Thalita Cirino do Nascimento

MODELLING THE MOLECULAR LIPOPHILICITY USING THE ONLINE CHEMICAL MODELING ENVIRONMENT (OCHEM)

This report presents a comprehensive account of the five-month internship project undertaken at the **Helmholtz Munich** under the supervision of **Dr. Igor Tetko**, serving as detailed documentation of the work completed during this period. It is a requisite for obtaining a **MSc. in Chemoinformatics and Physical Chemistry** degree within the Erasmus Mundus Joint Master's program of the *Università degli Studi di Milano* and *Université de Strasbourg*.

Université de Strasbourg

UNIVERSITAS STUDIORUM MEDIOLANENSIS

JUNE 2024

ACKNOWLEDGEMENTS

It is a deep pleasure for me to have a list of people to thank. First, I am very grateful to the Master's program coordinators for allowing me to resume my education four years after obtaining of my undergraduate degree, and to the European Union for financially supporting me during the entire duration of my master's studies.

Along the way, I have met talented and hard-working people who spare no effort in doing impeccable work. Starting with professors who have adapted their courses to accommodate international students in Milano, I would like to highlight a few names: Prof. Marco Vignati, Prof. Francesca Vasile, and Prof. Rocco Martinazzo. In the same context, I am also very grateful to my advisor Dr. Igor Tetko for his guidance, help, and patience during the preparation of this work and for the daily conversations during lunch and coffee breaks, which made my stay in Munich and learning much smoother and more pleasant.

This work was possible thanks to Dr. James Sangster, who compiled all the data, and Dr. Larisa Charochkina, who introduced it along with all the corresponding scientific articles into OCHEM.

I want to thank my colleagues (Xinyue, Ulrick and Leonardo) who accompanied me through this journey, whether it was hours spent in the library studying or at the bar hanging out. Also, many thanks to my close friends and family who are always there to help or just listen. And, finally, I want to thank myself for the tedious work of formatting the tables within this thesis.

TABLE OF CONTENTS

1 INTRODUCTION

The primary purpose of this project was to develop an accurate predictive model for molecular lipophilicity using the Online CHEmical Modeling environment (OCHEM) web-based platform.[1] The dataset used for the model training includes experimental octanol-water partition coefficient values for over 20k molecules collected by Dr. James Sangster. By leveraging the OCHEM platform and its integrated tools, we aimed to address the challenges associated with data quality, molecular representation standardisation, and the trade-off between model accuracy and interpretability. Through rigorous data preprocessing, exploration of various advanced machine learning methods based (mainly but not only) on representation learning by Deep Neural Networks (DNN), and consensus modelling approaches, the goal was also to create a reliable model capable of providing confidence estimations with well-defined applicability domain (AD). Below I present a literature review covering the main foundations of this project.

1.1 Background on molecular lipophilicity

The octanol-water partition coefficient — P_{ow} or K_{ow}, which is frequently used as $\log(P_{ow})$ for computational studies) — is the standard metric of molecular lipophilicity, *i.e.* the affinity of a molecule for an aqueous or lipophilic environment, as expressed in the equation 1.[2]

$$P_{ow} = \frac{[X^N]_{oct}}{[X^N]_{wat}} \tag{1}$$

This coefficient is determined for neutral species of the molecule ($\mathrm{X^N}$). In case a molecule is ionised, a closely related logD (octanol-water distribution coefficient) is reported, which accounts for the contribution of all the ionised species present at a given pH.[3] For a monoprotic acid or base, both coefficients can be related through a simple equation (2) derived from the Henderson-Hasselbalch's one:

$$D = \frac{P}{1 + 10^{\sigma(pH - pKa)}} \tag{2}$$

with **σ** taking the values +1 or -1 to account for acids and bases, respectively, where the ionisation constant (pK_a) is known and no partitioning of ionised species to octanol is considered.[3] Thus, logD has values less than or equal to logP, and if they are equal, it indicates that all the species are in their neutral form at the pH of the measurement. While equation 2 is suitable for molecules with one ionisable group, the relationship between logP and logD becomes more complex for molecules with multiple ionisable groups, and the distribution coefficient D accounts for the partition coefficient for the neutral species ($\mathrm{D^N}$) and all the n ionised species ($\mathrm{D^I_n}$) as shown in equation 3

$$D = D^N + D_1^I + D_2^I + \ldots + D_n^I \tag{3}$$

Also, the assumption that ionisable species do not partition to the octanol phase oversimplifies D and P relationship, and one should also account for it as described elsewhere.[4] Frequently logD is measured at physiological pH (around 7.4) for systemic circulation, but measurements at lower pH values, such as 2 or 3 to mimic gastric conditions, are also commonly reported. The accuracy of logP and logD interlaboratory measurements are estimated at 0.3-0.5 log units.[5]

Organic, environmental, and medicinal chemists consider logP and/or logD as key parameters. For organic chemists, choosing appropriate solvents for efficient dissolution, extraction, and purification of desired products during the synthetic process is crucial. The selection of solvents is directly driven by the solubility characteristics imparted by the compound's lipophilicity.[6] Environmental chemists monitor the lipophilicity of chemicals since it is closely related to the tendency for bioaccumulation and sorption to sediments.[7] This role of lipophilicity is important in assessing the environmental risks posed by chemicals and that is why logP is frequently used as a descriptor for modelling environmental properties of compounds, such as their biodegradability.[8,9] Moreover, along with polar surface area and molecular weight, the logP is an important physicochemical property frequently used to identify compounds with favourable pharmacokinetic properties such as drug-likeness and oral bioavailability.[10–12] Some popular criteria like Lipinski's Rule of Five (RO5) and the Ghose filter use calculated logP (clogP) values.[11,12] This is because molecular lipophilicity directly influences the compound's ADMETox (absorption, distribution, metabolism, excretion, toxicity) properties by contributing to its crossing through cell membranes, binding to proteins, and accumulation in adipose tissues.[11–13]

Although the importance of lipophilicity in the bioactivity of molecules was recognised more than 120 years ago through the pioneering work of Overton[14,15] and Meyer[16] on the relationships between olive oil/water partition coefficient and anaesthetic potency of chemicals, the wide interest of chemists in this topic can be traced back to the 1960s with the seminal publications of Hansch and Fujita on the relation of phenoxyacetic acid's physicochemical properties (including the partition coefficient) and its bioactivity,[17] which is considered to be the start of the quantitative structure-activity relationships (QSAR) era. Since then, molecular lipophilicity has been employed as a key property of drug candidates at the early discovery stage. However, experimentally measuring all the virtual chemical libraries partition coefficients is impossible, with a total chemical space estimated to comprise about 10^{60} theoretically accessible chemical compounds.[18] Therefore, reliable models for predicting molecular lipophilicity are much-demanded for modern drug discovery. In recent years there has been a great interest in developing such models.

1.2 Overview of classic methods to predict logP of compounds

The first method to estimate the partition coefficient from the chemical structure was published by Fujita et al.[19] It was limited to deriving a value from a scaffold whose logP value was previously known, based on substituent π-constants to correlate substituent effects. Subsequent to this work, numerous methods have been devised to predict $logP_{ow}$ values[20] and they are usually classified into two major categories: substructure- and property-based methods.[21–23]

1.2.1 Substructure-based methods

Substructure-based methods are typically known as “divide and conquer” approach. It comprises fragmental[24–26] and atom-based[27] methods. The first divides molecules into substructural fragments and uses additive fragment contributions along with correction factors to account for intramolecular interactions between fragments, while the second breaks down molecules to the atom level and then sums atomic contributions, usually correction factors for intramolecular interactions are not applied here.

1.2.2 Property-based methods

Property-based methods aim to directly model the underlying thermodynamic processes and physical properties that govern the partitioning of a molecule between the octanol and water phases. Instead of dismembering the molecule structurally, these methods use calculated molecular properties and solvation parameters to estimate the transfer of free energy between the two immiscible solvents. These methods are subdivided into: (i) empirical,[28] which uses experimental parameters like polarizability, hydrogen-bonding strength etc. that govern partitioning between octanol and water to derive empirical equations for logP; (ii) 3D structure-based[29] which relies on quantum or molecular mechanics (mainly but not only) to optimise molecular conformations and calculate lipophilicity-related properties; and (iii) topological descriptors-based relies on 2D descriptors like connectivity indices, E-state indices[30,31] etc. that encode size, shape and electronic properties to derive QSAR models for logP.

1.3 Deep Neural Networks and ensemble-based methods

Although traditional approaches are still widely used, they are increasingly being complemented or replaced by deep neural network (DNN) based methods trained on molecular fingerprints[32] or descriptors,[33] as well as representation (or feature) learning methods. The latter learns directly from internal features based on molecular representations such as SMILES[34] or molecular graphs (GNN)[35]. Previous studies[36] have reported that they are flexible and efficient in capturing complex relationships if given sufficient data. When combined into a consensus approach, NN-based methods have demonstrated a great potential to provide models with higher accuracy than individual models, as well as a broader applicability domain,[37,38] *i.e.* the part of the chemical space in which the model predictions are considered reliable.[39,40] Indeed, to predict a property accurately for a given chemical is possible only if the training set contains a representative set of compounds in the same or similar chemical classes.

1.4 Applicability domain and interpretation of models

Since it is unfeasible to predict logP with the same accuracy across the whole chemical space, more recently developed modelling approaches, such as those implemented into OCHEM, typically incorporate techniques to distinguish reliable predictions from unreliable ones.[38,41] These methods provide additional information to quantify the uncertainty associated with each prediction, enabling users to separate predictions that fall within an acceptable range of confidence from those that may be an extrapolation beyond the model's AD.

