

Prediction of logP for Pt(II) and Pt(IV) complexes: comparison of statistical and quantum-chemistry based approaches

Igor V. Tetko,^{1,*} Hristo P. Varbanov^{2,3}, Markus Galanski³, Mona Talmaciu,^{4,5} Jamie Platts,⁴ Mauro Ravera,⁶ Elisabetta Gabano⁶

1) Helmholtz Zentrum München - German Research Center for Environmental Health (GmbH), Institute of Structural Biology, Ingolstaedter Landstrasse 1, b. 60w, D-85764 Neuherberg, Germany and BigChem GmbH, Ingolstaedter Landstrasse 1, b. 60w, D-85764 Neuherberg, Germany

2) Institut des Sciences et Ingénierie Chimiques, Ecole Polytechnique Fédérale de Lausanne (EPFL), CH-1015, Lausanne, Switzerland

3) Institute of Inorganic Chemistry, University of Vienna, Währinger Strasse 42, A-1090 Vienna, Austria

4) School of Chemistry, Cardiff University, Park Place, Cardiff CF10 3AT, UK.

5) «Iuliu Hațieganu» University of Medicine and Pharmacy, Faculty of Pharmacy, Analytical Chemistry Department, Cluj-Napoca, Romania.

6) Dipartimento di Scienze e Innovazione Tecnologica, Università del Piemonte Orientale, Viale Teresa Michel 11, 15121 Alessandria, Italy.

Keywords: Pt complexes; Pt(II); Pt(IV); octanol/water partitioning; lipophilicity; QSPR; Quantitative Structure Property Prediction; neural networks; linear regression; quantum chemistry calculations

*Corresponding author: itetko@vcclab.org, Tel.: +49-89-3187-3575, Fax: +49-89-3187-3585

Abstract

The octanol/water partition coefficient, logP, is one of the most important physico-chemical parameters for the development of new metal-based anticancer drugs with improved pharmacokinetic properties. This study addresses an issue with the absence of publicly available models to predict logP of Pt(IV) complexes. Following data collection and subsequent development of models based on 187 complexes from literature, we validate new and previously published models on a new set of 11 Pt(II) and 35 Pt(IV) complexes, which were kept blind during the models development step. The error of the consensus model, 0.65 for Pt(IV) and 0.37 for Pt(II) complexes, indicates its good accuracy of predictions. The lower accuracy for Pt(IV) complexes was attributed to experimental difficulties with logP measurements for some poorly-soluble compounds. This model was developed using general-purpose descriptors such as extended functional groups, molecular fragments and E-state indices. Surprisingly, models based on quantum-chemistry calculations provided lower prediction accuracy. We also found that all the developed models strongly overestimate logP values for the three complexes measured in presence of DMSO. Considering that DMSO is frequently used as a solvent to store chemicals, its effect should not be overlooked when logP measurements by means of the shake flask method are performed. The final models are freely available at <http://ochem.eu/article/76903>.

Introduction

Platinum complexes form an important class of chemotherapeutics widely used in cancer treatment, acting mainly through binding to DNA in cell nuclei. Their ability to cross cell membranes by passive diffusion (amongst other pathways), and hence their cellular uptake is therefore an important aspect of drug design [1, 2]. The increase of cellular accumulation of platinum with increasing the lipophilicity was demonstrated for several classes of Pt(II) and Pt(IV) complexes [3, 4]. Moreover, correlation between their lipophilicity and in vitro cytotoxicity was observed for various series of analogues [3, 5, 6].

Lipophilicity is usually expressed by the n-octanol/water partition coefficient, P , that reflects the relative solubility of the drug in n-octanol (a model of the lipid bilayer of a cell membrane) and water (the medium inside and outside the cells) [7]. $\log P$ is one of the properties identified by Lipinski in the “Rule of 5” for drug-like molecules,[8] and is therefore, one of the most important physicochemical parameters in drug discovery studies, being related to the bioavailability of chemical compounds [9, 10]. Furthermore, lipophilicity is a crucial parameter for the development of orally administered drugs, while $\log P$ in the range between 0.5 and 3.5 can be considered as optimal for good oral bioavailability [11].

Nowadays, most anticancer drugs are administered intravenously, a route that leads to immediate and complete bioavailability, but can also be hazardous and of difficult management. This is particularly true for Pt(II)-based chemotherapeutics such as the well-known anticancer drug cisplatin [1]. In general, patients prefer oral medications over intravenous therapy, and, hence, the development of oral drugs will be very beneficial for them. In this context, Pt(IV) complexes are particularly interesting, being promising anticancer prodrug candidates for oral administration, due to their kinetic inertness (allowing stability in the gastro-intestinal tract) and the various possibilities for tuning of their lipophilicity [3, 5, 12, 13].

The reliable prediction of logP of Pt complexes deduced from their chemical structure is an important asset in the effort to create new metal-based anticancer drugs with improved pharmacokinetic properties. The majority of methods developed so far in that context are based on quantum chemical calculations. The development of statistical machine-learning methods, which are frequently used for prediction of logP for organic compounds, has been limited due to a small number of Pt complexes with measured logP values. Moreover difficulties with representation of this type of compounds, especially metal-ligand bonds, and calculation of descriptors also contribute to this problem. Indeed, only recently the chemoinformatics software providers, such as ChemAxon, have included support of this type of bond in their software. Still, this software does not support stereochemistry of coordination bonds.

There is a limited number of publications in the literature about models predicting logP values of Pt(II) and Pt(IV) complexes by using different descriptors and different mathematical approaches [4, 14-20]. However, these methods require the use of descriptors that are not easy to employ. The users should be proficient in quantum chemistry, and should be able to prepare chemical structures in the format required to run the algorithms. Moreover, the success of these predictions depends on the correct selection and optimization of 3D structure of molecules and they require considerable calculation time.

In our previous study [16] we developed a program to predict logP values of Pt(II)-complexes, which was based on the local correction of predictions provided by the ALOGPS 2.1 program [21]. This program is freely available at <http://www.vcclab.org/web/pt> site. On that website, the coordination bond in metal complexes has to be replaced by a single bond, which is recognized by the ALOGPS program. Strictly speaking, this is an incorrect representation of Pt complexes structure and requires a special attention from the users who would like to use this service.

Moreover, to the best of our knowledge, there is no on-line program predicting logP values of Pt(IV) complexes.

The main goal of this study is to extend our previous work by providing better and freely accessible tools to predict logP of both Pt(II) and Pt(IV) complexes and benchmark the developed algorithms against the state of the art of quantum-chemistry approaches. Another aim was to investigate whether performance of local, *i.e.* based only on logP data for Pt complexes, or global models, *i.e.* when extending set of Pt complexes with other logP data (*i.e.* of organic molecules), will provide the highest prediction accuracy.

The ability to predict relevant physicochemical or biological properties of Pt complexes would be an important step towards the rational design of new molecules, allowing potential drug candidates to be assessed before lengthy synthesis and testing [22]. “Virtual screening” is now a standard aspect of the search for new organic drugs, but has perhaps not yet caught on to the same extent for metal-based drugs, not least because of the lack of suitable methods.

Molecules and Methods

Training Set Data collection

Data used in this study were collected from 34 literature sources. The data were uploaded to the OCHEM web site [23] and duplicated measurements were eliminated by keeping the value closest to the average logP value for the considered molecule. In the case logP values were determined using a HPLC method, only values obtained from calibration curves logP vs. log k_w (extrapolated to 0% methanol in the mobile phase) were used. There were 14 stereoisomers (enantiomers and/or their mixtures) in the dataset. The logP values for many of these molecules were very similar, e.g. -

1.59 and -1.54 for (*SP-4-2*)-(cyclohexane-1*S*,2*S*-diamine)(ethanedioato)platinum(II) and (*SP-4-2*)-(cyclohexane-1*R*,2*R*-diamine)(ethanedioato)platinum(II) (oxaliplatin),[4] respectively. Considering that most of the used descriptors were 2D-based, we selected only one structure, which was provided by the more reliable source, as determined by the described experimental protocol. Where both values were provided within the same study, the lower logP value was chosen for consistency. In total, 100 and 87 structures were selected for Pt(II) and Pt(IV) complexes, respectively.

Training Datasets

Four datasets were used to develop the models: individual sets with Pt(II) and Pt(IV) complexes, a joint set with both types of complexes and a joint set merged with 12898 data points previously used to develop ALOGPS 2.1 program [21]. The ALOGPS 2.1 set did not include any metal complexes. The distribution of logP values for Pt compounds is shifted towards smaller logP as compared to that of organic compounds (**Figure 1**).

Benchmarking dataset

In total, 46 platinum complexes for which logP values were unknown during the model development step were used to benchmark the developed methods. The molecules as well as their experimental logP values are reported in Table 1. This dataset includes 35 Pt(IV) complexes measured in two labs. For Pt(II) complexes, we had only three values (**1-3**) measured using the shake-flask method. Therefore, we decided to include additionally eight new values, i.e. in total 11 Pt(II) complexes (**1-11**), for which logP values were determined using the RP-HPLC method (**4-11**).

Complexes **1-32** were obtained and analyzed at the Dipartimento di Scienze e Innovazione Tecnologica, Università del Piemonte Orientale (Alessandria, Italy), whereas compounds **33-46** were obtained at the Institute of Inorganic Chemistry, University of Vienna (Austria). Compounds **33-40** and **43-46**[24-27], as well as **12-32**[18, 28, 29] have been previously reported, but their

shake-flask log P values have been measured for this work. Complexes **1-3** and **41-42** are new and their synthesis and complete characterization are provided in the supporting information.

Complexes **4-5**[30] and **6-11**[31] are already known but their logP values have been determined for this work using a known RP-HPLC method [15, 32].

Experimental measurement protocols

Two different methods (shake-flask and RP-HPLC), applied according to different experimental protocols developed and accomplished in two different labs (in Austria and in Italy) were used for logP determination of the compounds from the test set in order to resemble the bias of data collected (from different sources) in the training sets.

LogP values of complexes **1-3** and **12-32** were obtained by using the shake-flask method by preparing a solution of each compound in n-octanol-saturated water ($[Pt_{stock}] \approx 0.1-0.2$ mM) and shaking it with water-saturated n-octanol for 30 min. The mixture was then centrifuged for 30 min at 4000 rpm. In general, the volumes of n-octanol (V_{oct}) and water (V_{water}) in such measurements are equal, but for compounds with rather extreme logP values different volumes are often used. In the actual case, for complexes **12-29** V_{water} was 0.6 mL and V_{oct} was 50 mL. In the case of rather insoluble complexes **30-32**, 5% DMSO was added to the n-octanol-saturated water to get reproducible measurements ($V_{water} = V_{oct} = 5$ mL) [33]. For Pt(II) complexes **1-3**, more reproducible results were obtained using $V_{water} = 3$ mL, $V_{oct} = 20$ mL and $[Pt_{stock}] = 1$ mM. Pt concentrations were measured by means of a Spectro Genesis ICP-OES spectrometer (Spectro Analytical Instruments, Kleve, Germany) equipped with a crossflow nebulizer. For the measurements the Pt 299.797 nm line was selected and a Pt standard stock solution of 1000 mg L⁻¹ was diluted in 1.0% v/v nitric acid to prepare calibration standards. In the case of better water-soluble complexes **12-29**, it was also possible to confirm the finding that an accurate RP-HPLC or ¹H NMR measurement in

the presence of a suitable internal standard can substitute the ICP experiment to determine the concentration of the aqueous solutions [34, 35]. This can be particularly useful in the case of large series of complexes.

