical Method for Structures, Vibrational Frequencies, and Noncovalent Interactions of Large Molecular Systems Parameterized for All spd-Block Elements (Z = 1-86), *J. Chem. Theory Comput.* 2017, *13*, 1989-2009; (b) Bannwarth, C.; Ehlert, S.; Grimme, S. GFN2-xTB—An Accurate and Broadly Parametrized Self-Consistent Tight-Binding Quantum Chemical Method with Multipole Electrostatics and Density-Dependent Dispersion Contributions, *J. Chem. Theory Comput.* 2019, *15*, 1652–1671; (c) Grimme, S. Exploration of Chemical Compound, Conformer, and Reaction Space with Meta-Dynamics Simulations Based on Tight-Binding Quantum Chemical Calculations, *J. Chem. Theory Comput.*, 2019, *155*, 2847-2862.

[2] Landrum, G. RDKit: Open-source cheminformatics. 2006; https://www.rdkit.org.

[3] Weininger, D. SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules, *J. Chem. Inf. Comput. Sci.* 1988, *28*, 31-36.

[4] Rogers, D.; Hahn, M. Extended-connectivity fingerprints. *J. Chem. Inf. Model.* 2010, *50*, 742–754.

[5] Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V., et al. Scikit-learn: Machine learning in Python. *J. Mach. Learn Res.* 2011, *12*, 2825–2830; https://scikit-learn.org.

[6] Breiman, L. Random forests. *Mach. Learn.* 2001, *45*, 5–32.

[7] https://www.v7labs.com/blog/convolutional-neural-networks-guide

[8] Ramsundar, B.; Eastman, P.; Walters, P.; Pande, V.; Leswing, K.; Wu, Z. Deep Learning for the Life Sciences: Applying Deep Learning to Genomics, Microscopy, Drug Discovery, and More, https://www.amazon.com/Deep-Learning-Life-Sciences-Microscopy/dp/1492039837.

[9] https://colab.research.google.com/notebooks/

[10] Yoon K. Convolutional neural networks for sentence classification. arXiv preprint arXiv:1408.5882 (2014).

[11] Ramsundar, B.; Liu, B.; Wu, Z.; Verras, A.; Tudor, M.; Sheridan, R. P.; Pande, V. Is multitask deep learning practical for pharma? *J. Chem. Inf. Model.* 2017, *57*, 2068−2076.

[12] Xiong, Z.; Wang, D.; Liu, X.; Zhong, F.; Wan, X.; Li, X.; Li, Z.; Luo, X.; Chen, K.; Jiang, H.; Zheng, M. Pushing the Boundaries of Molecular Representation for Drug Discovery with the Graph Attention Mechanism. *J. Med. Chem.* 2020, *63*, 8749–8760.

[13] Cho, H.; Choi, I. S. Three-Dimensionally Embedded Graph Convolutional Network (3DGCN) for Molecule Interpretation. *ChemMedChem,* 2019, arXiv:1811.09794.

[14] https://www.jetbrains.com/pycharm/

[15] https://docs.conda.io/en/latest/

[16] https://github.com/blackmints/3DGCN#readme

[17] Lundberg, S. M.; Lee, S.-I. A unified approach to interpreting model predictions. *Adv. Neural Inf. Process Syst.* 2017; 4765–4774; https://shap.readthedocs.io.

[18] Shapley, L. S. A value for n-person games. *Contributions to the Theory of Games* 1953, *2*, 307–317.

CompBat project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 875565. This document has been produced by the CompBat project. The content in this document represents the views of the authors, and the European Commission has no liability in respect of the content.

[19] Lundberg, S. M.; Erion, G.; Chen, H.; DeGrave, A.; Prutkin, J. M.; Nair, B.; Katz, R.; Himmelfarb, J.; Bansal, N.; Lee, S.-I. From local explanations to global understanding with explainable AI for trees. *Nat. Mach. Intell.* 2020, *2*, 2522–5839.

[20] Assary, R. S.; Brushett, F. R.; Curtiss, L. A. Reduction potential predictions of some aromatic nitrogen-containing molecules. *RSC Adv.* 2014, *4*, 57442–57451.

[21] Gunning, D.; Stefik, M.; Choi, J.; Miller, T.; Stumpf, S.; Yang, G.-Z. XAI-Explainable artificial intelligence. *Sci. Robot.* 2019, *4,* eaay7120.