1.5 Overview of available data to predict logP of compounds

While many databases with experimental values for lipophilicity such as PhysProp[42] (13688 compounds), AstraZeneca[43] (4200 compounds), OCHEM[1] and ChEMBL[44] databases are publically available, the demand for larger and more diverse datasets remains high. Such datasets are also crucial for deep learning approaches, which typically require large amounts of data to effectively capture complex relationships and patterns. Moreover, part of these databases overlaps with the original compilation by Hansch et al.,[10] who applied their broad experience to carefully select and document the most reliable values from various literature sources.

The classic experimental methods for determining logP are direct approaches such as the shake flask technique[45] and slow-stirring method, which are recommended by OECD Test Guideline No. 107 and 123, respectively.[46,47] However, these methods are rather slow and not easily parallelised for large data volume data analysis, which is required by industry. The indirect methods based on the retention time data by chromatographic techniques[3] enable rapid high-throughput analysis, generating a substantial volume of data that exceeds those measured by other methods. The retention time has exhibited a strong correlation with the apparent partition coefficient, as demonstrated by Win et al.[48] and Wang et al.[49] who incorporated liquid chromatography retention times as descriptors to enhance the accuracy of logD predictions.

For this project, we have access to a comprehensive database containing over 40k experimental logP/logD values (over 25k molecules) curated from literature sources over many years by Dr. James Sangster and collaborators from Sangster Research Laboratories[50] and thus represents a valuable opportunity to leverage advanced machine learning techniques to develop accurate and robust predictive models for molecular lipophilicity.

2 DATA

2.1 Sangster dataset

This dataset, initially compiled by Dr. J. Sangster, has been significantly extended since its initial publication.[50] It contains references to original publications, and has a rich annotation of experimental conditions for each record, in particular pH of measurements. In total, it consists of 25874 molecules (42402 data points) with at least one experimental lipophilicity value collected from an extensive collection of literature, encompassing over 3k books and scientific articles published up to 2018. Among these molecules, the disconnected structures or salts (1433), charged and strong acids (47), with more than 50 non-hydrogen atoms (544), and metal-containing (800) molecules were filtered out thus decreasing the dataset to 23167 molecules (39007 records from which 4202 were duplicates). The filtering was done since logP modelling of non-organic compounds as well as mixtures and salts cannot be easily accomplished using the methods used in our study while charged structures were likely ionised throughout all the pH range, such as strong acids and quaternary ammonium compounds. The search and exclusion of such molecules was done with help of SMARTS structural alerts implemented in OCHEM software.[51]

2.2 Data error correction

Experimental values collected from literature could be noisy for several reasons beyond simple experimental errors. Firstly, there is a lack of standardisation in endpoint annotation, with different sources reporting them as ratio (P), logarithm (log P), or percentage (%), leading to potential confusion. OCHEM uses automatic unit conversion and once data are introduced with correct units of measurement, they are automatically converted to the unit used for model development, which was logP in our case.

To identify errors in data, an initial consensus model was built. For all molecules with large prediction errors the source literature was revisited to check whether the logD/logP_{ow} value, pH, chemical structures, etc. and corrections were performed where necessary. It should be noted that we did not exclude any compounds with large model errors notwithstanding whether we could correct them or not.

2.3 Separation of neutral and ionised compounds: logP and logD sets

While examining compounds we realised that the dataset could possibly contain a mixture of logP and logD values. The pH of measurement was reported for 41.3 % of data points and the distribution of reported pH values is shown in Figure 1. Duplicated measurements (same pH, molecule and logD value) were counted only once.

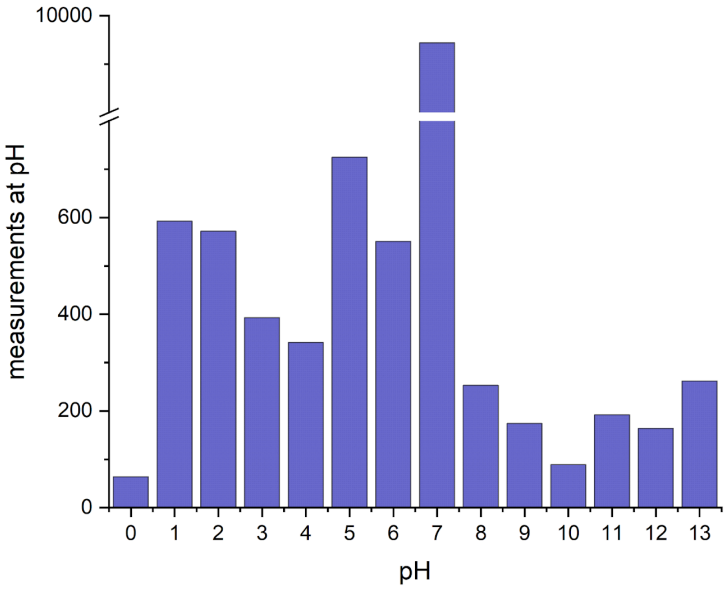

Figure 1. Distribution of pH-values for compounds with explicitly annotated pHs of measurement. The largest number of measurements (51.8%) was at pH 7.4.

Therefore, to ensure that the model is trained only with the intrinsic $logP_{ow}$ measures, the filtered data was split into a set with logP values (31048 data points, 19336 unique molecules) and a set with likely logD values (7954 data points, 5309 unique molecules). The split was based on predicted ionised state of compounds at the pH of measurements. Also a specific consideration was given to zwitterions. Although the overall charge is neutral, zwitterions are predominantly ionised in water at any pH. From the predicted pKa for each molecule, those with their most acidic pKa lower than 6 and their most basic pKa greater than 8 were classified as zwitterion (a total of 719 unique compounds) which were thus excluded from the logP set in order to verify how much they affect the statistical indicators of model quality (see Results).

All the records without an annotated pH were added to the logP set. Then, logD values were calculated across a pH range from 0 to 14 using ChemAxon calculator plugins (version 5.10.4) for those data points with annotated pH. If the pH at which the measurement was made fell within the range where the calculated logD was at its maximum, or had a value not less than 0.3 log units of the maximum (thus corresponding to the accuracy of experimental measurement), the experimental value was kept in the logP set, or, otherwise, it was saved into the logD set. This data analysis was implemented with Python 3.12 scripts and it is available on GitHub (https://github.com/t-cirino/master_thesis). The graphs plot in the Results section were done with OriginPro 10.1.5.132 version.

2.4 Correction of logD values ($logP_{adj}$ set)

The data collected by Dr. Sangster usually contained the pH at which the measurements were taken. In total, there were values for 5309 molecules that were not measured at the optimal pH (as predicted by ChemAxon) corresponding to the neutral form of molecules. These data points were removed from the logP set and formed the logD set. The compounds at logD set were ionised to a different degree and one could, in principle, convert logD values to the logP based on the equations reported in the Introduction.

However, to use these equations one also needs to know ionisation constants as well as partitioning coefficients of ionised species,[4] which were not available to us. While such calculations arenot difficult for compounds with only one ionisable group, the situation become much more complicated when compounds have several ionisable groups. Therefore, for this purpose we relied on ChemAxon software and used the difference between LogP and logD values predicted by ChemAxon (at the pH of measurement) to correct the experimental logD values to the octanol-partitioning coefficients using the following equation 4

$$\log P_{adj} = \log D_{exp} + \log D_{\mathrm{calc(max)}} - \log D_{\mathrm{calc(pHexp)}} \quad (4)$$

where $logP_{adj}$ corresponds to the adjusted experimental value, $logD_{exp}$ to the experimental value, $logD_{calc(max)}$, the maximum logD as predicted by ChemAxon across all pH (0-14) values, and $\mathrm{logD}_{\mathrm{calc(pHexp)}}$ to the calculated logD at the reported pH.

2.5 Time-split Test Set

An additional set was collected from ChEMBL and OCHEM. In both these databases we collected articles published after 2018 to create a prospective validation set. The same procedure as described above was used to split the dataset on logP and logD subsets containing 5443 and 152 molecules, respectively.

3 METHODS

This section describes the process of building a predictive model from the data preparation to model validation.

3.1 Data Preprocessing

The data quality is crucial in building accurate and reliable predictive models. This step was an extensive part of this work, consisting of molecular standardisation, verification of data in original data sources as well as separation of logP and logD values.

3.1.1 Molecular standardisation

Standardisation of chemical structure representation ensures coherence and uniformity when representing specific chemical groups for modelling approaches. OCHEM has a built-in workflow that was used to standardise the molecules before they were used to calculate descriptors and are used in other methods.

3.2 Model validation

Several machine learning techniques were evaluated using a 5-fold cross-validation (5CV) protocol. This validation consisted of repeating the entire model development process five times for different splits of the initial dataset into training and validation subsets. For each iteration, only the respective training subset (80% of the whole training set) was used to develop the model, while the validation subset (the remaining 20%) was used for evaluating the model's performance. Thus, the validation sets were never used to influence the model development process and thus were “blind” sets for each model developed with the respective training set. Moreover, as some molecules could have multiple measurements (records) in the training data set, OCHEM split data based on molecules (according to their first part of InChi keys which corresponds only to connectivity of molecules) rather than on data measurements (records). By skipping stereochemistry we thus ignored issues arising from possibly incorrect stereochemical information which would allow the same molecule to be part of the training and validation sets simultaneously. It should be mentioned that the use of 5CV corresponds to so-called “external cross-validation” and is different from so-called “internal validation”, where the validation procedure used to select model hyperparameters and/or descriptors could result in overfitting.[38] Actually, the latter term is more specific to the QSAR field and has in some cases led to overfitted results which gave rise to a scepticism about use of computational methods in the field. The final model was developed using all data and its performance was estimated using statistical parameters calculated following the 5CV protocol.