LogP values (Table 1) were calculated according to the formula, modified by taking into account the different employed volumes:[34]

$$\log P = \log \left(\frac{[Pt_{stock}] - [Pt_{aq}]}{[Pt_{aq}]} \times \frac{V_{water}}{V_{oct}} \right) \quad (1)$$

On the contrary, logP values of Pt(II) complexes **4-11** were determined with a RP-HPLC method, as follows. Aqueous solutions of the Pt(II) complexes (0.25 mM) were injected onto a C18 reverse phase HPLC column (250x4 mm LiChrospher 100 RP18 5 μ m column (Merck, Germany). The mobile phase was a 30:70 mixture of methanol and 15 mM HCOOH (flow rate = 0.75 mL min⁻¹, isocratic elution, UV-visible detector set at 210 nm). KI was the internal reference to determine the column dead-time (t_0). The retention time t_R of each complex was expressed in terms of log k' ($k' = (t_R - t_0) / t_0$), that in turn was used to calculate the corresponding log P based on the previous established relationship log P vs log k' [15, 32].

The n-octanol/water partition coefficients (logP) of compounds **33-46** were also determined using the classical shake-flask method, according to the OECD guidelines[36] with slight modifications, as described elsewhere [37]. Briefly, stock solutions of the complexes (~0.5 mM), freshly prepared in Milli-Q water (pre-saturated with n-octanol), were mixed with equal volumes of n-octanol (pre-saturated with Milli-Q water) and shaken for 1 h at RT. Platinum concentrations in the aqueous phase after phase separation by centrifugation ($[Pt_{aq}]$) and in the initial stock solutions ($[Pt_{stock}]$) were determined by means of ICP-MS. ICP-MS measurements were performed on an ICP-MS instrument Agilent 7500ce (Agilent Technologies, Waldbronn, Germany) equipped with a CETAC

ASX-520 autosampler and a MicroMist nebulizer at a sample uptake rate of approx. 0.25 mL min⁻¹. The instrument was tuned on a daily base in order to achieve maximum sensitivity and rhenium served as internal standard for platinum. Partition coefficients were calculated, according to the formula:

$$\log P = \log \left(\frac{[Pt_{stock}] - [Pt_{aq}]}{[Pt_{aq}]} \right) \quad (2)$$

All logP determinations were done in triplicate. Their mean values ± standard deviations are given in Table 1.

Pt(IV) complexes, as well as Pt(II) complexes featuring chelating carboxylates, are usually stable in solution, while dihalogenido (e.g. dichlorido) Pt(II) complexes are subjected to solvolysis (aquation or interaction with DMSO, when used for as solubilizing agent). In the experimental protocols used in both labs, solutions of Pt complexes were prepared freshly before the logP determination experiments. LogP values of dichloridoplatinum(II) complexes were measured by means of RP-HPLC, while freshly prepared solutions were promptly injected and analyzed within few minutes; no species that would result from solvolysis were detected during the experiment.

Representation of coordination (metal-ligand) bonds

The data preparation step included conversion of the structure to the same representation. Coordination bonds were represented using coordinate bond type (which was converted by ChemAxon to bond type #8 when exported and stored in sdf files) available in ChemAxon. This type is a non-standard one and it provides challenges with descriptor calculation, as the majority of descriptor calculation programs do not recognize it. To enable analysis of structures with different programs, an automatic conversion of coordinate bonds to single bonds (type #1) was implemented.

While this conversion introduced some error in chemical structure, it was consistent across all structures. Thus it provided the same bias, which was later accounted by data mining algorithm.

Descriptor calculation

Twelve different descriptor sets were used for the model development. The set of descriptors are listed in Table 2 together with references to the detailed description of algorithms. The descriptors are available online at OCHEM [23]. Dragon and CDK failed for calculation of 3D descriptors. Therefore only 2D based descriptors were calculated for them.

Unsupervised filtering of descriptors

Before model development, descriptors that had two or less non-zero values were eliminated. Moreover, we grouped together inter-correlated descriptors, *i.e.* those with linear correlation coefficient $R^2 > 0.95$, and selected for model development only one from each group.

Model development

Associative Neural Network [38-40] (ASNN) and Multiple Linear Regression Analysis (MLRA) were used to develop the methods analyzed in this study. Full details of model development protocol are described elsewhere [41, 42]. Briefly, each model consisted of 64 neural network models developed using the same architecture but with different random initializations of neural network weights. The performance of the ensemble models was evaluated using five-fold cross-validation (CV)[43] and bagging with 64 models [44]. In the former approach five models were built using 4/5 of initial training set and then used to predict the remaining 20% of data. Predicted values for the validation sets were used to estimate performance of models (CV results). In the bagging approach, 64 models were built. The training sets of models had the same size as the initial training set. Each set was created by random sampling with replacement of the initial training set. The detailed methodology used to develop the models is described elsewhere [42]. Neural networks

with 5 hidden neurons were used; this number was selected after performance analysis of models developed for all twelve descriptor sets. This architecture contributed the largest number of models with lowest Root Mean Squared Error (RMSE). The neural networks with other number of hidden neurons, e.g. 3 or 7, produced models with similar performance, *i.e.* the calculated result did not significantly depend in the used neural network architecture.

MLRA models were developed to provide mechanistic interpretation of calculated results. The step-wise regression was performed by eliminating on each step the least significant variable as identified by *t*-test. The variables significant at $p = 0.01$ were used in the final equations. All developed models estimate their applicability domains and accuracy of predictions using Leverage (MLRA), standard deviation of ensemble predictions (ASNN) and deviation of models in the consensus model.

Models based on quantum chemistry descriptors

SMILES strings for the benchmark data sets were converted to 3D structures through the CORINA web interface [45]. Previous experience indicates that descriptors such as polar surface area (PSA) are only weakly sensitive to conformational change,[46] so the CORINA-generated conformation was used without further searching. Where necessary, generated structures were manually corrected for issues such as hydrogen count or *cis/trans*-isomerism using GaussView4. These structures were then geometry optimized using PM6 [47] within MOPAC. Two previously published models were then applied to predict logP for the test set: the model denoted QC1 is based on PM6 electronic properties and it was trained on 28 Pt(IV) complexes [17].

$$\log P = 2.573 - 3.001 * q(\text{Pt}) - 24.41 * \max q(\text{H}) + 0.0204 * \text{Area_Cosmo} \quad (3)$$

The terms in the equation (1) include molecular surface area, calculated as part of the COSMO solvation scheme (Area_Cosmo), and Mulliken partial charges on Pt, $q(\text{Pt})$, and the most positively charged H-atom (max $q\text{H}$). The second model, denoted QC2, used exposed surface area-based descriptors, including total (TSA), polar (PSA) and platinum (PtSA)

$$\log P = -3.39 + 0.0099 \cdot \text{TSA} - 0.0171 \cdot \text{PSA} + 0.0543 \cdot \text{PtSA} \quad (4)$$

where all surface areas have units of \AA^2 . It was calculated from PM6-optimised geometries using our in-house modification of MOLVOL,[48] and was trained on $\log P$ for 24 Pt(II) complexes [14].

VCCLAB-Pt model

This model was developed in our previous study[16] with $N=65$ Pt(II) complexes. It is based on the ALOGPS 2.1 program,[21] which was initially developed to predict $\log P$ of organic molecules. Of course, the original version of this program was unable to predict metal-containing compounds and calculated $\text{RMSE} > 2$ log units. However, once it was augmented with $\log P$ data for $N=43$ Pt(II) complexes, it provided good prediction accuracy ($\text{RMSE} = 0.55$) for $N=12$ complexes blindly tested in the University of Wroclaw. The model is publicly available at <http://www.vcclab.org/web/pt> site.

Results and discussion

Performance of models based on combined set of Pt(II) and Pt(IV) complexes

For this analysis we used the combined set, containing both Pt(II) and Pt(IV) complexes. The RMSE calculated for different descriptors sets by using two methods, ASNN and MLRA, are shown in Table 2. As it is clear from the comparison of results, utilization of neural networks calculated lower RMSE compared to MLRA. The use of bagging further increased performance of models developed using ASNN. Only three types of descriptors, ChemAxon, Inductive and Adriana, were based on 3D structures. Models based on these descriptors had on average lower accuracies

compared to those based on 2D descriptors. This result may be connected with difficulties to generate 3D structures for platinum complexes and/or difficulty with some of these descriptor packages to work with coordination bonds.

The lowest RMSE=0.45±0.03 (N=187) was calculated using the extended functional groups (EFG)[49] descriptors and bagging. These descriptors count a number of functional groups in molecule. The initial functional groups were developed based on CheckMol program [50]. They were substantially extended to cover various chemical classes and to better represent hetero-aromatic molecules [49]. In total 583 functional groups are available at OCHEM web site as part of ToxAlerts [51]. The Pt(II) and Pt(IV) complexes from the training set were represented with 69 different functional groups. After the unsupervised filtering 40 functional groups were retained for model development.

The second best model with RMSE=0.46±0.04 was calculated with Isida Fragmentor descriptors,[52] developed at the *Laboratoire de Chemoinformatique* of the University of Strasbourg. In this approach each compound is split into Substructural Molecular Fragments (SMF). In our analysis SMF of lengths 2 to 4 were used. Each fragment type comprises a descriptor, with the number of occurrences of the fragment type as the respective descriptor value. In this study, we used the sequence fragments composed of atoms and bonds.

E-state indices[53] are one of the most successful type of descriptors, which was used to model logP of chemical compounds since the last century [54]. These indices combine electronic and topological properties of the analyzed molecules. The indices are calculated for atom or atomic bonds and are summed over the same atom/bond types in a molecule for modeling. In this study we used an extended set of E-state indices,[55] which was developed to better cover the environment of amino, hydroxyl, and carbonyl groups of molecules. E-state indices types as well as Fragmentor

descriptors are generated automatically based on the atoms present in the molecules. Thus, they also covered Pt atoms and bond types in both Pt(II) and Pt(IV) complexes.

The combination of three models built on EFG, Fragmentor descriptors and E-state indices in a consensus model provided the lowest RMSE = 0.45 ± 0.03 log units for the combined set (for the cross-validation protocol). The same combination of models for bagging calculated lower RMSE = 0.41 ± 0.03 . An attempt to further improve the accuracy of this model by including other models into the consensus was not successful. This was also the case when a combined set of descriptors, which includes E-state, Fragmentor descriptors and EFG was used. For this analysis the descriptors from these three sets were merged in one and models were developed using the same protocols as applied for the individual sets. The error of these new models RMSE = 0.55 ± 0.04 (for CV RMSE = 0.47 ± 0.04 for bagging) were larger compared to the respective consensus RMSEs. Thus, for all further studies we used consensus models built on EFG, Fragmentor descriptors and E-state indices.