CompBat project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 875565. This document has been produced by the CompBat project. The content in this document represents the views of the authors, and the European Commission has no liability in respect of the content.

compbat

Appendix 1: Consortium

COMPUTER AIDED DESIGN FOR NEXT GENERATION FLOW BATTERIES
COMPBAT

List of participants

Participant No.	Participant organisation name	Country
1 (Coordinator)	**Aalto Korkeakoulusaatio sr** Aalto University (Aalto)	Finland
2	**Természettudományi Kutatóközpont** Research Centre for Natural Sciences (TTK)	Hungary
3	**Uppsala Universitet** Uppsala University (UU)	Sweden
4	**Universita Di Pisa** Pisa University (UNIPI)	Italy
5	**Skolkovo Institute of Science and Technology** (SKOLTECH)	Russia
6	**Jyvaskylan Yliopisto** University of Jyväskylä (JYU)	Finland
7	**Turun Yliopisto** University of Turku (UTU)	Finland

CompBat project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 875565. This document has been produced by the CompBat project. The content in this document represents the views of the authors, and the European Commission has no liability in respect of the content.

compbat

Appendix 2. Database in tabulated form

The computed reduction potentials and solubility data are provided in separate Excel documents (*DB-II-pot.csv* and *DB-II-solv.csv*). The first few lines of these tables are shown below for illustration. The reduction potentials are given in V, the Gibbs free energies of solvation are in kcal/mol, SMILES strings and the number of heavy atoms are in the last two columns.

Table DB-II-pot.csv

	NR	POT	R3	R1	R2	CHG	SYS	SMILES	NUMHEAVY
0	1	-1.31	COOH	P1	a1	0	pyr1	Cc1c[n+](C)c(C)c(O)c1C(=O)[O-]	13
1	2	-1.26	COOH	P1	a2	0	pyr1	CCc1c[n+](C)c(C)c(O)c1C(=O)[O-]	14
2	3	-1.36	COOH	P1	a3	0	pyr1	Cc1c(O)c(C(=O)[O-])c(C(C)C)c[n+]1C	15
3	4	-1.37	COOH	P1	a4	0	pyr1	Cc1c(O)c(C(=O)[O-])c(C(C)(C)C)c[n+]1C	16
4	5	-1.28	COOH	P1	a5	0	pyr1	CCCc1c[n+](C)c(C)c(O)c1C(=O)[O-]	15
5	6	-1.28	COOH	P1	a6	0	pyr1	Cc1c(O)c(C(=O)[O-])c(CC(C)C)c[n+]1C	16
6	7	-1.34	COOH	P1	a7	0	pyr1	Cc1c(O)c(C(=O)[O-])c(CC(C)(C)C)c[n+]1C	17
7	8	-1.27	COOH	P1	a8	0	pyr1	CCCCc1c[n+](C)c(C)c(O)c1C(=O)[O-]	16
8	9	-1.2	COOH	P1	b1	0	pyr1	Cc1c(O)c(C(=O)[O-])c(-c2ccccc2)c[n+]1C	18
9	10	-1.22	COOH	P1	b2	0	pyr1	Cc1ccc(-c2c[n+](C)c(C)c(O)c2C(=O)[O-])cc1	19
10	11	-1.21	COOH	P1	b3	0	pyr1	CCc1ccc(-c2c[n+](C)c(C)c(O)c2C(=O)[O-])cc1	20
11	12	-1.22	COOH	P1	b4	0	pyr1	Cc1c(O)c(C(=O)[O-])c(-c2ccc(C(C)C)cc2)c[n+]1C	21
12	13	-1.21	COOH	P1	b5	0	pyr1	Cc1c(O)c(C(=O)[O-])c(-c2ccc(C(C)(C)C)cc2)c[n+]1C	22