In addition to 5CV a prospective test set comprising 5443 molecules was used. The molecules in this set did not overlap with those in the training set, which was checked using International Chemical Identifier (InChi) hash keys when ignoring stereochemistry of molecules.

3.3 Regression models quantitative quality indicators

There are several statistical indices designed to assess a regression model's predictive ability. The most well-known is the coefficient of determination, designated in this study as Q^2. This coefficient measures a fraction of variance in the data covered by the model. The closer this value is to 1, the more accurate the model's predictions for the training set are, indicating the model fitting. Another important index is the root mean square error (RMSE) and the mean absolute error (MAE). The determination coefficient Q^2, RMSE and MAE are described below in equations 5 to 7, respectively.

$$Q^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{5}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2} \tag{6}$$

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i| \tag{7}$$

where N is the total number of data points (observations), y_i the measured value for the (i)-th data point, $\hat{y}_i$ the predicted value for the (i)-th data point, and $\bar{y}$ the mean of all the measured values in the dataset

3.4 Molecular descriptors

Descriptor-based models (section 3.4.1) rely on the conversion of molecules into numerical vectors, *i.e.* each molecule is represented by the same size set of numbers in a multidimensional space. These numbers, known as molecular descriptors, encode molecules' characteristics and serve as the coordinate axes in this space. The goal is to establish a mathematical relationship between a specific property of the molecule P and its structural descriptors D, expressed as $\mathrm{P} = f(\mathrm{D})$. All descriptor packages described below are implemented into OCHEM. Some descriptors required 3D structures, which were calculated by RDKit (v. 2023.09.5).

CDK Descriptors (CDK23) accounts for 256 descriptors computed by the CDK (version 2.8) software.[52] It comprises constitutional, topological, geometric, electronic, and hybrid molecular descriptors. All these descriptors are sensitive to the geometry of the molecules and require the generation of 3D conformations.

Dragon7 Descriptors (Dragon7)[53] includes a large set of molecular descriptors (5270) computed using the commercial Dragon (version 7) software[54]. Some descriptors required 3D conformations.

EPA descriptors are calculated based on the 2D structure of molecules and include a wide range of molecular features developed by the US Environmental Protection Agency as part of their Toxicity Estimation Software Tool (TEST).[55] These descriptors are designed to predict various toxicological endpoints and physicochemical properties of chemical compounds.

ISIDA Fragment Descriptors (ISIDA-Fr)[56] are based on the occurrence of specific substructures within a molecule, accounting for sequence of atoms and bonds in a linear or circular disposition. In this study, the size of fragments ranged from two to four.

JPlogP[57] is a set of descriptors used to develop a model to predict lipophilicity of compounds with the same name.

Mold2 Descriptors (MolD2)[58] is a comprehensive set of 777 molecular descriptors calculated using the Mold2 software. It includes a wide variety of topological indices and counts for different types of atoms, bonds, functional groups, and other structural features.

MordRed[59] comprises 1613 two-dimensional descriptors and 213 three-dimensional ones. It was developed aiming to address inefficiencies in other descriptors previously developed (PaDEL, CDK, ChemoPy).

OEstate (Optimised Electrotopological State) package is a set of molecular descriptors that were developed using extended electrotopological state (E-state) indices. These descriptors were originally introduced by Hall and Kier[60], which were extended to capture detailed information about amino, carbonyl, and hydroxy groups as described elsewhere.[30]

PaDEL2 is a new open source version of PADEL[61] that currently offers 1875 descriptors (1444 are up to two dimensions and 431 are three dimensional) and 12 types of fingerprints with a total of 16092 bits which was implemented within OCHEM to support the modern version of Java.

RDKit descriptors[62] package used in this work includes 472 one and two dimensional and 933 tree dimensional descriptors as well as several topological fingerprints.. Many of these descriptors are similar to those offered by the Dragon descriptor package

ECFP4 (RDKit)[63] Extended-Connectivity Fingerprints were also calculated with (a radius of two Morgan fingerprints, corresponding to diameter 4 in ECFP notation) considering their popularity and wide use in the drug discovery field. However, models calculated with this type of descriptors had the lowest accuracy and thus were not included into modelling.

Unsupervised filtering was implemented to eliminate redundancy by removing descriptors with less than two different values per training dataset and those with strong correlation (Pearson correlation > 0.95) and variance smaller than 0.01.

3.5 Computational methods for logP prediction

OCHEM offers several tens of different methods that can be used to correlate logP with the structure of molecules including both descriptor-based and representation learning methods.

3.5.1 Descriptor-based methods

Previous studies have indicated that E-state indices[60] have yielded highly predictive models for logP.[30,64] These descriptors available as part of the OEstate package were therefore selected to estimate the performance of different descriptor-based methods. Initially, we investigated eight methods, which included

both linear (MLRA, PLS) and non-linear (kNN, ASNN, DNN, XGboost, CatBoost) methods as described in the Results and Discussion sections.

k Nearest Neighbors (kNN) is a simple method used for both classification and regression tasks. Its algorithm determines the class or property value of a new data point based on the similarity between the input data and the training examples. It takes the k nearest data points in the feature space of the training data. The "nearness" is typically measured using distance metrics such as Euclidean distance or correlation coefficients like the Pearson correlation coefficient.

Multiple Linear Regression Analysis (MLRA) is a method that helps identify the most important variables in a dataset by gradually removing less significant ones. It does this by evaluating each variable's contribution to the overall regression model through the application of a significance test. This process simplifies the model, making it easier to understand and interpret the relationships between the variables and the outcome being studied.

Partial Least Squares Regression (PLS)[65] is particularly effective for handling datasets where the number of variables is large compared to the number of dataset samples. It performs a dimension reduction by creating latent variables that capture the most relevant information in the data, which can effectively handle highly correlated original variables. The optimal number of latent variables is determined through cross-validation, ensuring that the model improves its predictive ability without overfitting.

Random Forest Regression (RFR)[66] is an ensemble model that combines multiple decision trees. As a first step, the RFR algorithm creates numerous decision trees using a randomly selected subset of the available features and training samples. The randomised selection decreases the correlation among trees and ensures that they capture different aspects of the relationship between the input features and the target property value. It reduces overfitting and improves generalisation. When making predictions, each tree in the forest independently generates an output based on the input features. The final prediction is then obtained by aggregating the outputs of all the trees, typically by taking the average in the case of regression tasks. This ensemble approach helps to reduce the potential bias of individual trees.

Scikit gradient-boosted trees, such as XGBoost[67] and CatBoost[68], are ensemble learning methods that combine multiple decision trees, but unlike RFR, where the trees are built independently and in parallel, gradient-boosted trees construct the ensemble sequentially, with each new tree interactively attempting to correct the errors made by the previous trees. Finally, after converging, the final prediction is obtained by averaging the predictions of all the trees in the ensemble.

Associative Neural Network (ASNN)[69] is an ensemble method that combines multiple neural networks with kNN. The predictions of networks for each molecule create a new “ensemble” space. The kNN algorithm is then applied in this space to perform a bias correction of the global model for each specific data point based on the errors of the closest neighbours in the ensemble space.

Deep learning Neural Networks (DNN)[70] have several hidden layers in their architecture, which is particularly useful when dealing with complex relationships between molecular descriptors and the target property/activity being predicted. As the input data passes through the hidden layers, the network learns to combine these descriptors, forming a hierarchical internal representation of the chemical structures. The DNN architecture used in this work is composed of six hidden layers densely connected with ReLU activation function and Adam optimization method.[70]

3.5.2 Representation learning methods

In addition to the descriptor-based methods we also used representation learning methods based on the representation of molecules as graphs (KGCNN) as well as text (Transformer-CNN and its variation CNF2). Below is a brief overview of these methods.

Keras Graph Convolution Neural Networks (KGCNN)[35] is a package that provides layer classes for building graph convolution models. It supports multiple backends, including PyTorch, TensorFlow, and Jax. Aiming to efficiently process and analyse multiple graphs simultaneously, KGCNN uses the concept of disjoint and batched graphs. The first has a single graph that consists of smaller disjoint sub-graphs, while the second has multiple graphs stacked along a batch dimension to form a single data structure. After preliminary evaluation of the initial dataset, seven methods were selected (Table 1).

Table 1. Overview of the KGCNN-derived methods applied to the final model. For more detailed descriptions of KGCNN algorithms, see the GitHub webpage (github.com/aimat-lab/gcnn_keras)

KGCNN method	Abbr.	ref
Attentive Fingerprints	AttFP	[71]
Message Passing Neural Network	ChemProp	[72]
Graph ATtention networks	GAT	[73]
GAT version 2	GATv2	[74]
Graph Isomorphism Network	GIN	[75]
Pre-trained GIN	GINE	[76]
Hamiltonian Neural Networks	HamNet	[77]

Transformer Convolutional Neural Networks (TRANS-CNN)[34] is NLP (Natural Language Processing) method that was pre-trained over 1.7M molecules from the ChEMBL database to learn the task of canonisation of chemical structures. The learned latent representation that is used as input to one dimensional Convolutional Neural Network (CNN) and its output is correlated with target properties of molecules using fully connected neural networks.

Transformer Convolutional Neural Network Fingerprint (CNF)[78] extends the previous method and uses a combination of several CNNs with different receptive fields to provide a richer representation (fingerprint) to be correlated with target properties of molecules.

3.6 Consensus modelling

In addition to the development of individual models, consensus models were also built. A simple aggregating approach, which averages the contributions of all individual methods, was applied. The consensus had better performance than any individual approach due to the correction of variance and bias of each individual model.