Comparison of models' performances based on the different sets of molecules

We developed models using individual sets consisting of only Pt(II) or Pt(IV) complexes. In general, the RMSE calculated for Pt(II) data were significantly higher compared to those calculated for the Pt(IV) set (see Figures S2 and S3). These results may indicate the higher heterogeneity of data in the Pt(II) set as compared to those in the Pt(IV) set. For CV models, the RMSE = 0.52 ± 0.04 (N=100) calculated for dataset with Pt(II) compounds was not significantly different compared to RMSE = 0.48 ± 0.03 , calculated for the subset with Pt(II) complexes using the combined Pt(II) + Pt(IV) dataset. For Pt(IV) complexes (N=87) very similar RMSE values 0.38 ± 0.04 and 0.40 ± 0.04 were calculated for models developed using the Pt(IV) dataset and for the Pt(IV) subset of the combined set. Thus, development of individual models based on each subset or development of a

single combined model based on both Pt sets provided similar performance. Similar performance of models developed using the individual Pt sets and the combined set were also observed for the bagging validation protocol.

We have further investigated this finding by developing a model, extending the combined Pt set with 12898 data points previously used to develop ALOGPS 2.1 program [21]. Since this dataset was much larger, we developed only CV models for it. The consensus model with RMSE = 0.37 ± 0.01 had a similar performance to the models developed only for Pt complexes. It had CV RMSE of 0.55 ± 0.08 and 0.41 ± 0.03 for Pt(II) and Pt(IV) complexes, respectively. Thus, like with the previous analysis, the RMSE of models developed with even larger set were not significantly different.

The model developed for Pt(II) complexes in this study was based on the largest published set of Pt(II) complexes (N=100). Its performance RMSE = 0.48 was similar to the results, RMSE = 0.43 (N=43) and RMSE = 0.61 (N=12) for training and test sets, respectively, calculated in our previous study (VCCLAB-Pt model) [16].

Identification of outlying compounds

Several compounds appeared to have large errors in different models. In total 11 molecules had absolute errors between predicted and measured experimental values more than 1 log unit for the analyzed consensus models. An appearance of molecules with large differences between predicted and calculated values is expected considering statistical properties of the distribution of errors in the developed models. Indeed, most of molecules were outliers in one or two models only. However, two molecules (*SP-4-3*)-(ethane-1,2-diamine)(2-ethoxy-3-carboxypropanoato)platinum(II) and (*SP-4-1*)-diiodidobis(*N*-methyl-5-nitroimidazole)platinum(II) were consistent outliers across the models. This result may signal some problems with experimental measurements for these compounds.

Analysis of predicted Molecular Matched Pairs

To further understand this result, we analyzed the Molecular Matched Pairs (MMPs) plot[56] for the consensus model, ASNN3, developed with the combined Pt(II) + Pt(IV) set. While MMPs is a widely used method in drug discovery,[57] predicted MMPs[56] is a new tool to provide analysis of the developed models. MMP indicates the change of activity of a molecule due to a change in a single group. They are grouped by transformations, which indicate chemical groups that are changed. Thus, MMPs are related to the analyzed set of molecules with experimental values. Predicted MMP indicate the change in the activity as calculated with the corresponding model. They are related to a model and can be calculated for any set of molecules following an application of the model. The calculation of MMPs is based on the fragmentation and indexing of molecules using an approach developed by Hussain and Rea. [58] In order to prevent combinatorial explosion, only molecules with less than 40 rotatable bonds are considered by the algorithm while the variable part of the molecules is restricted to 10 atoms [56]. Apart from these limitations there are no other parameters of the algorithm. Technically, the MMPs for models are available in the lower left corner of each model. In case if some molecules are not yet indexed, the user can manually submit them using the respective button of the MMP plot.

The points close to the diagonal line indicate MMPs (Figure 2), which were correctly learnt and thus were reproduced by the model. The model correctly predicted the sign of the change of logP for MMPs in quadrants one and three. The closer is a point to the diagonal of these quadrants the better it was reproduced by the model. On the contrary, the points in quadrants two and four indicate predicted MMPs for which the model incorrectly calculated the direction of the change of the logP of a molecule.

Both outlying molecules contributed a number of MMPs in second and fourth quadrants, which were thus predicted by the model to have an opposite change in activity. One such MMP corresponds to transformation with substitution of hydrogen [H] to ethyl [CC] group (Figure 3). There were 14 MMPs corresponding to this transformation in the analyzed dataset. The consensus model learnt this transformation and correctly predicted direction of the change of the logP values for 13 out of 14 MMPs. These MMPs are shown as green dots in the first quadrant of the plot. For MMP of (SP-4-3)-(ethane-1,2-diamine)(2-hydroxy-3-carboxypropanoato)platinum(II) to (SP-4-3)-(ethane-1,2-diamine)(2-ethoxy-3-carboxypropanoato)platinum(II) shown on the plot, this MMP is shown to decrease experimentally measured logP values. The corresponding MMP is highlighted in the fourth quadrant of the plot thus signaling that the model predicted a different sign for this transformation. Addition of two carbons increases logP on average by 0.95 log units (based on 1295 MMPs for ALOGPS2.1 dataset), but for this particular pair it reduced logP by 1.24 log units. Such change seems likely to be contributed by an experimental error in the measured values. While we could have suspected such problems, analysis of predicted MMPs provided additional structural information to confirm this observation. The plot of predicted versus experimental MMPs allowed easy identification of the suspicious transformation, which is likely to be due to experimental error with measurement of logP value for (SP-4-3)-(ethane-1,2-diamine)(2-ethoxy-3-carboxypropanoato)platinum(II).

However, not all outlying compounds indicate erroneous data. For the other outlying molecule, MMPs identified that change of (SP-4-2)-dichloridobis(*N*-methyl-5-nitroimidazole)platinum(II) (experimental logP = -0.49) to (SP-4-1)-diiodidobis(*N*-methyl-5-nitroimidazole)platinum(II) (experimental logP = 0.95) increased logP values by 1.44 log units, whereas the model predicted a decrease of -0.12 log units. In this transformation, two [Cl] atoms attached to Pt(II) were changed to two [I] atoms. The dataset did not contain other MMPs with this transformation and there were only four molecules with [I] in the whole dataset. The much larger ALOGPS 2.1 set contains about

100 MMPs with [Cl] -> [I] transformation. This transformation on average increased logP value by 0.4 log units thus contributing about 0.8 log units for the change of two Cl atoms. This difference is comparable with observed $\Delta\log P = 1.44$. Moreover, the measured logP values were for complexes with different geometry (*cis* and *trans*), which could also contribute further difference in logP values. It is interesting that ALOGPS 2.1 correctly learnt [Cl] -> [I] MMPs, the majority of which were in the first quadrant of the plot. Thus, in the current example the limited experimental data were not sufficient to correctly learn the observed dependencies or even their sign. While it is still possible that one or both of the experimental values could contain experimental errors, the provided analysis is not sufficient to exclude one of the molecules.

Indeed, if we consider the same both MMPs for model developed using the combined set of ALOGPS 2.1 + Pt(II) + Pt(IV) compounds, the first MMP (*SP-4-3*)-(ethane-1,2-diamine)(2-hydroxy-3-carboxypropanoato)platinum(II) to (*SP-4-3*)-(ethane-1,2-diamine)(2-ethoxy-3-carboxypropanoato)platinum(II)) still stays as an outlier. The logP change due to the second MMP ((*SP-4-2*)-dichloridobis(*N*-methyl-5-nitroimidazole)platinum(II) to (*SP-4-1*)-diiodidobis(*N*-methyl-5-nitroimidazole)platinum(II)) is, however, correctly predicted with this model and the corresponding predicted MMP appears near the diagonal of the first quadrant of the plot. Thus, the use of larger set of molecules allowed the model to better learn dependencies for underrepresented transformations available in the data. This result may suggest that models built on the larger set of molecules could potentially have a more robust prediction accuracy compared to those developed with more focused set of molecules.

Analysis of MLRA models

The MLRA models, as exemplified by Table 2, had lower accuracy compared to ASNN. Some of the differences in performances between ASNN and MLRA models are dramatic. For example, the model based on E-state indices was the third best model for ASNN method (RMSE=0.54) while it was the worst for MLRA method (RMSE=20). At least for some descriptors, low accuracy of

MLRA models stemmed from a few predictions with very large positive or negative logP values. For example model based on E-state indices predicted logP of 315 and 33.2 for (*SP-4-2*)-dichloridobis(1,2-dimethyl-5-nitroimidazole)platinum(II) and (*SP-4-2*)-dichloridobis(*N*-methyl-5-bromoimidazole)platinum(II) which have logP of -0.33 and -0.08, respectively. Such large errors can be contributed to problems with small number of compounds in the training set and overfitting of MLRA models due to variable selection. The estimation of applicability domain for these molecules using Leverage clearly indicated that such prediction had large leverage values and thus were non-reliable ones. Another way to deal with this problem could be to limit the range of predicted logP values to that of the training set. If such restriction was added, RMSE of E-state model decreased to 1.1 log unit.

The best models calculated using MLRA were based on CDK[58] and EFG[49] descriptors. The CDK-based model was based on 18 descriptors, which also included prediction of lipophilicity using ALogP algorithm. The model based on functional group counts included only nine variables and thus it is more easily interpretable.

$$\begin{aligned} \log P = & -2.25 + 3.85*(\text{unsaturated six-membered heterocycles with two heteroatoms}) + \\ & 0.176*(\text{carboxylic acid derivatives}) + 0.119*(\text{halogen derivatives (alkyl or aryl)}) + \\ & 0.0816*(\text{non-metal atoms}) + 0.0621*(\text{aromatic atoms}) - 0.538*(\text{carboxylic acid amides}) - \\ & 0.466 * (\text{hydroxy compounds: alcohols or phenols}) - 0.386*(\text{six-membered heterocycles with} \\ & \text{three heteroatoms}) - 0.316*(\text{chalcogens (oxygen group)}) \end{aligned} \quad (5)$$

In the above equation (5) the coefficients near to each group are proportional to the contribution of the corresponding group to the logP. The coefficients look logical and reflect an intuitive expectation for the contribution of respective groups to the total lipophilicity.

Benchmarking of models for prediction of new molecules

To further assess the predictive ability of models developed here, and to compare them to models from our previous work, we applied each model to a test set of complexes (see Table 1) not used in training developed in our previous publications. The first three models were applied “as is” and no tuning of their parameters was performed for this study.

Performance of VCCLAB-Pt model

This model performed well for prediction of Pt(II) complexes, but failed to accurately predict Pt(IV) complexes. This result was expected since this model did not have any Pt(IV) complexes in its training set and thus it was unable to correctly account for this class of compounds.