Table DB-II-sol.csv

	NR	Solv-XTB	R3	R1	R2	CHG	SYS	SMILES	NUMHEAVY
0	1	-37.7572	COOH	P1	a1	0	pyr1	Cc1c[n+](C)c(C)c(O)c1C(=O)[O-]	13
1	2	-38.7137	COOH	P1	a2	0	pyr1	CCc1c[n+](C)c(C)c(O)c1C(=O)[O-]	14
2	3	-39.2184	COOH	P1	a3	0	pyr1	Cc1c(O)c(C(=O)[O-])c(C(C)C)c[n+]1C	15
3	4	-39.6737	COOH	P1	a4	0	pyr1	Cc1c(O)c(C(=O)[O-])c(C(C)(C)C)c[n+]1C	16
4	5	-38.8333	COOH	P1	a5	0	pyr1	CCCc1c[n+](C)c(C)c(O)c1C(=O)[O-]	15
5	6	-39.0885	COOH	P1	a6	0	pyr1	Cc1c(O)c(C(=O)[O-])c(CC(C)C)c[n+]1C	16
6	7	-39.0366	COOH	P1	a7	0	pyr1	Cc1c(O)c(C(=O)[O-])c(CC(C)(C)C)c[n+]1C	17
7	8	-39.0988	COOH	P1	a8	0	pyr1	CCCCc1c[n+](C)c(C)c(O)c1C(=O)[O-]	16
8	9	-40.4976	COOH	P1	b1	0	pyr1	Cc1c(O)c(C(=O)[O-])c(-c2ccccc2)c[n+]1C	18
9	10	-40.9504	COOH	P1	b2	0	pyr1	Cc1ccc(-c2c[n+](C)c(C)c(O)c2C(=O)[O-])cc1	19
10	11	-41.0157	COOH	P1	b3	0	pyr1	CCc1ccc(-c2c[n+](C)c(C)c(O)c2C(=O)[O-])cc1	20
11	12	-41.3542	COOH	P1	b4	0	pyr1	Cc1c(O)c(C(=O)[O-])c(-c2ccc(C(C)C)cc2)c[n+]1C	21
12	13	-41.3956	COOH	P1	b5	0	pyr1	Cc1c(O)c(C(=O)[O-])c(-c2ccc(C(C)(C)C)cc2)c[n+]1C	22

CompBat project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 875565. This document has been produced by the CompBat project. The content in this document represents the views of the authors, and the European Commission has no liability in respect of the content.

Appendix 3: Nested cross-validation

A schematic illustration of the implemented nested cross-validation workflow is illustrated in Figure A.1. In the outer loop, the data is split into *k* (stratified) folds. Each fold is in turn selected as a held-out test (*test*) set while the remaining folds form the training and validation (*trainval*) set. The formed *trainval* set is split in the inner loop into *l* (stratified) folds. Each fold is in turn selected as a held-out validation (*val*) set, while the remaining folds are combined into a training (*train*) set on which the model is trained using different hyperparameter settings. Each model is validated on the *val* set and the defined loss function is evaluated and averaged across all inner CV folds to determine the best hyperparameter combination. With the best hyperparameters, the model is trained on the full *trainval* set and finally tested on the outer loop *test* set withheld before entering the inner loop. Averaging the model loss on the test set over all outer CV folds yields an unbiased performance estimation of the employed ML algorithm (RF) including error bars reflecting the algorithm stability with respect to different input data. Herein, SHAP values for each test set sample are also computed in the outer CV loop, thus enabling an exhaustive analysis and explanation of all predictions for each molecule in the full dataset. Importantly, feature importances are always evaluated on samples unseen by a particular model, thus avoiding bias associated with the trainingset instances to which the models could theoretically overfit.

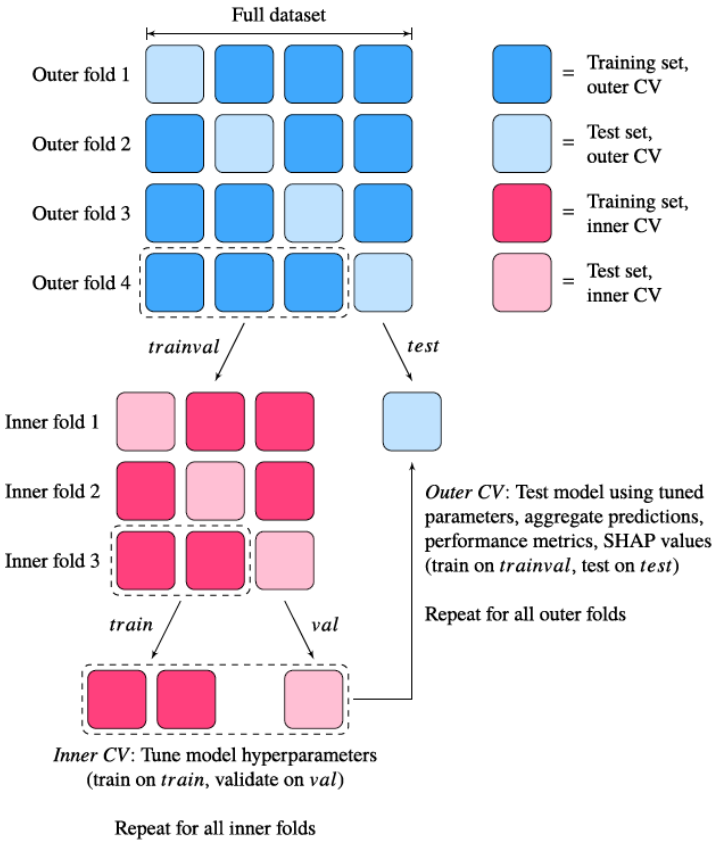

Figure. A.1: Nested cross-validation exemplified for the case of 4x3-fold CV.