3.7 Applicability domain and estimation of accuracy of predictions

The AD study was carried out based on the “distance to model” (DM) approach to estimate the accuracy of individual predictions for regression models.[38] A larger DM corresponds to a larger dissimilarity of the test compound to the model and thus the model is expected to have a lower accuracy for such chemicals. Among the several DMs supported by the OCHEM, the one based on the standard deviation (CONSENSUS-STD) of an ensemble of models used in consensus[38] was applied. It measured (dis)agreement among all the models, ensuring discrimination between low (all models disagree) and high-confidence predictions. The average error increased with DM and was used to identify outlying predictions, which had errors too large to be produced using Gaussian Distribution.[79] Such measurements could likely be experimental errors due to e.g, incorrect measurements or incorrect annotation of pH values of measurements. The DM value which covered 95% of data in the training set was used to determine the applicability domain (AD) of the model.

3.8 Tautomers subset canonicalisation

OCHEM contained molecules in different tautomeric forms, which could influence the accuracy of developed models. Thus it was interesting to see whether using the standardised tautomeric forms could improve modelling as claimed in another study by Ulrich et al,[16] where the authors used data augmentation considering all potential tautomeric forms of the chemicals to train a representation learning DNN model to predict logP. They stated that “the performance of all models increases if the randomly selected tautomer is first transformed into the major tautomer using JChem”.

Although the same tool (JChem by ChemAxon) was sought to carry out this analysis, it was not provided within the education package. Therefore, to canonicalize the training and test sets, the RDKit Python library was applied, specifically the 2020.03 release[17] and the `rdkit.Chem.MolStandardize` module with tools for normalising molecules defined by SMARTS patterns. The **`Enumerate`**`()` and **`Canonicalize`**`()` functions provided by this module were employed to find molecules with multiple tautomeric forms and the canonical tautomer, respectively. As mentioned by the authors, the algorithm used by the tool was based on Sitzmann et al.[47] While this algorithm may not provide the most stable tautomer for all molecules, it was designed to produce “reasonable” tautomers, that is, with aromatic rings scoring the most, and then other more stable tautomeric forms over the other (*i.e.* keto over iminol, amide over iminol, etc.). The RDKit library was chosen over other libraries that use quantum calculations to find the most stable tautomer,[48,49] since they were all designed for the gas phase and may yield different results in aqueous solution and varying pH levels.

4 RESULTS

4.1 Selection of methods

The performance of all descriptor-based methods described in section 3.4.1 were assessed individually with OEstate as the initial molecular descriptor. OEstate was selected since it contributed highly predictive models for logP in the previous studies.[30,64] The results of the training set 5CV for the best-performing models are shown in Table 2. For more details over all models, see Discussion section.

Table 2. Quantitative indicators of 5CV performance of the best descriptor-based machine learning methods using OEstate molecular descriptors

Method	Whole set (23167)[a]		LogP set (19336)[a]		
	RMSE	Q^2	RMSE	RMSE (zwitt. excl.)	Q^2
DNN	0.805 ± 0.005	0.826 ± 0.002	0.626 ± 0.006	0.610 ± 0.006	0.890 ± 0.002
CatBoost	0.796 ± 0.005	0.830 ± 0.002	0.625 ± 0.005	0.607 ± 0.006	0.891 ± 0.002

[a]The number of unique compounds used for each model are indicated in the parentheses.

CatBoost and DNN had similar performances for the whole Sangster set and its logP subset. Their statistical parameters were significantly better than other methods for both RMSE and coefficient of determination. Based on this analysis we decided to use these two methods for further study. While the removal of potentially ionised compounds (logD subset) significantly decreased (by about 0.2 log units on average) the RMSE for the logP set, the exclusion of zwitterion just affected it slightly (RMSE decreased for 0.01-0.02 log units). And we decided in turn to work only with the logP set and to keep zwitterions in the following studies. The selected methods were applied to other descriptors available on the OCHEM web site. The 5CV performances for the logP subset are shown in Table 3.

With an RMSE < 0.63 delimited as a threshold, seven descriptors combined with DNN and twelve with CatBoost (highlighted at Table 3) provided best results. A similar analysis was performed for representation learning methods. Following that analysis we selected seven graph-neural network approaches which provided the highest accuracy for both of these sets. These methods were used in combination with Transfomer-CNN and Transforner-CNF as representation learning approaches for the further analyses.

Table 3. Quantitative indicators of the 5CV performance for models trained on the logP dataset using CatBoost and DNN methods with different sets of molecular descriptors.

	DNN		CatBoost	
	RMSE	**Q^2**	**RMSE**	**Q^2**
2D Descriptors				
EPA	0.608 ± 0.007	0.896 ± 0.002	0.602 ± 0.006	0.898 ± 0.002
MORDRED (2D)	0.617 ± 0.005	0.893 ± 0.002	0.614 ± 0.005	0.894 ± 0.002
PaDEL (2D)	0.625 ± 0.005	0.891 ± 0.002	0.612 ± 0.005	0.895 ± 0.002
OEstate	0.626 ± 0.005	0.890 ± 0.003	0.625 ± 0.005	0.891 ± 0.002
MOLD2	0.630 ± 0.007	0.889 ± 0.002	0.625 ± 0.006	0.890 ± 0.002
ISIDA fragments	0.642 ± 0.006	0.885 ± 0.002	0.614 ± 0.005	0.885 ± 0.002
JPlogP	0.632 ± 0.006	0.888 ± 0.003	0.640 ± 0.005	0.885 ± 0.002
3D Descriptors				
CDK23 (3D)	0.624 ± 0.006	0.891 ± 0.002	0.606 ± 0.006	0.897 ± 0.002
RDKit (3D)	0.669 ± 0.006	0.875 ± 0.00	0.616 ± 0.006	0.893 ± 0.002
PaDEL2 (3D)	0.627 ± 0.006	0.89 ± 0.002	0.612 ± 0.005	0.895 ± 0.002
MORDRED (3D)	0.617 ± 0.006	0.893 ± 0.002	0.615 ± 0.005	0.894 ± 0.002
Dragon7 (3D)	0.649 ± 0.006	0.882 ± 0.002	0.616 ± 0.006	0.894 ± 0.002
alvaDesc(3D)	0.649 ± 0.007	0.882 ± 0.002	0.608 ± 0.005	0.896 ± 0.002

4.2 Consensus modelling

The statistical parameters of the consensus models (see Table 4) showed that inclusion of models based on 3D descriptors did not improve the performance of descriptor based models. Considering that the generation of 3D structures requires additional steps for 2D to 3D conversion, we decided to work only with 2D descriptors.

Table 4. Statistical parameters of consensus models built using different sets of models.

Models in the consensus	**RMSE**	**Q^2**
2D descriptor-based (10)[1]	0.570 ± 0.006	0.909 ± 0.002
3D descriptor-based (9)	0.580 ± 0.006	0.906 ± 0.002
All descriptor-based (19)	0.573 ± 0.006	0.909 ± 0.002
Representation learning (9)[2]	0.551 ± 0.006	0.915 ± 0.002
2D descriptor-based + Representation learning (19)	0.552 ± 0.005	0.914 ± 0.002

[1] Ten 2D descriptor-based models included DNN and CatBoost methods applied to 2D descriptors sets with the highest performances as highlighted in grey in Table 3. [2]The representation learning methods included seven KGCNN approaches from Table 1 as well as Transfomer-CNN and CNF methods.

The representation learning methods provided higher performance (lower RMSE and higher q^2) than descriptor-based methods, with the most performant (KCGNN ChemProp) with a Q^2 of 0.902 ± 0.002. However, the differences in statistical coefficients for both groups of these methods were non-significant. A combination of both these types of models in a common consensus did not improve its performance compared to the results based on the consensus of representation learning methods alone. Still, to have a higher diversity of models we used consensus based on 2D and representation learning methods, which is referred to in the following sections as the final consensus model.

4.3 Manual correction of data

The consensus models built for logP and logD subsets were used to identify potential data errors as indicated in section 2.2. The literature was revisited to confirm (or not) whether the outliers were errors. Most of the corrections corresponded to verification and correction (if required) pH values. If after correction we found that measurements were performed for neutral species, the molecules were moved from logD to logP sets and vice versa. This analysis constituted one of the most time-consuming parts of this study and led to the separation of molecules on the logP and logD sets as described in the Data section.

While the data collected by James Sangster were usually of high quality, we still found 380 data points with pH corrected, mostly wrongly attributed when it was already previously adjusted through equations described in the Introduction. Moreover, the chemical structure was rectified for 24 data points.

4.4 Exclusion of outliers

The corrections implemented during the data preprocessing predominantly addressed the errors arising from data entry rather than those from the experiment itself or due to different experimental protocols. Thus even after revisiting original articles some of errors could be still present in the curated dataset.

The consensus standard deviation (CONSENSUS-STD) was used to exclude outlying molecules. We used the same procedure described by Tetko et al.[79] where errors which were unlikely to be produced by mixture of gaussian distribution, which described evolution of errors as a function of DM,[38] were identified and excluded. In four iterations, 12048 outlying data points were excluded; therefore, 19000 data points (18945 unique compounds out of 19366) remained in the training set. The retrained model with excluded outlying data points decreased its RMSE by 32% while increasing Q^2 by 5%, suggesting that the retrained model could explain a greater proportion of the variance in the logP values (see Table 5). Moreover, the performance for the test set was also improved and the AD coverage for this set increased by >9%. Thus exclusion of outliers increased accuracy of the model.