Performance of models based on quantum chemical descriptors

Table 3 reports RMSE for two models based on quantum chemical descriptors, built on 3D structures from PM6 optimization. In general, model QC1 performs poorly with RMSE over 2 log units for both test sets. This model employs electronic descriptors such as dipole moment and atomic partial charges, and was specifically designed for Pt(IV) complexes. As such, the poor performance observed here is slightly surprising. Model QC2, which is based on exposed surface areas and was initially developed for Pt(II) complexes, performs rather better, though still resulting in rather large RMSE. Within the test set, significant variation in the quality of prediction of values from shake flask (RMSE = 0.47) and HPLC (RMSE = 1.06) is evident. Interestingly, this model performs at the same level for Pt(II) and Pt(IV) complexes, despite the fact that no Pt(IV) complexes were used in training sets for this approach.

Performance of models developed in this study

The models calculated good accuracy of predictions for both Pt(II) and Pt(IV) complexes. The accuracy of prediction for Pt(II) complexes was higher compared to that for Pt(IV). Actually, for

Pt(II) complexes all models provided similar prediction accuracy. The accuracy of prediction was lower for Pt(IV) data, with the best models based on Pt(IV) or their combination with Pt(II) data. The extension of the dataset with logP values of 12.9k organic compounds did not provide significantly improved RMSE from the best model. It is interesting that the model developed using only values for Pt(II) complexes did not fail completely, as was the case with VCCLAB-Pt. We assign this difference to the fact that models developed in this study used new descriptors to account for the molecular environment, whereas the previous VCCLAB-Pt model was based on similarity, which may not work so well for extrapolation.

MMP analysis allowed identification of several non-expected changes to logP, which can be possibly attributed to some experimental problems with data. For example the change of CH₃ → phenyl (**26** to **27**) decreased logP by 0.29. In general, phenyl, which is more lipophilic fragment compared to single atom of CH₃ increases logP. Another issue arose with three complexes **30–32**. In this case, a decrease of logP with the increase of the alkyl chain length connected to benzene was observed, which does not fit the usual pattern. Actually, these three compounds had very low solubility and even though all experiments were performed with maximum care, the measurement of their logP was difficult, poorly reproducible and required the use of co-solvent (5% DMSO) to avoid precipitation. The co-solvent could introduce some bias in the measurements: even though present in low percentage, it might affect the partitioning and therefore the final logP value of the Pt complexes, being themselves partitioned according to their logP. Indeed, DMSO has a negative logP = -1.35 [55] as it is presented more in the water than in the octanol phase. Therefore, the partitioning of analyzed compounds in DMSO can shift them from the octanol to the water phase and thus can artificially decrease the apparent logP values, determined by means of the shake-flask method as compared to the true values. The strength of this effect will depend on the relative solubility of the investigated compound in the analyzed media, i.e., water, octanol and DMSO. The computational models were developed using experimental values measured in pure octanol/water

phase and thus were unbiased with respect to the presence of DMSO. Therefore their predictions could be used to confirm this effect. All four models, ASNN1-ASNN4, as well as both quantum-chemistry and VCCLAB-Pt models predicted higher logP (the average differences in logP values were $\Delta\log P > 1$) than the observed experimental values for compounds **30-32**. This result is in agreement with the hypothesis about the influence of DMSO on logP measurements. If we assume that the predicted experimental values for complexes **30-32** have the same accuracy as for other test set compounds, \sim RMSE = 0.6 - 0.7 log units, the observed difference of >1 log unit is dominated by the effect of co-solvent. Considering that DMSO is frequently used as a solvent for storage of chemical compounds, its presence during the experimental logP measurements can significantly bias them. This hypothesis requires more careful experimental investigation.

If we exclude these three compounds the RMSE of the best performing for Pt(IV) set models, ASNN1 and ASNN3, decreased to RMSE = 0.58, such that their performance became similar to that for Pt(II) complexes.

Influence of chiral centers

The statistical methods used for logP modeling (Estate indices, fragmentor and EFG) only use 2D representation of molecules and thus are insensitive to differences in stereochemistry and do not distinguish stereoisomers (for, e.g. isomeric complexes with N,N' disubstituted ethylenediamine) respectively. However, we also addressed this effect for 3D algorithms. MOPAC PM6-DH2 geometry optimisations were performed on cis and trans isomers of [Pt(1,4-dimethyl-en)Cl₂] (compound **6** from the test set). The trans form is 2.8 kJ mol⁻¹ more stable than cis, at the semi-empirical level used. Descriptors calculated from these calculations varied only very slightly between isomers: the volume of the trans isomer is 1.3 Å³ less than that of the cis (actual volumes are cis 213.29, trans 211.95 Å³ so the change is 0.6%), while atomic charges differ by less than 0.01 e. This leads to a change in predicted logP of 0.24 for QC1 model, i.e. significantly less than error

associated with predictive models. However, this result can be of interest for some practical applications of both diastereoisomers.

Development of the final model

The final model was developed using all Pt data available in this study with an exception of complexes 30-32 from the test set and ((*SP*-4-3)-(ethane-1,2-diamine)(2-ethoxy-3-carboxypropanoato)platinum(II) from the training set, which may have problems with experimental values as described above. It has CV RMSE = 0.39 ± 0.02 that is similar to the RMSE = 0.41 ± 0.03 calculated using the training set. Since it was developed with the biggest set of Pt complexes, the model has the largest applicability domain [59] and can be used to predict properties of novel Pt(II) and Pt(IV) complexes.

Conclusions

The blind benchmarking of models developed to predict logP of Pt(II) and Pt(IV) complexes was performed using data coming from two different experimental laboratories. The calculated results showed that existing modeling techniques allow an accurate prediction of logP for platinum complexes. We studied two approaches to predict logP of Pt compounds: based on representation of chemical structures using general sets of descriptors as well as those exploring quantum-chemistry parameters of molecules. In this study the models based on quantum-chemistry descriptors provided lower prediction ability compared to those based on simpler descriptors. The poor performance of the QC models seems to be related to a lack of diversity in the original training sets, which did not contain several of the functional groups present in the test set used here.

The local models developed with Pt data only (ASNN1 - ASNN3) had lower RMSE for prediction of test set compounds compared to the model based on combined set of organic molecules and Pt complexes (ASNN4) for Pt(IV) complexes. However, similar accuracy of both types of models was

observed for Pt(II) complexes. Thus, this study was unable to conclude which approach, i.e., development of global models using diverse sets of molecules or just making focused models based on compounds sharing same framework or scaffold, is the better one. More data and more diverse dataset are required to make conclusions about the advantages of each methodology.

The MMP approach allowed an easy explanation of outlying compounds. While one can easily find errors on the predicted versus experimental plot, their explanation can be difficult. An analysis of such molecules with help of chemical transformation patterns allows an easy identification and explanation of outlying molecules, as it was shown in this work.

We also concluded that addition of DMSO prior to shake-flask experiments could lower octanol/water partition coefficients compared to the true values measured in the absence of the co-solvent. The observed differences between predicted and measured logP values, which were on average more than one log unit, confirmed the theoretical analysis. Additional studies are required to verify whether the presence (as residual solvent, which was used to store chemical compounds) or intended use of DMSO as a co-solvent can significantly bias the measured logP values.

The main motivation of this study was to develop a publicly available model for prediction of logP of Pt(IV) complexes. The goal was fully achieved and the developed model is publicly available at <http://ochem.eu/article/76903>. It can be used to predict logP of new platinum (both Pt(II) and Pt(IV)) complexes as well as to screen large virtual libraries of Pt complexes. Moreover, we published 45 new logP values for Pt complexes, i.e. contributed almost 25% of new values. The academic community can use these values to develop new and/or compare and improve the existing approaches. Publishing of models and providing online data is an important way to share the results of research, to make them available to academy and industry and to allow their re-use by the readers will contribute certainly to the development of the field of computational chemistry [60].

Table of Abbreviations

ASNN	Associative Neural Networks
CV	Five fold cross-validation
MLRA	Multiple Linear Regression Analysis
MMP	Molecular Matched Pair
OCHEM	On-line Chemical Modeling Environment
r^2	Coefficient of determination
RMSE	Root Mean Squared Error
EFG	Extended Functional Groups

Acknowledgements

The authors would thank Mrs. Tanzem Haque for her help with data collection. We also thank Dr. Yuri Sushko, Dr. Sergey Novotarskyi and Mr. Robert Körner for their work on the development of OCHEM platform and collection of some Pt data, which made this study possible. Mag. Sarah Theiner is acknowledged for her assistance in ICP-MS measurements. H.V is thankful for the financial support of the Austrian Science Fund (FWF, Schrödinger fellowship J3577-B13).

References

- [1] L. Kelland *Nat. Rev. Cancer.* 7 (2007) 573-584.
- [2] Y. Jung, S.J. Lippard *Chem. Rev.* 107 (2007) 1387-1407.
- [3] M.R. Reithofer, A.K. Bytzeck, S.M. Valiahdi, C.R. Kowol, M. Groessl, C.G. Hartinger, M.A. Jakupec, M. Galanski, B.K. Keppler *J. Inorg. Biochem.* 105 (2011) 46-51.
- [4] S.P. Oldfield, M.D. Hall, J.A. Platts *J. Med. Chem.* 50 (2007) 5227-5237.
- [5] H. Varbanov, S.M. Valiahdi, A.A. Legin, M.A. Jakupec, A. Roller, M. Galanski, B.K. Keppler *Eur. J. Med. Chem.* 46 (2011) 5456-5464.
- [6] P. Gramatica, E. Papa, M. Luini, E. Monti, M.B. Gariboldi, M. Ravera, E. Gabano, L. Gaviglio, D. Osella *J. Biol. Inorg. Chem.* 15 (2010) 1157-1169.
- [7] X. Liu, B. Testa, A. Fahr *Pharm. Res.* 28 (2011) 962-977.
- [8] C.A. Lipinski, F. Lombardo, B.W. Dominy, P.J. Feeney *Adv. Drug Deliv. Rev.* 46 (2001) 3-26.
- [9] A. Leo, C. Hansch, D. Elkins *Chem. Rev.* 61 (1971) 525-616.
- [10] C. Hansch *Acc. Chem. Res.* 2 (1969) 232-239.
- [11] G.L. Patrick, J. Spencer (Ed.), *An introduction to medicinal chemistry*, Oxford University Press, New York, 2009.
- [12] M.D. Hall, T.W. Hambley *Coord. Chem. Rev.* 232 (2002) 49-67.
- [13] C.F. Chin, Q. Tian, M.I. Setyawati, W. Fang, E.S. Tan, D.T. Leong, W.H. Ang *J. Med. Chem.* 55 (2012) 7571-7582.
- [14] J.A. Platts, D.E. Hibbs, T.W. Hambley, M.D. Hall *J. Med. Chem.* 44 (2001) 472-474.
- [15] J.A. Platts, S.P. Oldfield, M.M. Reif, A. Palmucci, E. Gabano, D. Osella *J. Inorg. Biochem.* 100 (2006) 1199-1207.
- [16] I.V. Tetko, I. Jaroszewicz, J.A. Platts, J. Kuduk-Jaworska *J. Inorg. Biochem.* 102 (2008) 1424-1437.
- [17] J.A. Platts, G. Ermondi, G. Caron, M. Ravera, E. Gabano, L. Gaviglio, G. Pelosi, D. Osella *J. Biol. Inorg. Chem.* 16 (2011) 361-372.