CompBat project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 875565. This document has been produced by the CompBat project. The content in this document represents the views of the authors, and the European Commission has no liability in respect of the content.

Appendix 4: Hyperparameters of DeepChem methods

For ML models used via DeepChem the default hyperparameters were utilized, which are listed below.

TextCNN:

n_embedding: 75(deafault) - Length of embedding vector.

num_filters(list): [100, 200, 200, 200, 200, 100, 100, 100, 100, 100, 160, 160] (deafault) – Properties of filters used in the conv net (The number of filters is the number of neurons, since each neuron performs a different convolution on the input to the layer).

dropout(list or float): 0.25(deafault) - Dropout rate (Dropout is a technique where randomly selected neurons are ignored during training). The length of this list should equal len(layer_sizes). Alternatively this may be a single value instead of a list, in which case the same value is used for every layer.

mode: regression

AttentiveFPModel:

num_layers: 2(deafault) – Number of graph neural network layers, i.e. number of rounds of message passing.

num_timesteps: 2(deafault) – Number of time steps for updating graph representations with a GRU(Gated recurrent unit).

graph_feat_size: 200(deafault) – Size for graph representations.

dropout: 0(deafault) – Dropout probability. The length of this list should equal len(layer_sizes). Alternatively this may be a single value instead of a list, in which case the same value is used for every layer.

mode: regression(deafault)

number_atom_features:30(default, from featurizer) – The length of the initial atom feature vectors.

number_bond_features: 11(default, from featurizer) – The length of the initial bond feature vectors.

self_loop: True(default) – Whether to add self loops for the nodes, i.e. edges from nodes to themselves. When input graphs have isolated nodes, self loops allow preserving the original feature of them in message passing.

RobustMultitaskRegressor

layer_sizes(list): [1000](default) – the size of each dense layer in the network. The length of this list determines the number of layers.

weight_init_stddevs: 0.02(default) – the standard deviation of the distribution to use for weight initialization of each layer. The length of this list should equal len(layer_sizes). Alternatively this may be a single value instead of a list, in which case the same value is used for every layer.

CompBat project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 875565. This document has been produced by the CompBat project. The content in this document represents the views of the authors, and the European Commission has no liability in respect of the content.

bias_init_consts: 1.0(default) – the value to initialize the biases in each layer to. The length of this list should equal len(layer_sizes). Alternatively this may be a single value instead of a list, in which case the same value is used for every layer.

weight_decay_penalty: 0.0(default) – the magnitude of the weight decay penalty to use

weight_decay_penalty_type: l2(default) – the type of penalty to use for weight decay, either 'l1' or 'l2'

dropouts: 0.5(default) – the dropout probablity to use for each layer. The length of this list should equal len(layer_sizes). Alternatively this may be a single value instead of a list, in which case the same value is used for every layer.

activation_fns: tf.nn.relu(default) – the Tensorflow activation function to apply to each layer. The length of this list should equal len(layer_sizes). Alternatively this may be a single value instead of a list, in which case the same value is used for every layer.

bypass_layer_sizes: [100](default) – the size of each dense layer in the bypass network. The length of this list determines the number of bypass layers.

bypass_weight_init_stddevs: [0.02](default) – the standard deviation of the distribution to use for weight initialization of bypass layers. same requirements as weight_init_stddevs

bypass_bias_init_consts: [1.0](default) – the value to initialize the biases in bypass layers same requirements as bias_init_consts

bypass_dropouts: [0.5](default) – the dropout probablity to use for bypass layers. same requirements as dropouts

CompBat project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 875565. This document has been produced by the CompBat project. The content in this document represents the views of the authors, and the European Commission has no liability in respect of the content.

Appendix 5: Hyperparameters of the 3DGCN model

Hyperparameter optimization was carried out in 3DGCN studies. The training was performed with the optimized model, with the following hyperparameters, indicating the name of the parameters in the project in parentheses:

- Batch size (batch): 16
- Size of convolutional filter (units_conv): 128
- Size of filter in fully connected layers (units_dense): 128
- Number of convolutional layer (num_layers): 4
- Pooling technique: sum
- Fold of cross-validation: 5

CompBat project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 875565. This document has been produced by the CompBat project. The content in this document represents the views of the authors, and the European Commission has no liability in respect of the content.