Table 5. Statistical parameters of the consensus logP model before and after the exclusion of outlying data points

Set	5CV consensus		logP Test set		Test set within AD		
	RMSE	Q^2	RMSE	Q^2	RMSE	Q^2	fraction
All data, (31048)[a]	0.552 ± 0.005	0.914 ± 0.002	0.87 ± 0.01	0.716 ± 0.007	0.74 ± 0.01	0.762 ± 0.008	0.798
Outliers excluded, (19000)[a]	0.436 ± 0.005	0.943 ± 0.001	0.82 ± 0.01	0.729 ± 0.008	0.71 ± 0.01	0.775 ± 0.008	0.872

[a]The number of records used for each consensus model is indicated in the parentheses.

4.5 LogD set analysis

The original logD data predicted with the consensus logP model had a high RMSE of 1.77 ± 0.01 logP unit and a negligible Q^2 value of 0.0, indicating poor predictive performance (Figure 4A). After converting logD to logP (see section 2.2), the adjusted logP values were predicted by logP model (Figure 4B) with a significantly lower RMSE of 1.27 ± 0.02 logP unit and a significantly higher $Q^2 = 0.56 \pm 0.01$. These results demonstrate that the adjusted partition coefficient values were much more similar to the logP. The better fit of the data points over the trained model (in grey) is evident when visualising the predicted versus measured plot below.

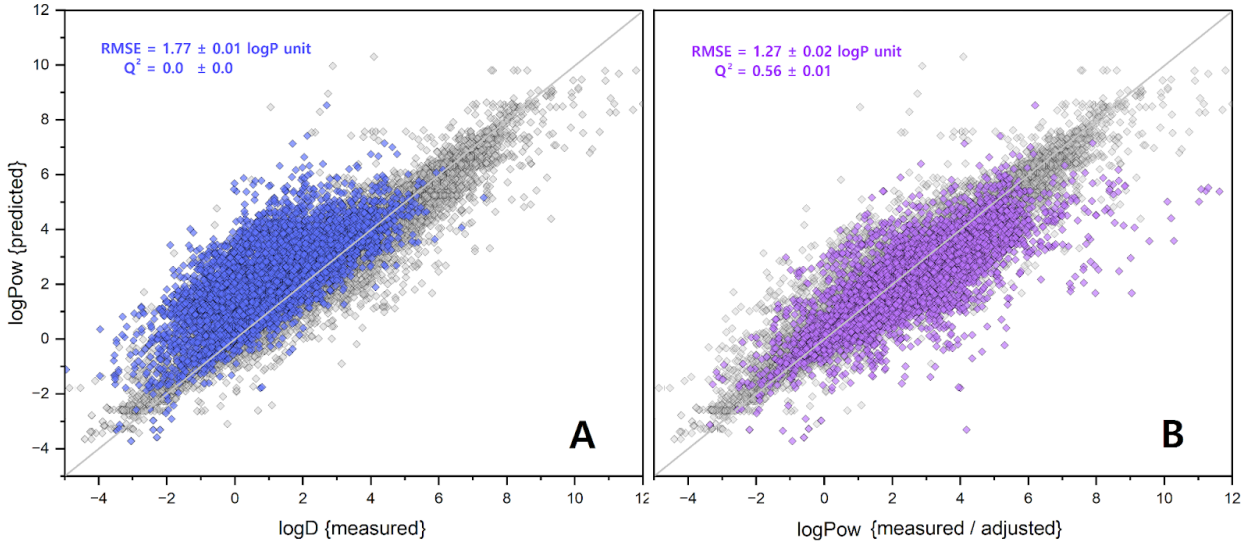

Figure 4. The predicted logP (y-axis) versus measured logP/logD (x-axis) scattered plots. Each rhombus represents a compound. The rhombi in grey (◇) are training set logP values predicted using the 5CV for both plots. The plot A depicts the original logD values as test set with rhombi in blue (◆), while the plot on the plot B with purple rhombus (◆) represents the adjusted logP values.

We also developed consensus models using logD and adjusted logP values which were then applied to predict logP values. As shown in Table 6, the model based on original logD data had much higher RMSE than the one developed using adjusted logP values. Analogously the logD model predicted molecules from the logP set with higher error (Figure 5A) than the one developed using adjusted logP values (Figure 5B).

Table 6. Models quality indicators for the consensus and the validation for the logP as test set

Model training set	Consensus		LogP set	
	RMSE	Q^2	RMSE	Q^2
Original LogD	0.93 ± 0.001	0.715 ± 0.006	1.295 ± 0.009	0.493 ± 0.006
Adjusted from LogD	0.88 ± 0.01	0.788 ± 0.007	0.913 ± 0.07	0.748 ± 0.004

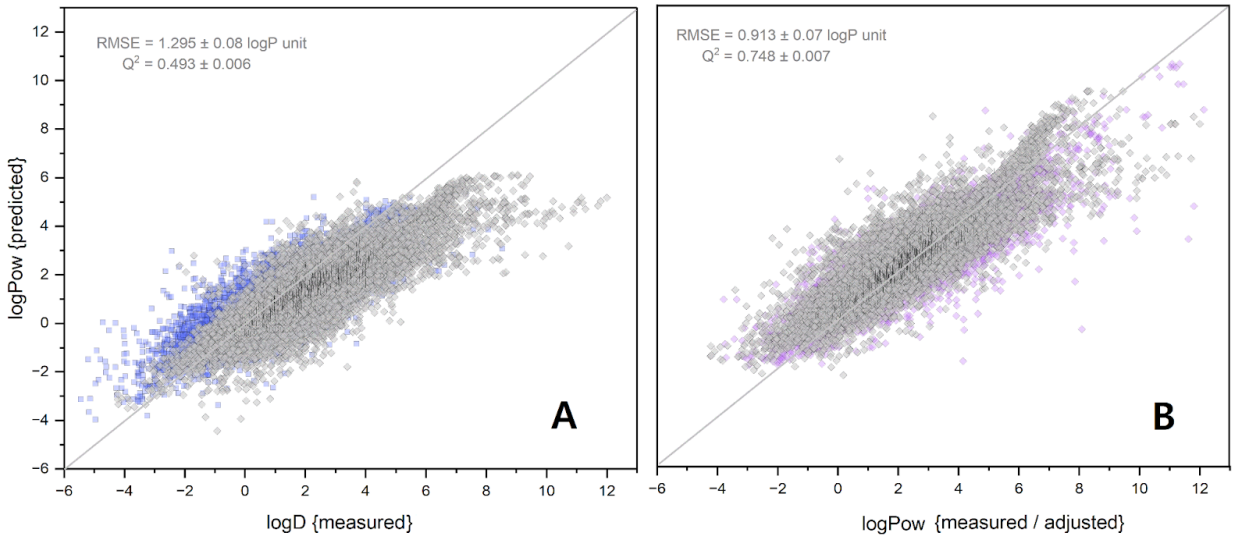

Figure 5. In this predicted versus measured scattered plot, the rhombi grey (◇) are the logP test set in both plots. The graphs differ regarding the training set: the plot A depicts the original logD values with rhombi in blue (◆), while the plot B with purple rhumbi (◆) representing the adjusted logP values.

4.6 Tautomers canonisation and subset analysis

The `Enumerate()` function was initially applied to 23167 molecules in the Sangster dataset input as SMILES strings, resulting in the identification of 5983 molecules (over a quarter) that exhibited two or more tautomeric forms. Then, the `Canonicalize()` function was applied to ensure that the same canonical tautomer for a molecule was always obtained, regardless of the input tautomer or atom ordering. This operation identified only 32 molecules (out of 5983) not in a canonical tautomeric form. The same process was applied over the test set, and 1966 out of 5443 molecules were identified as tautomers.

To understand how the canonical representation of tautomers affects the overall performance, two training sets were used: the original molecules before standardisation and the molecules after RDKit standardisation to the canonical tautomer representation. Correspondingly, two test sets were created - one with non-canonical tautomers and one with canonical tautomers. The non-canonical models were validated on the non-canonical test set, while the canonical models were validated on the canonical test set. The consensus models were built using the same methods and descriptors as described in section 4. Its quantitative indicators of quality for canonical and non-canonical are shown in Table 7, as well as the test set validation indicators.

Table 7. Consensus model quality indicators for the tautomer logP subset and using the correspondent test set validation.

Models set	Consensus		The test set	
	RMSE	R^2	RMSE	Q^2
Non-canonical	0.73 ± 0.01	0.817 ± 0.005	0.84 ± 0.02	0.55 ± 0.005
Canonical	0.72 ± 0.01	0.819 ± 0.005	0.85 ± 0.02	0.54 ± 0.02

These results suggest that the canonicalisation with RDKit did not provide a significant improvement in the overall performance, which could be due to the very small number of compounds that were not in canonical form.

5 DISCUSSION

5.1 LogP and logD separation sets

The accuracy of prediction models heavily relies on the quality and consistency of the training data. One of the possible sources of inaccuracy when developing models for $logP_{ow}$ arises from its definition: the intrinsic lipophilicity is the partition coefficient for n-octanol/water of the **neutral** species in both phases. Thus, when a molecule has ionisable groups, coexisting in both neutral and ionised forms depending on the pH of the medium, this distinction is crucial when building models for logP prediction, as the inclusion of logD data introduces significant errors as shown in Table 2. Our study addresses this issue by proposing a methodology to separate neutral from ionised species and to convert the categorised logD values to their corresponding logP values. This correction step proved to be highly effective, as evidenced by the significant improvement in the predictive performance of the adjusted logP model on the original logD one.

When the distribution of logP/logD values for each set (Figure 6) is examined, it can be observed that both sets follow a normal distribution. However, the mean, which also corresponds to the maximum distribution, is shifted to lower logD values for the logD set compared to the logP set. This shift is attributed to the presence of ionised species in the logD set, as they inherently have lower logD values compared to their neutral counterparts.