- [18] G. Ermondi, G. Caron, M. Ravera, E. Gabano, S. Bianco, J.A. Platts, D. Osella Dalton. Trans. 42 (2013) 3482-3489.
- [19] A.A. Toropov, A.P. Toropova, E. Benfenati J. Math. Chem. 46 (2009) 1060-1073.
- [20] P. Sarmah, R.C. Dea J. Comput. Aided. Mol. Des. 23 (2009) 343-354.
- [21] I.V. Tetko, V.Y. Tanchuk J. Chem. Inf. Comput. Sci. 42 (2002) 1136-1145.
- [22] A.M. Montana, C. Batalla Curr. Med. Chem. 16 (2009) 2235-2260.
- [23] I. Sushko, S. Novotarskyi, R. Korner, A.K. Pandey, M. Rupp, W. Teetz, S. Brandmaier, A. Abdelaziz, V.V. Prokopenko, V.Y. Tanchuk, R. Todeschini, A. Varnek, G. Marcou, P. Ertl, V. Potemkin, M. Grishina, J. Gasteiger, C. Schwab, I.I. Baskin, V.A. Palyulin, E.V. Radchenko, W.J. Welsh, V. Kholodovych, D. Chekmarev, A. Cherkasov, J. Aires-de-Sousa, Q.Y. Zhang, A. Bender, F. Nigsch, L. Patiny, A. Williams, V. Tkachenko, I.V. Tetko J. Comput. Aided. Mol. Des. 25 (2011) 533-554.
- [24] M.R. Reithofer, S.M. Valiahdi, M.A. Jakupec, V.B. Arion, A. Egger, M. Galanski, B.K. Keppler J. Med. Chem. 50 (2007) 6692-6699.
- [25] H.P. Varbanov, S.M. Valiahdi, C.R. Kowol, M.A. Jakupec, M. Galanski, B.K. Keppler Dalton. Trans. 41 (2012) 14404-14415.
- [26] H.P. Varbanov, M.A. Jakupec, A. Roller, F. Jensen, M. Galanski, B.K. Keppler J. Med. Chem. 56 (2013) 330-344.
- [27] H.P. Varbanov, S. Goschl, P. Heffeter, S. Theiner, A. Roller, F. Jensen, M.A. Jakupec, W. Berger, M. Galanski, B.K. Keppler J. Med. Chem. 57 (2014) 6751-6764.
- [28] M. Ravera, E. Gabano, S. Bianco, G. Ermondi, G. Caron, M. Vallaro, G. Pelosi, I. Zanellato, I. Bonarrigo, C. Cassino, D. Osella Inorg. Chim. Acta 432 (2015) 115-127.
- [29] V. Gandin, C. Marzano, G. Pelosi, M. Ravera, E. Gabano, D. Osella ChemMedChem 9 (2014) 1299-1305.
- [30] G. Caron, M. Ravera, G. Ermondi Pharm. Res. 28 (2011) 640-646.

- [31] M. Milanesio, E. Monti, M.B. Gariboldi, E. Gabano, M. Ravera, D. Osella *Inorg. Chim. Acta* 361 (2008) 2803-2814.
- [32] A. Ghezzi, M. Aceto, C. Cassino, E. Gabano, D. Osella *J. Inorg. Biochem.* 98 (2004) 73-78.
- [33] P.D. Braddock, T.A. Connors, M. Jones, A.R. Khokhar, D.H. Melzack, M.L. Tobe *Chem. Biol. Interact.* 11 (1975) 145-161.
- [34] R. Kizu, T. Nakanishi, K. Hayakawa, A. Matsuzawa, M. Eriguchi, Y. Takeda, N. Akiyama, T. Tashiro, Y. Kidani *Cancer Chemother. Pharmacol.* 43 (1999) 97-105.
- [35] M. Coluccia, A. Nassi, A. Boccarelli, D. Giordano, N. Cardellicchio, D. Locker, M. Leng, M. Sivo, F.P. Intini, G. Natile *J. Inorg. Biochem.* 77 (1999) 31-35.
- [36] OECD Guidelines for the Testing of Chemicals, Section 1. Physical-Chemical properties.
<http://dx.doi.org/10.1787/20745753>.
- [37] S. Theiner, H.P. Varbanov, M. Galanski, A.E. Egger, W. Berger, P. Heffeter, B.K. Keppler *J. Biol. Inorg. Chem.* 20 (2015) 89-99.
- [38] I.V. Tetko *Neural Process. Lett.* 16 (2002) 187-199.
- [39] I.V. Tetko *J. Chem. Inf. Comput. Sci.* 42 (2002) 717-728.
- [40] I.V. Tetko *Methods Mol. Biol.* 458 (2008) 185-202.
- [41] S. Vorberg, I.V. Tetko *Mol. Inf.* 33 (2014) 73-85.
- [42] I.V. Tetko, Y. Sushko, S. Novotarskyi, L. Patiny, I. Kondratov, A.E. Petrenko, L. Charochkina, A.M. Asiri *J. Chem. Inf. Model.* 54 (2014) 3320-3329.
- [43] I.V. Tetko, V.P. Solov'ev, A.V. Antonov, X. Yao, J.P. Doucet, B. Fan, F. Hoonakker, D. Fourches, P. Jost, N. Lachiche, A. Varnek *J. Chem. Inf. Model.* 46 (2006) 808-819.
- [44] L. Breiman *Machine Learn.* 24 (1996) 123-140.
- [45] Online Demo - Fast 3D Structure Generation with CORINA. http://www.molecular-networks.com/online_demos/corina_demo.
- [46] P. Ertl, B. Rohde, P. Selzer *J. Med. Chem.* 43 (2000) 3714-3717.
- [47] J.J. Stewart *J. Mol. Model.* 13 (2007) 1173-1213.

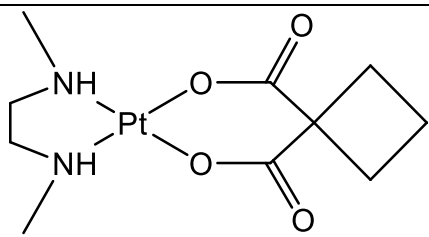
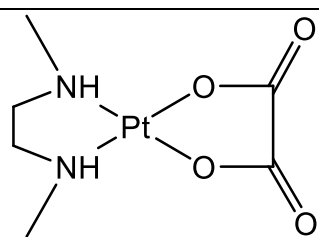
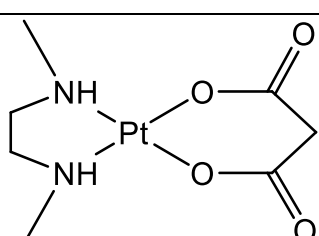
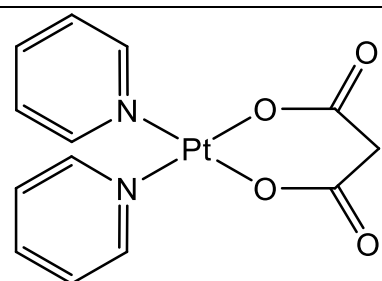
- [48] R.A. Saunders, J.A. Platts *New J. Chem.* 28 (2004) 166-172.
- [49] E. Salmina, N. Haider, I.V. Tetko *Molecules* (2015).
- [50] N. Haider *Molecules* 15 (2010) 5079-5092.
- [51] I. Sushko, E. Salmina, V.A. Potemkin, G. Poda, I.V. Tetko *J. Chem. Inf. Model.* 52 (2012) 2310-2316.
- [52] A. Varnek, D. Fourches, D. Horvath, O. Klimchuk, C. Gaudin, P. Vayer, V. Solov'ev, F. Hoonakker, I.V. Tetko, G. Marcou *Curr. Comp. Aid. Drug Design* 4 (2008) 191-198.
- [53] L.H. Hall, L.B. Kier *J. Chem. Inf. Comput. Sci.* 35 (1995) 1039-1045.
- [54] J.J. Huuskonen, A.E. Villa, I.V. Tetko *J. Pharm. Sci.* 88 (1999) 229-233.
- [55] J.J. Huuskonen, D.J. Livingstone, I.V. Tetko *J. Chem. Inf. Comput. Sci.* 40 (2000) 947-955.
- [56] Y. Sushko, S. Novotarskyi, R. Korner, J. Vogt, A. Abdelaziz, I.V. Tetko *J. Cheminform.* 6 (2014) 48.
- [57] A.G. Dossetter, E.J. Griffen, A.G. Leach *Drug Discov. Today* 18 (2013) 724-731.
- [58] C. Steinbeck, Y. Han, S. Kuhn, O. Horlacher, E. Luttmann, E. Willighagen *J. Chem. Inf. Comput. Sci.* 43 (2003) 493-500.
- [59] I.V. Tetko, P. Bruneau, H.W. Mewes, D.C. Rohrer, G.I. Poda *Drug Discov. Today* 11 (2006) 700-707.
- [60] I.V. Tetko *J. Comput. Aided. Mol. Des.* 26 (2012) 135-136.
- [61] N.I. Zhokhova, I.I. Baskin, V.A. Palyulin, A.N. Zefirov, N.S. Zefirov *Dokl. Chem.* 417 (2007) 282-284.
- [62] V. Potemkin, M. Grishina *Drug Discov. Today* 13 (2008) 952-959.
- [63] A. Cherkasov *Curr. Comp. Aid. Drug Design* 1 (2005) 21-42.
- [64] J. Gasteiger *J. Med. Chem.* 49 (2006) 6429-6434.
- [65] R. Todeschini, V. Consonni (Ed.), *Handbook of Molecular Descriptors*, WILEY-VCH, Weinheim, 2000.
- [66] A. Bender, H.Y. Mussa, R.C. Glen, S. Reiling *J. Chem. Inf. Comput. Sci.* 44 (2004) 1708-1718.

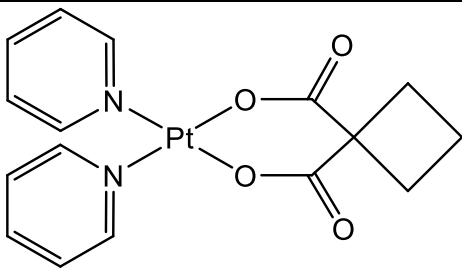
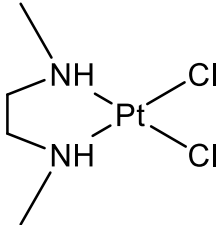
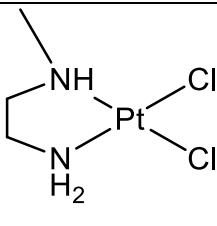
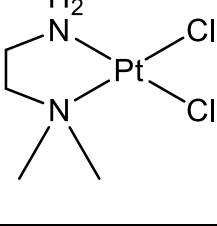
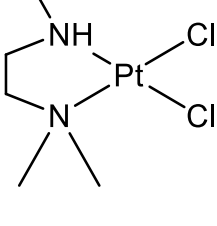
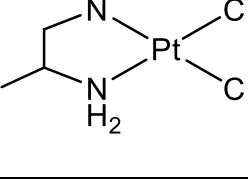
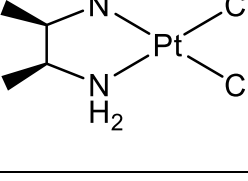
[67] D. Rogers, M. Hahn J. Chem. Inf. Model. 50 (2010) 742-754.

