Journal of Inorganic Biochemistry 269 (2025) 112890

Contents lists available at ScienceDirect

Journal of Inorganic Biochemistry

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

Online OCHEM multi-task model for solubility and lipophilicity prediction of platinum complexes

Nesma Mousa [a], Hristo P. Varbanov [b], Vidya Kaipanchery [c], Elisabetta Gabano [d], Mauro Ravera [e,*], Andrey A. Toropov [f], Larisa Charochkina [g], Filipe Menezes [h], Guillaume Godin [i], Igor V. Tetko [h,j,**]

[a] Freie Universität Berlin, Fachbereich Biologie, Chemie, Pharmazie, Takustr. 3, 14195 Berlin, Germany
[b] Institute of Pharmacy/Pharmaceutical Chemistry, University of Innsbruck, Center for Chemistry and Biomedicine, Innrain 80 - 82/IV, 6020 Innsbruck, Austria
[c] Jerzy Haber Institute of Catalysis and Surface Chemistry, Polish Academy of Sciences, Niezapominajek 8, Krakow 30239, Poland
[d] Dipartimento per lo Sviluppo Sostenibile e la Transizione Ecologica, Università del Piemonte Orientale, Piazza S. Eusebio 5, 13100 Vercelli, Italy
[e] Dipartimento di Scienze e Innovazione Tecnologica, Università del Piemonte Orientale, Viale Teresa Michel 11, 15121 Alessandria, Italy
[f] Istituto Di Ricerche Farmacologiche Mario Negri IRCCS, Via Mario Negri 2, 20156 Milan, Italy
[g] V.P. Kukhar Institute of Bioorganic Chemistry and Petrochemistry, National Academy of Sciences of Ukraine, Academician Kukhar Str. 1, Kyiv 02094, Ukraine
[h] Institute of Structural Biology, Molecular Targets and Therapeutics Center, Helmholtz Munich - Deutsches Forschungszentrum Für Gesundheit Und Umwelt (GmbH), 86764 Neuherberg, Germany
[i] Osmo Labs, PBC, 450 E 29th St, New York, USA
[j] BIGCHEM GmbH, Valerystr. 49, 85716 Unterschleißheim, Germany

ARTICLE INFO

Keywords:
Platinum Pt(II)/Pt(IV) complexes
Water solubility
Lipophilicity
Consensus model
Neural networks
Representation learning

ABSTRACT

Predicting the solubility and lipophilicity of platinum(II, IV) complexes is essential for prioritizing potential anticancer candidates in drug discovery. This study introduces the first publicly available online model for predicting the solubility of platinum complexes, addressing the lack of literature and models in this regard. Using a time-split dataset, we developed a consensus model with a Root Mean Squared Error (RMSE) of 0.62 through 5-cross-validation on a training set of 284 historical compounds (solubility data reported prior to 2017). However, the RMSE increased to 0.86 when applied to a prospective test set of 108 compounds reported after 2017. Further analysis of the high prediction errors revealed that these inaccuracies are primarily attributed to the under-representation of novel chemical scaffolds, particularly Pt(IV) derivatives, in the training sets. For instance, a series of eight phenanthroline-containing compounds, not covered by the training set's chemical space, had an RMSE of 1.3. When the model was redeveloped using a combined dataset, the RMSE of this series significantly decreased to 0.34 under the same validation protocol. Additionally, we developed an interpretable linear model to identify structural features and functional groups that influence the solubility of platinum complexes. We further validated the correlation between solubility and lipophilicity, consistent with the Yalkowsky General Solubility Equation. Building on these insights, we developed a final multitask model that simultaneously predicts solubility and lipophilicity as two endpoints with RMSE = 0.62 and 0.44, respectively. The data and final developed model is available at https://ochem.eu/article/31.

Abbreviations: ADMET, Absorption, Distribution, Metabolism, Excretion, and Toxicity; GIT, gastrointestinal tract; MPNN, message passing neural network; OCHEM, Chemical Modelling environment; QSPR, quantitative structure-property relationship; SDF, Structure-Data Files; CNN, Convolutional Neural Network; SMILES, Simplified Molecular Input Line Entry System; SMARTS, SMILES-Arbitrary-Target Specification; ASNN, Associative Neural Network; RF, Random Forest; OOB, out-of-bag; InChi, International Chemical Identifier; RMSE, Root Mean Squared Error; MP, Melting Point; EFG, Extended Functional Group; HS, high specificity; LS, low specificity.

* Correspondence to: Mauro Ravera, Dipartimento di Scienze e Innovazione Tecnologica, Università del Piemonte Orientale, Viale Teresa Michel 11, 15121 Alessandria, Italy.

** Correspondence to: Igor V. Tetko, Institute of Structural Biology, Molecular Targets and Therapeutics Center, Helmholtz Munich - Deutsches Forschungszentrum für Gesundheit und Umwelt (GmbH), Ingolstädter Landstr. 1, 86764, Neuherberg, Germany.