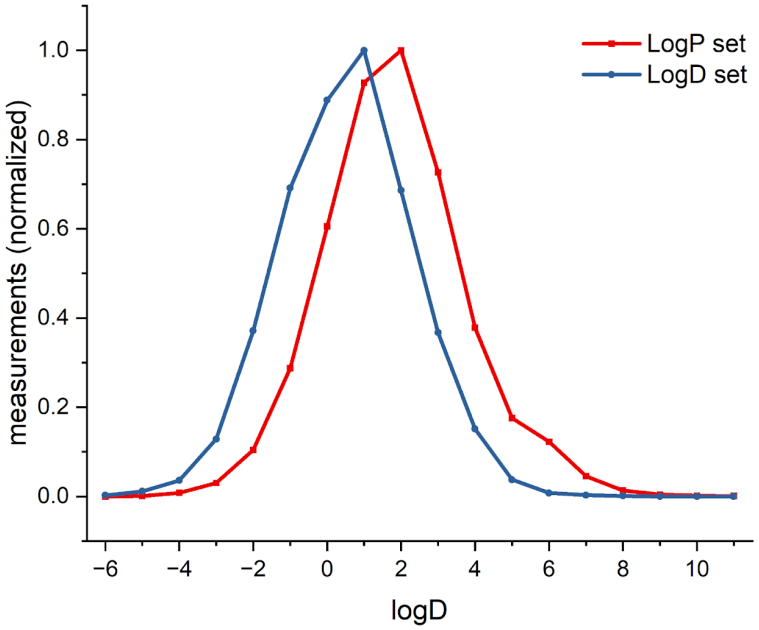

Figure 6. Distributions plotted on a graph with the x-axis representing the measured distribution coefficient and the y-axis representing the number of measurements normalized from 0 to 1. The red curve corresponds to the measurements in the "LogP set" while the blue curve corresponds to the "LogD set".

To better understand which functional groups are overrepresented in each set, the setCompare[1] OCHEM built-in tool was applied. Overrepresentation was measured as the ratio of percentages of each group in both logP and logD sets. For the logD set, the main functional groups overrepresented are

pnictogens, nitrogen and phosphorus, primary and secondary amines, carboxylic acids, and heterocyclic compounds. In contrast, the logP set had hydrocarbons, aromatic (without heteroatom), halogen derivatives, and ether overrepresented. This observation aligns with the expectation that ionisable groups are more prevalent in the logD set, as they are sensitive to pH changes, while non-ionizable groups are more likely to be found in the logP set.

5.2 Comparison of descriptor-based methods

The outset study done to select the best method (section 4.1) demonstrated a clear trade-off between model accuracy and interpretability, as shown in Figure 7. The most accurate were NN and ensemble-based methods, representing more advanced and complex machine learning techniques. Conversely, simpler methods like kNN, MLR, and PLS exhibited lower predictive performances.

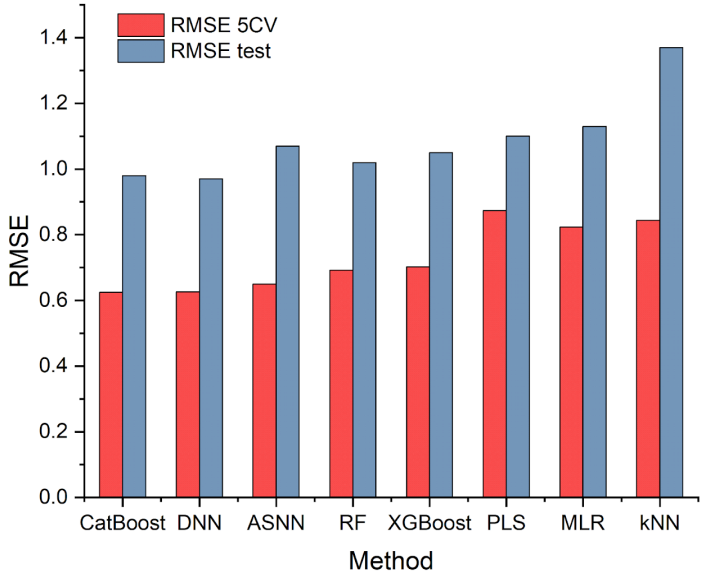

Figure 7. Bar chart comparing the performance of different descriptor-based methods when combined with the OEstate molecular descriptor. The chart displays the RMSE values for each method. The red bars represent the RMSE obtained through 5CV, while the blue bars represent the RMSE obtained from validation using the test set.

This observation aligns with the discussion by Wu et al.[85] in their study on the Tox21 data sets, where they have shown that more complex models like Support Vector Machines (SVM) and DNN generally attained higher accuracy at the expense of reduced interpretability and transparency, earning them the metaphorical label of "black boxes".[86] This lack of transparency in how NNs arrive at their predictions contrasts with the less accurate machine learning methods like decision trees, linear regression and k-nearest neighbours (k-NN), which are considered by many people as more explainable. Although these algorithms with a simple architecture may not be the most powerful, they are indeed more explicable when using interpretable descriptors, that is, their behaviour can be understood and accepted by humans. In practice, there is often a trade-off, necessitating a balanced approach. High accuracy is crucial for reliable predictions, but interpretability enables trust,[85] understanding of the model's reasoning, and potential insights

into the underlying chemistry. The optimal strategy is in line with Occam's razor, *i.e.* it may involve using simpler, interpretable models with acceptable accuracy and employing more complex approaches like DNN or ensemble techniques only when substantial accuracy gains can be achieved. Alternatively, methods to improve the interpretability of complex models could be explored.

5.3 Representation learning versus descriptor-based DNN methods

Representation learning methods have gained significant attention in recent years for their ability to learn directly from molecular structures. Their ability to learn the intrinsic patterns and relationships from raw molecular data, without the need for explicit descriptor engineering, presents a significant advantage. The superiority of representation learning methods observed in this study aligns with the growing trends in the field of cheminformatics and drug discovery and previous studies on the same topic.[34]

The most performant representational learning model was based on the Message Passing Neural Network (KGCNN ChemProp)[72] method, which operates on molecular graphs and has each node receiving messages from its neighbours to then update its representation based on the aggregated message. It combines convolution centered on bonds and descriptors. This architecture promotes it adaptabilidy in learning from a hybrid representation, merging task-specific encodings and fixed descriptors.

5.4 Exclusion of outliers

We showed that the exclusion of outliers improved the performance of models. Thus retaining outliers can result in modelling noise and reduce the model's overall performance. However, at the same time removing too many data points can lead to a loss of the model's applicability domain coverage. In this study, we aimed to achieve a compromise that addressed both of these considerations, and thus the removal of outliers was terminated when the joint accuracy of the model on training + removed set started to increase (in our case after 4 iterations). This provided a good compromise as evidenced by the improved performance of the model on the test set after excluding the outliers (see Table 5).

The setCompare[1] tool was applied between this outlying set and the set free of outliers. The analysis revealed that carboxylic acids, phenols, primary and secondary amines, amino acids, and halogen derivatives were overrepresented in the excluded outlying data. All of these functional groups, except for halogen derivatives, may correspond to measurements without annotated pH that were erroneously attributed to the logP set. The overrepresentation of halogen derivatives is expected since measurements of apolar halogen derivatives tend to be inaccurate.[87]

6 CONCLUSION AND PERSPECTIVES

The primary objective of this study was achieved by building a predictive model for molecular lipophilicity with a good prediction accuracy: the 5CV Q^2 = 0.943 ± 0.001 while the model predicted test set with Q^2 of 0.729 ± 0.008. A critical step towards this goal was the separation of logP and logD data. Moreover, we proposed to convert logD values to their corresponding logP values, which significantly improved the model's predictive performance, highlighting the importance of data consistency and adherence to the fundamental definition of logP as the partition coefficient for neutral species. The final consensus model was further enhanced by the exclusion of outliers, which ensured the mitigated potential sources of noise that could adversely impact model performance.

Notably, the consensus model combined descriptor-based and representation learning methods, showing that an ensemble model capable of leveraging the strengths of different approaches was an effective strategy for achieving good predictive metrics and a well-defined applicability domain, demonstrating its potential for practical applications in various fields. The representation learning methods were the most performant methods and the best individual performance was achieved by the KGCNN ChemProp[72] based model. Although this field needs to be matured,[88] the fast development of representation learning approaches and emergence of novel architectures, which can not only incorporate natively 3D conformation of molecules[89] but also provide explanation of models,[34,90] will further advance the field of structure-property prediction to a new level: accurate, explainable and trustful predictions to further accelerate drug discovery. It looks like these methods can do it better than the traditional machine learning approaches.

The modelling methodology described in this work was conducted aiming to achieve good accuracy, however, in order to enhance interpretability, a linear model using functional groups[91] as interpretable descriptors could be an approach to be explored in a future work.