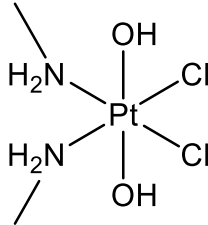
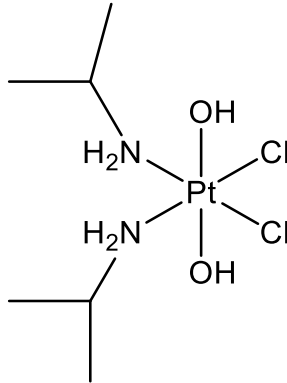
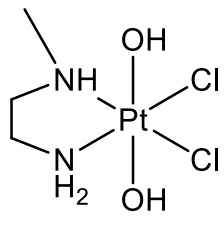
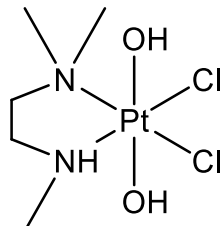
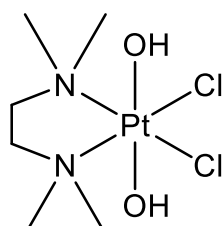
[68] L.H. Hall, L.M. Hall SAR QSAR Environ. Res. 16 (2005) 13-41.

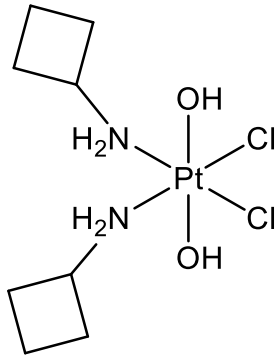
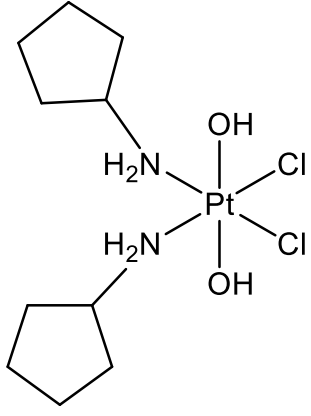
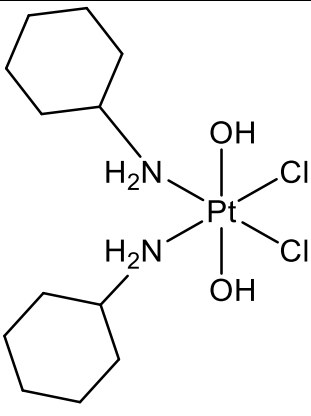
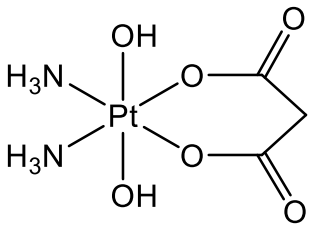
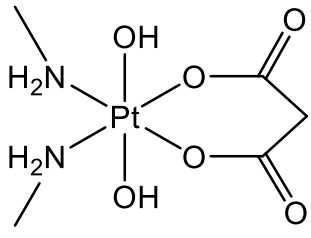
[69] J. Sadowski, J. Gasteiger, G. Klebe J. Chem. Inf. Comput. Sci. 34 (1994) 1000-1008.

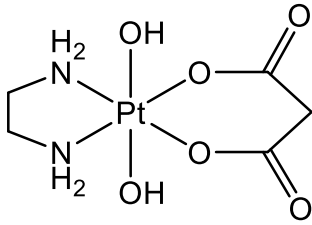
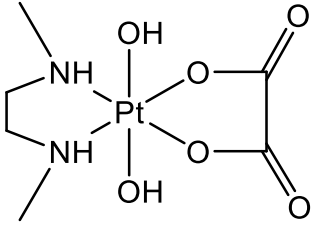
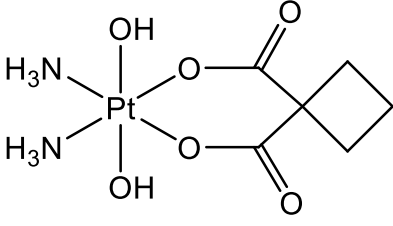
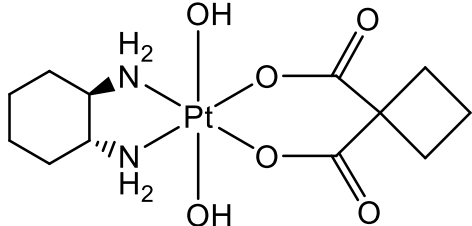
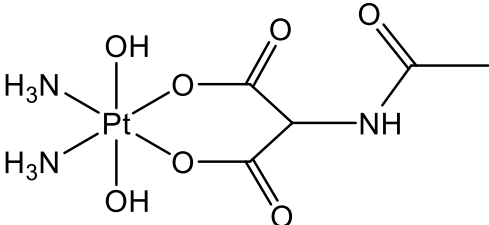
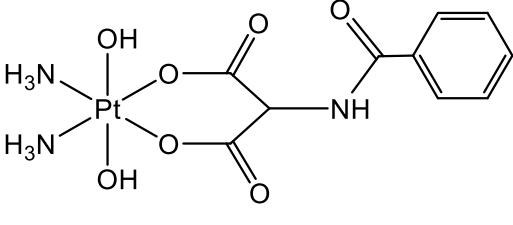
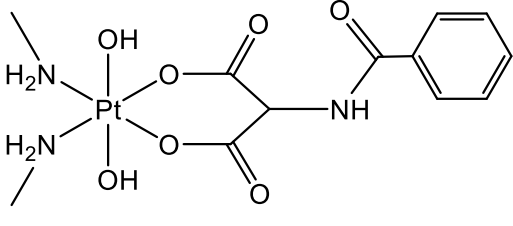
Table 1. Pt (II/IV) complexes, together with their experimental logP values, used as the benchmarking dataset for the models.

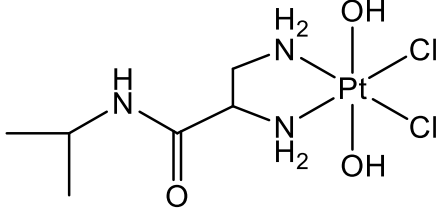
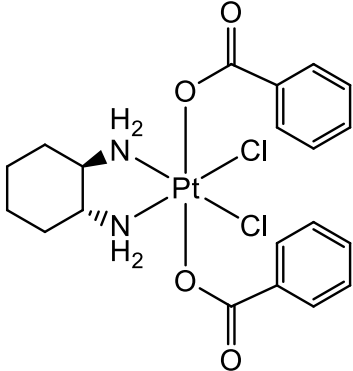
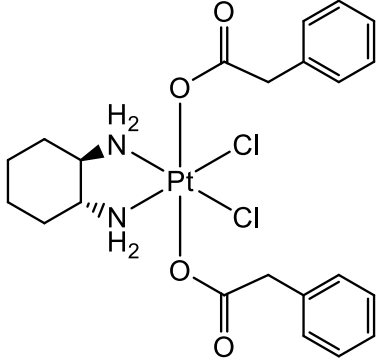
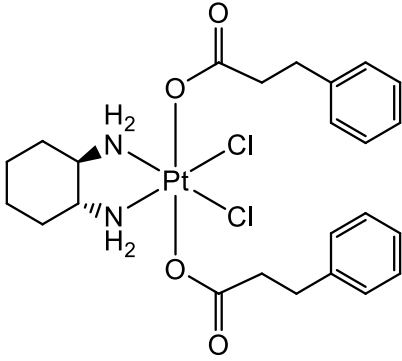
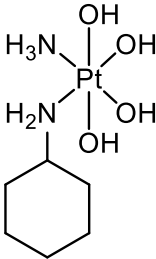
N	Name	Formula	LogP
Pt(II) complexes			
1 ^a	(<i>SP-4-2</i>)-(cyclobutane-1,1-dicarboxylato)(<i>N,N'</i> -dimethylethane-1,2-diamine)platinum(II)		-1.72 ± 0.06
2 ^a	(<i>SP-4-2</i>)-(N,N'-dimethylethane-1,2-diamine)ethanediatoplatinum(II)		-2.36 ± 0.36
3 ^a	(<i>SP-4-2</i>)-(N,N'-dimethylethane-1,2-diamine)propanedioatoplatinum(II)		-2.15 ± 0.13
4	(<i>SP-4-2</i>)-bis(pyridine)propanedioatoplatinum(II)		-1.08 ± 0.09

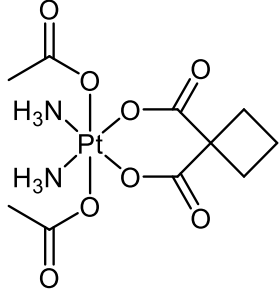
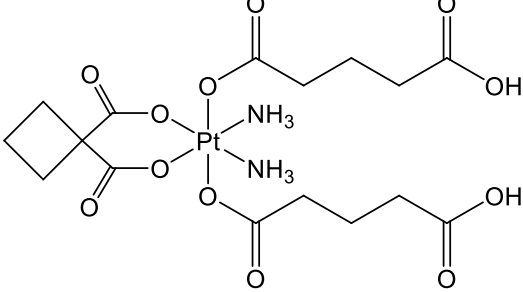
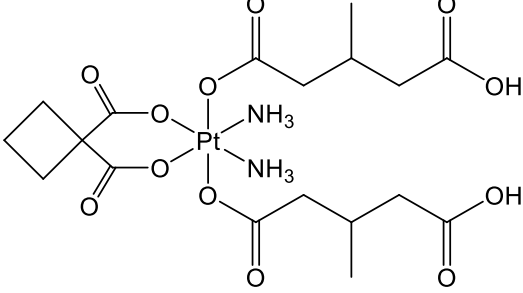
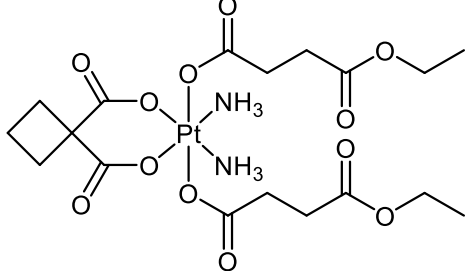
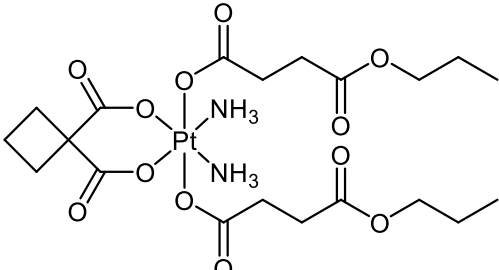
5	(<i>SP-4-2</i>)-(cyclobutane-1,1-dicarboxylato)bis(pyridine)platinum(II)		$-0.06 \pm 0.04 \times 10^{-1}$
6	(<i>SP-4-2</i>)-dichlorido(<i>N,N</i> -dimethylethane-1,2-diamine)platinum(II)		-1.66 ± 0.14
7	(<i>SP-4-3</i>)-dichlorido(<i>N</i> -methylethane-1,2-diamine)platinum(II)		-1.78 ± 0.15
8	(<i>SP-4-3</i>)-dichlorido(<i>N,N</i> -dimethylethane-1,2-diamine)platinum(II)		-1.51 ± 0.12
9	(<i>SP-4-3</i>)-dichlorido(<i>N,N,N'</i> -trimethylethane-1,2-diamine)platinum(II)		-1.22 ± 0.07
10	(<i>SP-4-3</i>)-dichlorido(propane-1,2-diamine)platinum(II)		-2.00 ± 0.18
11	(<i>SP-4-2</i>)-dichlorido(<i>meso</i> -butane-2,3-diamine)platinum(II)		-1.87 ± 0.16
Pt(IV) complexes			