REFERENCES

1. Sushko, I. *et al.* Online chemical modeling environment (OCHEM): web platform for data storage, model development and publishing of chemical information. *J Comput Aided Mol Des* **25**, 533–554 (2011).
2. Pliška, V., Testa, B. & Waterbeemd, H. van de. *Lipophilicity in Drug Action and Toxicology*. (VCH, Weinheim, 1996).
3. *Molecular Drug Properties: Measurement and Prediction*. (Wiley, 2007). doi:10.1002/9783527621286.
4. Lombardo, F., Faller, B., Shalaeva, M., Tetko, I. & Tilton, S. The Good, the Bad and the Ugly of Distribution Coefficients: Current Status, Views and Outlook. in *Methods and Principles in Medicinal Chemistry* (ed. Mannhold, R.) 407–437 (Wiley, 2007). doi:10.1002/9783527621286.ch16.
5. Tetko, I. V. & Poda, G. I. Application of ALOGPS 2.1 to predict log D distribution coefficient for Pfizer proprietary compounds. *J Med Chem* **47**, 5601–5604 (2004).
6. Wypych, G. *Handbook of Solvents*. (ChemTec Publishing, Toronto, 2014).
7. Koelmans, A. A. Limited reversibility of bioconcentration of hydrophobic organic chemicals in phytoplankton. *Environ Sci Technol* **48**, 7341–7348 (2014).
8. Vorberg, S. & Tetko, I. V. Modeling the Biodegradability of Chemical Compounds Using the Online CHEmical Modeling Environment (OCHEM). *Mol Inform* **33**, 73–85 (2014).
9. Min, K., Cuiffi, J. D. & Mathers, R. T. Ranking environmental degradation trends of plastic marine debris based on physical properties and molecular structure. *Nat Commun* **11**, 727 (2020).
10. Hansch, C., Leo, A. & Hoekman, D. H. *Exploring QSAR*. (ACS, Washington, 1995).
11. Lipinski, C. A., Lombardo, F., Dominy, B. W. & Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Advanced Drug Delivery Reviews* **23**, 3–25 (1997).
12. Ghose, A. K., Viswanadhan, V. N. & Wendoloski, J. J. A Knowledge-Based Approach in Designing Combinatorial or Medicinal Chemistry Libraries for Drug Discovery. 1. A Qualitative and Quantitative Characterization of Known Drug Databases. *J. Comb. Chem.* **1**, 55–68 (1999).
13. Hansch, C. & Leo, A. *Substituent Constants for Correlation Analysis in Chemistry and Biology*. (Wiley, New York, 1979).
14. Overton, E. Ueber die osmotischen Eigenschaften der Zelle in ihrer Bedeutung für die Toxikologie und Pharmakologie: mit besonderer Berücksichtigung der Ammoniake und Alkaloide. *Zeitschrift für Physikalische Chemie* **22U**, 189–209 (1897).
15. Overton, C. E. *Studien Über Die Narkose: Zugleich Ein Beitrag Zur Allgemeinen Pharmakologie*. (G. Fischer, 1901).
16. Meyer, H. Zur Theorie der Alkoholnarkose: Erste Mittheilung. Welche Eigenschaft der Anästhetica bedingt ihre narkotische Wirkung? *Archiv f. experiment. Pathol. u. Pharmakol* **42**, 109–118 (1899).
17. Hansch, C., Maloney, P. P., Fujita, T. & Muir, R. M. Correlation of Biological Activity of Phenoxyacetic Acids with Hammett Substituent Constants and Partition Coefficients. *Nature* **194**, 178–180 (1962).
18. Kirkpatrick, P. & Ellis, C. Chemical space. *Nature* **432**, 823–823 (2004).
19. Fujita, Toshio., Iwasa, Junkichi. & Hansch, Corwin. A New Substituent Constant, π, Derived from Partition Coefficients. *J. Am. Chem. Soc.* **86**, 5175–5180 (1964).
20. Mannhold, R., Poda, G. I., Ostermann, C. & Tetko, I. V. Calculation of molecular lipophilicity: State-of-the-art and comparison of log P methods on more than 96,000 compounds. *J Pharm Sci* **98**, 861–893 (2009).
21. Buchwald, P. & Bodor, N. Octanol-water partition: searching for predictive models. *Curr Med Chem* **5**, 353–380 (1998).
22. Carrupt, P., Testa, B. & Gaillard, P. Computational Approaches to Lipophilicity: Methods and Applications. in *Reviews in Computational Chemistry* (eds. Lipkowitz, K. B. & Boyd, D. B.) vol. 11 241–315 (Wiley, 1997).
23. Klopman, G. & Zhu, H. Recent Methodologies for the Estimation of N-Octanol / Water Partition

Coefficients and their Use in the Prediction of Membrane Transport Properties of Drugs. *MRMC* **5**, 127–133 (2005).

24. Leo, A. J. Calculating log Poct from structures. *Chem. Rev.* **93**, 1281–1306 (1993).
25. Leo, A. J. & Hoekman, D. Calculating log P(oct) with no missing fragments;The problem of estimating new interaction parameters. *Perspectives in Drug Discovery and Design* **18**, 19–38 (2000).
26. Petrauskas, A. A. & Kolovanov, E. A. ACD/Log P method description. *Perspectives in Drug Discovery and Design* **19**, 99–116 (2000).
27. Wildman, S. A. & Crippen, G. M. Prediction of Physicochemical Parameters by Atomic Contributions. *J. Chem. Inf. Comput. Sci.* **39**, 868–873 (1999).
28. Raevsky, O. A., Trepalin, S. V., Trepalina, H. P., Gerasimenko, V. A. & Raevskaja, O. E. SLIPPER-2001 − Software for Predicting Molecular Properties on the Basis of Physicochemical Descriptors and Structural Similarity. *J. Chem. Inf. Comput. Sci.* **42**, 540–549 (2002).
29. Hornig, M. & Klamt, A. COSMO *f rag* : A Novel Tool for High-Throughput ADME Property Prediction and Similarity Screening Based on Quantum Chemistry. *J. Chem. Inf. Model.* **45**, 1169–1177 (2005).
30. Huuskonen, J. J., Livingstone, D. J. & Tetko, I. V. Neural Network Modeling for Estimation of Partition Coefficient Based on Atom-Type Electrotopological State Indices. *J. Chem. Inf. Comput. Sci.* **40**, 947–955 (2000).
31. Tetko, I. V., Tanchuk, V. Yu., Kasheva, T. N. & Villa, A. E. P. Estimation of Aqueous Solubility of Chemical Compounds Using E-State Indices. *J. Chem. Inf. Comput. Sci.* **41**, 1488–1493 (2001).
32. Prasad, S. & Brooks, B. R. A deep learning approach for the blind logP prediction in SAMPL6 challenge. *J Comput Aided Mol Des* **34**, 535–542 (2020).
33. Wu, K., Zhao, Z., Wang, R. & Wei, G. Top *P* – *S* : Persistent homology-based multi-task deep neural networks for simultaneous predictions of partition coefficient and aqueous solubility. *J Comput Chem* **39**, 1444–1454 (2018).
34. Karpov, P., Godin, G. & Tetko, I. V. Transformer-CNN: Swiss knife for QSAR modeling and interpretation. *J Cheminform* **12**, 17 (2020).
35. Reiser, P., Eberhard, A. & Friederich, P. Graph neural networks in TensorFlow-Keras with RaggedTensor representation (kgcnn). *Software Impacts* **9**, 100095 (2021).
36. Wu, Z. *et al.* MoleculeNet: a benchmark for molecular machine learning. *Chem. Sci.* **9**, 513–530 (2018).
37. Zhu, H. *et al.* Combinatorial QSAR Modeling of Chemical Toxicants Tested against Tetrahymena pyriformis. *J. Chem. Inf. Model.* **48**, 766–784 (2008).
38. Tetko, I. V. *et al.* Critical assessment of QSAR models of environmental toxicity against Tetrahymena pyriformis: focusing on applicability domain and overfitting by variable selection. *J Chem Inf Model* **48**, 1733–1746 (2008).
39. Gombar, V. K. & Enslein, K. Assessment of n-octanol/water partition coefficient: when is the assessment reliable? *J Chem Inf Comput Sci* **36**, 1127–1134 (1996).
40. Eros, D. *et al.* Reliability of logP predictions based on calculated molecular descriptors: a critical review. *Curr Med Chem* **9**, 1819–1829 (2002).
41. Sushko, I. *et al.* Applicability Domains for Classification Problems: Benchmarking of Distance to Models for Ames Mutagenicity Set. *J. Chem. Inf. Model.* **50**, 2094–2111 (2010).
42. Syracuse Research Corporation (SRC). The Physical Properties Database (PHYSPROP).
43. Hersey, A. *ChEMBL Deposited Data Set - AZ_dataset*. https://www.ebi.ac.uk/chembl/doc/inspect/CHEMBL3301361 (2015) doi:10.6019/CHEMBL3301361.
44. Zdrazil, B. *et al.* The ChEMBL Database in 2023: a drug discovery platform spanning multiple bioactivity data types and time periods. *Nucleic Acids Research* **52**, D1180–D1192 (2024).
45. Andrés, A. *et al.* Setup and validation of shake-flask procedures for the determination of partition coefficients (logD) from low drug amounts. *European Journal of Pharmaceutical Sciences* **76**, 181–191 (2015).
46. OECD. *Test No. 123: Partition Coefficient (1-Octanol/Water): Slow-Stirring Method*. (OECD, 2022). doi:10.1787/9789264015845-en.

47. OECD. *Test No. 107: Partition Coefficient (n-Octanol/Water): Shake Flask Method*. (OECD, 1995). doi:10.1787/9789264069626-en.

48. Win, Z.-M., Cheong, A. M. Y. & Hopkins, W. S. Using Machine Learning To Predict Partition Coefficient (Log *P*) and Distribution Coefficient (Log *D*) with Molecular Descriptors and Liquid Chromatography Retention Time. *J. Chem. Inf. Model.* **63**, 1906–1913 (2023).

49. Wang, Y. *et al.* LogD7.4 prediction enhanced by transferring knowledge from chromatographic retention time, microscopic pKa and logP. *J Cheminform* **15**, 76 (2023).

50. Sangster, J. M. Octanol-water partition coefficients of simple organic compounds. *ACS Journal of Physical Chemistry* **18**, (1989).

51. Sushko, I., Salmina, E., Potemkin, V. A., Poda, G. & Tetko, I. V. ToxAlerts: A Web Server of Structural Alerts for Toxic Chemicals and Compounds with Potential Adverse Reactions. *J. Chem. Inf. Model.* **52**, 2310–2316 (2012).