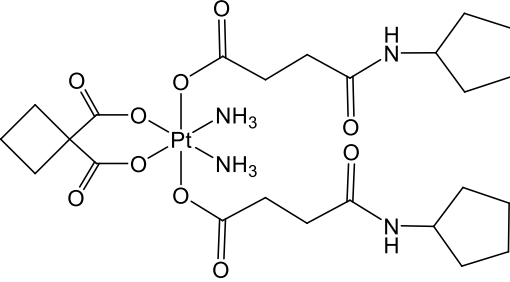
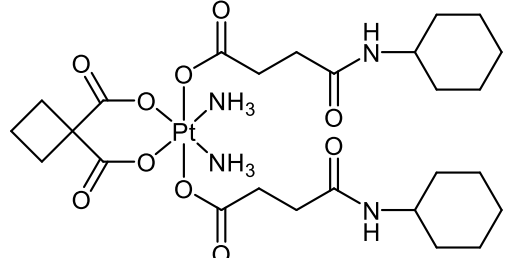
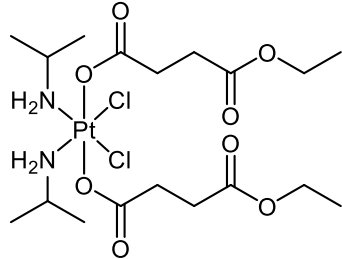
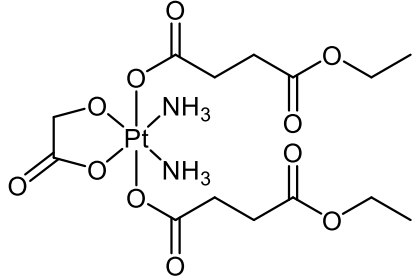
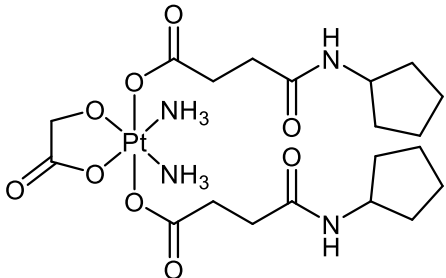
12	(OC-6-33)- dichloridodihydroxidobis(methylamine)platinum(IV)		-2.75 ± 0.20
13	(OC-6-33)- dichloridodihydroxidobis(isopropylamine)platinum(IV)		-1.26 ± 0.28
14	(OC-6-43)- dichloridodihydroxido(<i>N</i> -methylethane-1,2-diamine)platinum(IV)		-2.81 ± 0.03×10 ⁻¹
15	(OC-6-43)- dichloridodihydroxido(<i>N,N,N'</i> -trimethylethane-1,2-diamine)platinum(IV)		-2.60 ± 0.06
16	(OC-6-33)- dichloridodihydroxido(<i>N,N,N',N'</i> -tetramethylethane-1,2-diamine)platinum(IV)		-2.23 ± 0.17

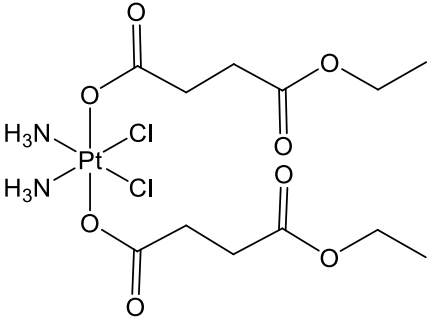
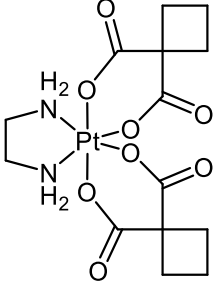
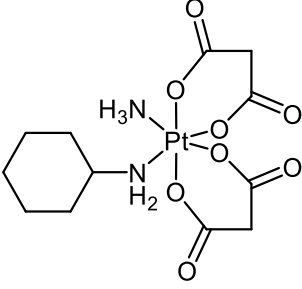
17	(OC-6-33)- dichloridobis(cyclobutanamine)dihydroxidoplatinum(IV)		-0.54 ± 0.08
18	(OC-6-33)- dichloridobis(cyclopentanamine)dihydroxidoplatinum(IV)		-0.23 ± 0.23
19	(OC-6-33)- dichloridobis(cyclohexanamine)dihydroxidoplatinum(IV)		0.57 ± 0.19
20	(OC-6-33)- diamminedihydroxido(propanedioato)platinum(IV)		-2.55 ± 0.02
21	(OC-6-33)- dihydroxidobis(methylamine)(propanedioato)platinum(IV)		-2.53 ± 0.23

22	(OC-6-33)-(ethane-1,2-diamine)dihydroxido(propanedioato) platinum(IV)		-2.50 ± 0.04
23	(OC-6-33)-(ethanedioato)dihydroxido(<i>N,N</i> -dimethylethane-1,2-diamine)platinum(IV)		-2.69 ± 0.01
24	(OC-6-33)-diammine(cyclobutane-1,1-dicarboxylato)dihydroxidoplatinum(IV)		-2.25 ± 0.20
25	(OC-6-33)-(cyclobutane-1,1-dicarboxylato)(cyclohexane-1 <i>R</i> ,2 <i>R</i> -diamine)dihydroxidoplatinum(IV)		-1.72 ± 0.05
26	(OC-6-33)-(2-acetamidomalonato)diamminedihydroxidoplatinum(IV)		-1.89 ± 0.09
27	(OC-6-33)-diammine(2-benzamidomalonato)dihydroxidoplatinum(IV)		-2.18 ± 0.17
28	(OC-6-33)-(2-benzamidomalonato)dihydroxidobis(methylamine)platinum(IV)		-1.36 ± 0.08

29	(OC-6-33)-(2,3-diamino- <i>N</i> -isopropylpropanamide)dichloridodihydroxidoplatinum(IV)		-2.34 ± 0.02
30 ^b	(OC-6-33)-bis(benzoato)dichlorido(cyclohexane-1 <i>R</i> ,2 <i>R</i> -diamine)platinum(IV)		1.11 ± 0.11
31 ^b	(OC-6-33)-dichlorido(cyclohexane-1 <i>R</i> ,2 <i>R</i> -diamine)bis(2-phenylacetato)platinum(IV)		0.92 ± 0.01
32 ^b	(OC-6-33)-dichlorido(cyclohexane-1 <i>R</i> ,2 <i>R</i> -diamine)bis(3-phenylpropanoato)platinum(IV)		0.48 ± 0.04 × 10 ⁻¹
33	(OC-6-32)-ammine(cyclohexylamine)tetrahydroxidoplatinum(IV)		-1.70 ± 0.29

34	(OC-6-33)- bis(acetato)diammine(cyclobutane- 1,1-dicarboxylato) platinum(IV)		-1.42 ± 0.06
35	(OC-6-33)-diamminebis(4- carboxybutanoato)(cyclobutane-1,1- dicarboxylato) platinum(IV)		-1.42 ± 0.03
36	(OC-6-33)-diamminebis(4-carboxy- 3-methylbutanoato) (cyclobutane- 1,1-dicarboxylato) platinum(IV)		-0.48 ± 0.07
37	(OC-6-33)-diammine(cyclobutane- 1,1-dicarboxylato)bis((4-ethoxy)-4- oxobutanoato)platinum(IV)		0.06 ± 0.08
38	(OC-6-33)-diammine(cyclobutane- 1,1-dicarboxylato)bis((4-propyloxy)- 4-oxobutanoato)platinum(IV)		0.95 ± 0.06

39 ^c	(OC-6-33)-diammine(cyclobutane-1,1-dicarboxylato)bis((4-cyclopentylamino)-4-oxobutanoato) platinum(IV)		0.77 ± 0.06
40	(OC-6-33)-diammine(cyclobutane-1,1-dicarboxylato)bis((4-cyclohexylamino)-4-oxobutanoato) platinum(IV)		1.33 ± 0.03
41	(OC-6-33)-dichloridobis((4-ethoxy)-4-oxobutanoato) bis(isopropylamine)platinum(IV)		1.30 ± 0.11
42	(OC-6-42)-diamminebis((4-ethoxy)-4-oxobutanoato)glycolatoplatinum(IV)		-1.24 ± 0.07
43	(OC-6-42)-diamminebis((4-cyclopentylamino)-4-oxobutanoato)glycolatoplatinum(IV)		-0.39 ± 0.12

44	(OC-6-33)-diamminedichlorido bis((4-ethoxy)-4-oxobutanoato) platinum(IV)		-0.36 ±0.02
45	(OC-6-22)-bis(1,1'-cyclobutandicarboxylato)ethane-1,2-diamineplatinum(IV)		-0.66 ± 0.01
46	(OC-6-32)-ammine(cyclohexylamine)bis(malonato)platinum(IV)		-1.00 ±0.02

^a These complexes are obtained as a 9:1 mixture of inseparable isomers. ^b Rather insoluble complexes that required the use of a small amount of co-solvent (DMSO, 5%) in the measurements.

^c logP value is from ref [37] (published after the model development step).

Table 2. RMSE of models developed with different sets of descriptors for the combined set Pt(II) + Pt(IV) with ASNN and MLRA methods.

Descriptors	type	N	ASNN CV		ASNN Bagging		MLRA CV		N'
			RMSE	r ²	RMSE	r ²	RMSE	r ²	
Fragmentor[52]	2D	140	0.53	0.88	0.46	0.91	0.84	0.70	14
GSFrag[61]	2D	171	0.76	0.76	0.67	0.81	1.10	0.50	45
Mera, Mersy[62]	3D ^b	229	0.80	0.73	0.81	0.72	1.08	0.51	10
ChemAxon	3D ^b	79	0.70	0.80	0.56	0.87	1.09	0.51	9
Inductive[63]	3D ^b	38	0.68	0.80	0.58	0.86	1.08	0.51	18
Adriana[64]	2D	108	1.00	0.52	0.90	0.56	1.1	0.3	10
Dragon[65] ^a	2D	803	0.58	0.86	0.54	0.88	2.00	<0	98
CDK[58] ^a	2D	39	0.62	0.84	0.53	0.88	0.72	0.78	18
MolPrint[66]	2D	234	0.77	0.75	0.71	0.79	1.28	0.30	15
EFG[49]	2D	40	0.48	0.90	0.45	0.91	0.78	0.74	9
ECFP4[67]	2D	282	0.67	0.81	0.61	0.84	1.15	0.44	18
Estate[68]	2D	90	0.54	0.88	0.48	0.90	20	<0	54
Consensus: Estate + EFG + Fragmentor	2D		0.45	0.92	0.41	0.93			

^a Calculation of 3D descriptors was not possible. Majority of calculated descriptors were based on 3D structure of molecules while some individual descriptors were also based on 2D representation. ^b3D structures were generated using CORINA [69]. N is number of descriptors after the unsupervised filtering, RMSE is Root Mean Squared Error, r² is coefficient of determination, ASNN is Associative Neural Network, MLRA is Multiple Linear Regression Analysis, CV is Cross-Validation protocol and N' is number of descriptors in MLRA model.