52. Steinbeck, C. *et al.* The Chemistry Development Kit (CDK): An Open-Source Java Library for Chemo- and Bioinformatics. *J. Chem. Inf. Comput. Sci.* **43**, 493–500 (2003).

53. Todeschini, R. & Consonni, V. *Handbook of Molecular Descriptors*. (John Wiley & Sons, 2008).

54. Mauri, A., Consonni, V., Pavan, M. & Todeschini, R. Dragon software: An easy approach to molecular descriptor calculations. *Match* **56**, 237–248 (2006).

55. US Environmental Protection Agency. EPA C.C.T.E. Toxicity estimation software tool (TEST). (2022).

56. Varnek, A. *et al.* ISIDA - Platform for Virtual Screening Based on Fragment and Pharmacophoric Descriptors. *CAD* **4**, 191–198 (2008).

57. Plante, J. & Werner, S. JPlogP: an improved logP predictor trained using predicted data. *J Cheminform* **10**, 61 (2018).

58. Hong, H. *et al.* Mold ^{2} , Molecular Descriptors from 2D Structures for Chemoinformatics and Toxicoinformatics. *J. Chem. Inf. Model.* **48**, 1337–1344 (2008).

59. Moriwaki, H., Tian, Y.-S., Kawashita, N. & Takagi, T. Mordred: a molecular descriptor calculator. *J Cheminform* **10**, 4 (2018).

60. Hall, L. H. & Kier, L. B. Electrotopological State Indices for Atom Types: A Novel Combination of Electronic, Topological, and Valence State Information. *J. Chem. Inf. Comput. Sci.* **35**, 1039–1045 (1995).

61. Yap, C. W. PaDEL-descriptor: An open source software to calculate molecular descriptors and fingerprints. *J Comput Chem* **32**, 1466–1474 (2011).

62. *The RDKit Book.* (2024).

63. Rogers, D. & Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **50**, 742–754 (2010).

64. Tetko, I. V. & Tanchuk, V. Y. Application of associative neural networks for prediction of lipophilicity in ALOGPS 2.1 program. *J Chem Inf Comput Sci* **42**, 1136–1145 (2002).

65. Wold, S., Sjöström, M. & Eriksson, L. PLS-regression: a basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems* **58**, 109–130 (2001).

66. Breiman, L. Random Forests. *Machine Learning* **45**, 5–32 (2001).

67. Chen, T. & Guestrin, C. XGBoost: A Scalable Tree Boosting System. (2016) doi:10.48550/ARXIV.1603.02754.

68. Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V. & Gulin, A. CatBoost: unbiased boosting with categorical features. (2017) doi:10.48550/ARXIV.1706.09516.

69. Tetko, I. V. Associative neural network. *Methods Mol Biol* **458**, 185–202 (2008).

70. Sosnin, S., Karlov, D., Tetko, I. V. & Fedorov, M. V. Comparative Study of Multitask Toxicity Modeling on a Broad Chemical Space. *J. Chem. Inf. Model.* **59**, 1062–1072 (2019).

71. Xiong, Z. *et al.* Pushing the Boundaries of Molecular Representation for Drug Discovery with the Graph Attention Mechanism. *J. Med. Chem.* **63**, 8749–8760 (2020).

72. Yang, K. *et al.* Analyzing Learned Molecular Representations for Property Prediction. *J. Chem. Inf. Model.* **59**, 3370–3388 (2019).

73. Veličković, P. *et al.* Graph Attention Networks. (2017) doi:10.48550/ARXIV.1710.10903.

74. Brody, S., Alon, U. & Yahav, E. How Attentive are Graph Attention Networks? (2021) doi:10.48550/ARXIV.2105.14491.

75. Xu, K., Hu, W., Leskovec, J. & Jegelka, S. How Powerful are Graph Neural Networks? (2018) doi:10.48550/ARXIV.1810.00826.

76. Hu, W. *et al.* Strategies for Pre-training Graph Neural Networks. (2019) doi:10.48550/ARXIV.1905.12265.

77. Li, Z., Yang, S., Song, G. & Cai, L. HamNet: Conformation-Guided Molecular Representation with Hamiltonian Neural Networks. (2021) doi:10.48550/ARXIV.2105.03688.

78. Makarov, D. M., Fadeeva, Yu. A., Shmukler, L. E. & Tetko, I. V. Beware of proper validation of models for ionic Liquids! *Journal of Molecular Liquids* **344**, 117722 (2021).

79. Tetko, I. V., M. Lowe, D. & Williams, A. J. The development of models to predict melting and pyrolysis point data associated with several hundred thousand compounds mined from PATENTS. *J Cheminform* **8**, 2 (2016).

80. Ulrich, N., Goss, K.-U. & Ebert, A. Exploring the octanol–water partition coefficient dataset using deep learning techniques and data augmentation. *Commun Chem* **4**, 90 (2021).

81. Landrum, G. *et al.* rdkit/rdkit: 2020_03_1 (Q1 2020) Release. [object Object] https://doi.org/10.5281/ZENODO.3732262 (2020).

82. Sitzmann, M., Ihlenfeldt, W.-D. & Nicklaus, M. C. Tautomerism in large databases. *J Comput Aided Mol Des* **24**, 521–551 (2010).

83. Wieder, M., Fass, J. & Chodera, J. D. Fitting quantum machine learning potentials to experimental free energy data: Predicting tautomer ratios in solution. Preprint at https://doi.org/10.1101/2020.10.24.353318 (2020).

84. Liu, Z., Zubatiuk, T., Roitberg, A. & Isayev, O. Auto3D: Automatic Generation of the Low-Energy 3D Structures with ANI Neural Network Potentials. *J. Chem. Inf. Model.* **62**, 5373–5382 (2022).

85. Wu, L. *et al.* Trade-off Predictivity and Explainability for Machine-Learning Powered Predictive Toxicology: An in-Depth Investigation with Tox21 Data Sets. *Chem. Res. Toxicol.* **34**, 541–549 (2021).

86. Sushko, Y. *et al.* Prediction-driven matched molecular pairs to interpret QSARs and aid the molecular optimization process. *J Cheminform* **6**, 48 (2014).

87. Brooke, D. N., Dobbs, A. J. & Williams, N. Octanol: Water partition coefficients (P): Measurement, estimation, and interpretation, particularly for chemicals with P > 105. *Ecotoxicology and Environmental Safety* **11**, 251–260 (1986).

88. Hartog, P. B. R., Krüger, F., Genheden, S. & Tetko, I. V. Using test-time augmentation to investigate explainable AI: inconsistencies between method, model and human intuition. *J Cheminform* **16**, 39 (2024).

89. Cremer, J., Medrano Sandonas, L., Tkatchenko, A., Clevert, D.-A. & De Fabritiis, G. Equivariant Graph Neural Networks for Toxicity Prediction. *Chem. Res. Toxicol.* acs.chemrestox.3c00032 (2023) doi:10.1021/acs.chemrestox.3c00032.

90. Gallegos, M., Vassilev-Galindo, V., Poltavsky, I., Martín Pendás, Á. & Tkatchenko, A. Explainable chemical artificial intelligence from accurate machine learning of real-space chemical descriptors. *Nat Commun* **15**, 4345 (2024).

91. Haider, N. Functionality Pattern Matching as an Efficient Complementary Structure/Reaction Search Tool: an Open-Source Approach. *Molecules* **15**, 5079–5092 (2010).

APPENDIX

Appendix A: Internship information

A1. About the host organisation

Located in Neuherberg, a suburb of Munich, the *Helmholtz Munich* (HZM) is part of the Helmholtz Association, Germany's largest scientific organisation. The centre has several groups conducting research in artificial intelligence (AI) for health and drug discovery. One of these groups based at the HZM Institute of Structural Biology has been led by Dr. Igor Tetko since 2001. He has a solid academic background with an MSc degree from the Moscow Institute of Physics and Technology, a PhD in Bioorganic Chemistry from the Ukrainian Academy of Sciences, and a *habilitation à diriger des recherches* (HDR) in Chemoinformatics from the University of Strasbourg. Moreover, with over 200 publications, Dr. Tetko carries extensive experience and expertise in machine learning (ML) tools for computer-aided drug discovery and life sciences.

A2. Activities assigned

This project was carried through a sequence of tasks that are described in chronological sequence below:

- Data preprocessing and cleaning, including molecular standardisation and input error correction.
- Separation of logP from logD values according to the experimental pH and calculated pKa values.
- Evaluation and comparison of various ML techniques and molecular descriptors for lipophilicity prediction.
- Development of consensus models combining multiple methods to improve accuracy and AD.
- Implementation of AD assessment and outlier exclusion techniques.
- Adjustment of each logD experimental data point to its estimated intrinsic lipophilicity (logP).
- Exploration of tautomer canonicalisation and its impact on model performance.
- (In progress) Development of an interpretable linear model using functional group descriptors.

A3. Skills acquired

The theoretical knowledge acquired to carry out this project includes a molecular lipophilicity understanding, ML methods and molecular descriptors used in QSAR predictive models, consensus modelling approaches for combining multiple models, techniques for assessing AD and handling outliers, and validation of QSAR models. All the algorithms for building models and their validation are implemented into the OCHEM, thus no model development skills were needed. Moreover, the following practical skills were acquired (or improved):

- Familiarity with OCHEM tools for chemoinformatics and modelling
- Importing data from experimental chemical databases like ChEMBL and OCHEM
- Development of Python scripts for the manipulation and analysis of chemical data.
- Exploring the RDKit library for identifying tautomers in a dataset and for tautomeric canonicalisation.