Table 3. RMSE of models analyzed in the current study

Model	Training data set	ref	Training set, CV		Test set predictions	
			Pt(II)	Pt(IV)	Pt(II)	Pt(IV)
VCCLAB-Pt(II)	12.9k organic molecules + 67 Pt(II)	[16]	-	-	0.48± 0.06	2.0 ± 0.2
QC1	28 Pt(IV)	[17]	-	0.35	2.7 ± 0.1	2 ± 0.1
QC2	24 Pt(II)	[14]	0.35	-	1.3 ± 0.1	0.9 ± 0.1
ASNN1	87 Pt(IV)	This work	-	0.36 ± 0.04	0.47 ± 0.08	0.66 ± 0.08
ASNN2	100 Pt(II)		0.50 ± 0.04	-	0.55 ± 0.07	0.81 ± 0.07
ASNN3	87 Pt(IV) + 100 Pt(II)		0.45 ± 0.04	0.36 ± 0.04	0.37 ± 0.07	0.65 ± 0.06
ASNN4	12.9k organic molecules + 87 Pt(IV) + 100 Pt(II)		0.55 ± 0.07	0.41 ± 0.04	0.36 ± 0.06	0.72 ± 0.07

Figure captions:

Synopsis for the Graphical Abstract:

Performance of the consensus model for the training and test set complexes. The logP values of the test set compounds were kept blind during the model development steps.

Figure 1. Histogram of distribution of logP values in the training sets of the ALOGPS 2.1 (A) and of the Pt compounds (B).

Figure 2. Model and predicted MMP plot calculated for it. Left panel shows predicted (y) versus experimental (x) values for consensus logP model, ASNN3, developed with combined set of Pt(II) and Pt(IV) molecules. Green corresponds to Pt(II) complexes, while red corresponds to Pt(IV) complexes. The right panel shows predicted MMP (each dot is one MMP), which are present in the data and were learnt by the model. The points near to the diagonal correspond to MMPs, which were correctly learnt by the model. Both plots are screenshots obtained using OCHEM software and are accessible at the profile of ASNN3 model available at <http://ochem.eu/article/76903>.

Figure 3. Analysis of outlying molecules using predicted MMPs plot. Green dots on the plot indicate MMPs corresponding to a transformation, which changes hydrogen to an ethyl group. The highlighted dot and the shown MMP indicate a pair of molecules one of which ((SP-4-3)-(ethane-1,2-diamine)(2-ethoxy-3-carboxypropanoato)platinum(II), on the left) has a possibly incorrectly measured experimental value. The MMP plot is a screenshot, which was obtained using OCHEM. It is available at the lower right corner of the ASNN3 model available at <http://ochem.eu/article/76903>.

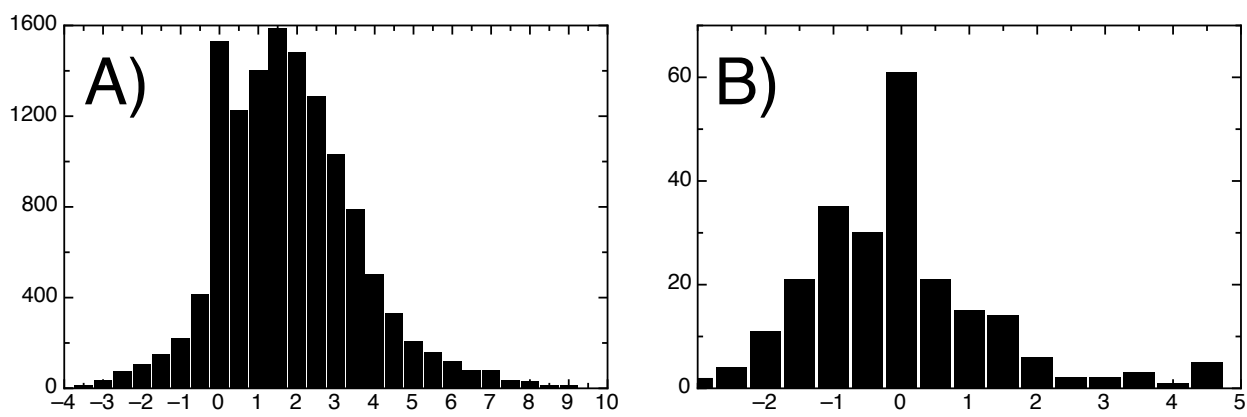


Figure 1.

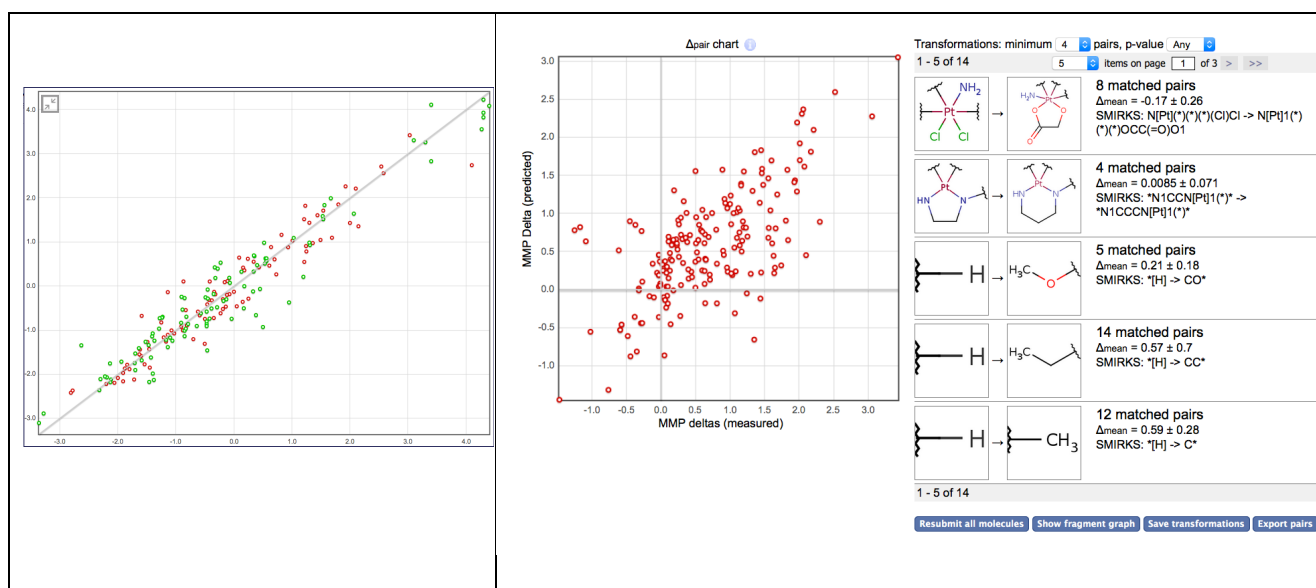


Figure 2.

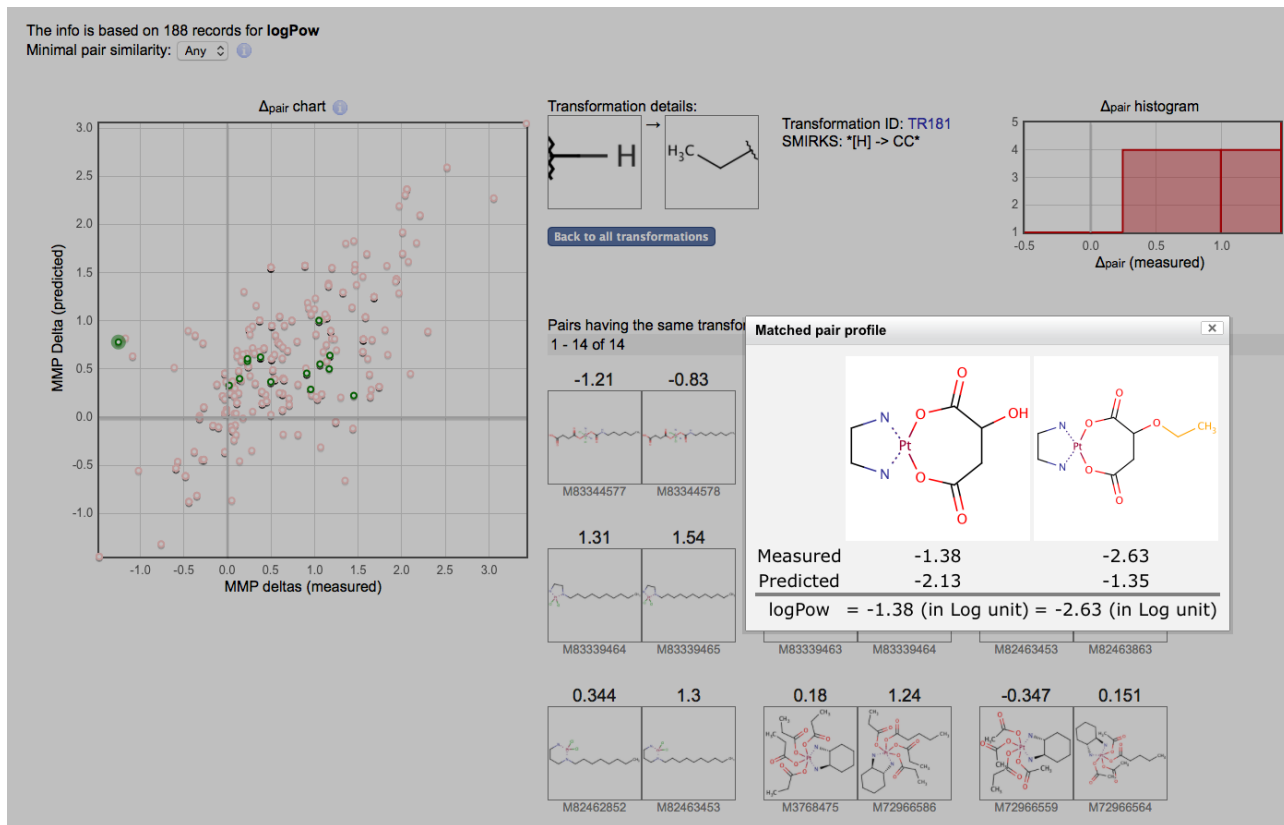


Figure 3.