E-mail addresses: mauro.ravera@uniupo.it (M. Ravera), itetko@vcclab.org (I.V. Tetko).

https://doi.org/10.1016/j.jinorgbio.2025.112890
Received 30 December 2024; Received in revised form 5 March 2025; Accepted 6 March 2025
Available online 10 March 2025

1. Introduction

Aqueous solubility is a crucial property of a drug, impacting its administration route, bioavailability and key pharmacokinetic properties, namely ADMET (Absorption, Distribution, Metabolism, Excretion, and Toxicity). [1,2] Drugs in solid dosage forms administered orally first disintegrate into smaller parts or primary particles, allowing drug molecules to dissolve more readily in the gastrointestinal (GIT) fluids than from an intact tablet. This molecular dissolution of the drug is then followed by its penetration through the intestinal barrier. [3] If the solubility or rate of dissolution is too low in the water-based GIT fluids, the drug molecule will primarily be excreted without entering the bloodstream or reaching its site of action, thereby rendering it ineffective. [1]

In drug discovery, understanding the physicochemical properties of candidate molecules in the early stages can help identify and eliminate molecules that are unlikely to succeed in later stages of development. [1] The importance of aqueous solubility for drug design is considered throughout the drug development pipeline, from assay development, ADMET optimization, and formulation design to the dosage form selection stages. In this respect, *in-silico* estimation of the aqueous solubility for the large number of drug candidates typically considered during lead optimization (before synthesis) is essential for streamlining and accelerating this traditionally lengthy and costly process. It is currently estimated that 40 % of marketed drug products, and 70–90 % of drug candidates in development stages have a low aqueous solubility. This results in low bioavailability, which often necessitates higher doses in the final formulations in order to achieve the desired therapeutic effect. However, this can also increase the risk of toxicity. [3]

In this study, we use quantitative structure-property relationship (QSPR) modelling to predict the intrinsic aqueous solubility of platinum complexes. By quantifying the intrinsic aqueous solubility, we aim to answer the question: "How much of a molecule can dissolve in water under thermodynamic equilibrium at a given temperature, assuming no ionization or other interactions?". In other words, intrinsic aqueous solubility provides insights into the inherent ability of the neutral form of a compound to dissolve in water without external influences like pH adjustments or complexation. For an ionizable molecule, intrinsic solubility is defined as the concentration of the unionized molecule in a saturated aqueous solution at thermodynamic equilibrium at a given temperature. [4] Hence, it represents the lowest solubility of a compound across all pH levels. This definition gives each compound a single, unique value, which is advantageous for computational modelling, and stands in contrast to other types of solubility, e.g., kinetic and equilibrium solubility, which depend on the pH used to perform the measurements and therefore give rise to many different values for ionizable compounds. [5]

Solubility can be expressed using various terminologies, such as molality (*m*), volume fraction (*v/v*), parts of solvent (*ppx*), percentage (% *w*/w, % v/v), molarity (*M*, in mol/L), and others. [3] In modelling, a standard practice is to use the logarithm (log) of solubility (*S*, in mol/L). This standardized unit is important to ensure consistency and avoid discrepancies in reported experimental values, which might otherwise appear in different units, such as mM, mg/mL, or μg/mL. [6] In practice, around 85 % of drugs have a solubility between −1 and − 5 log(mol/L). [7] This solubility range reflects a compromise between the polarity necessary for reasonable aqueous solubility and the hydrophobicity necessary for acceptable membrane transport. [7]

Recent years have seen substantial advancements in modelling methods for predicting aqueous solubility of organic molecules. [1,7,8] Many new approaches in this field have been developed, including artificial neural networks and deep learning, [9,10] which leverage raw molecular structures and have shown promise for the prediction of various molecular properties, including aqueous solubility, especially for large datasets. Comparative analyses between the deep residual network (ResNet) architecture and shallow neural networks, have been conducted by Cui et al. [11] New techniques include graph-based neural networks, [12] message passing neural networks (MPNNs) [13], and MPNN models with self-attention [14] as well as Transformers. [15] These efforts generally rely on datasets, ranging from 100 to about 1 k molecules, [8] with the exception of several recent works, [11,16,17] which have leveraged datasets of around 10,000 molecules, and the Kaggle Kinetic Solubility Prediction Challenge, which utilized a dataset of 100 k compounds. [18] The use of more advanced representations of chemical structures, SELFIES, allowed Yüksel et al. [19] to improve the performance of solubility models in comparison to traditional use of SMILES. These studies demonstrate an increasing interest in developing new methods for solubility prediction for organic molecules. However, to our knowledge, no published solubility prediction models have specifically addressed platinum complexes.

It was not until the 1960s that metal-containing/organometallic drugs reached a milestone in cancer treatment, when Barnett Rosenberg accidentally discovered the anticancer properties of *cis*-diamminedichloridoplatinum(II), known as cisplatin. [20] Cisplatin binds covalently to DNA, leading to DNA damage and subsequent cell death. [21] Despite its efficacy, cisplatin is associated with high systemic toxicity, especially nephrotoxicity. [21] Additionally, it has a water solubility of 1 mg/mL (−2.47 log(mol/L)), poor lipophilicity, high reactivity and severe side effects arising from premature aquation, and non-selective binding to biomolecules. [22,23] This prompted the development of new metallodrugs based on the structure activity relationship (SAR) approach, as postulated by Cleare and Hoeschele [24], using cisplatin as a structural scaffold. This strategy led to the development of carboplatin (*cis*-diammine(1,1-cyclobutanedicarboxylato) platinum(II)), a complex with higher water solubility (17.2 mg/mL or − 1.32 log(mol/L)), higher stability and lower toxicity. A decade later, Oxaliplatin (trans-L-(1*R*,2*R*-diaminocyclohexane)oxalatoplatinum(II)), which has a water solubility of 6.1 mg/mL (−1.81 log(mol/L)), was developed, with the aim of circumventing tumor resistance to cisplatin. [20]

Why does the search for therapeutic platinum complexes continue? Despite the widespread use of existing platinum-based drugs in more than half of all cancer chemotherapy treatments, [25] their clinical application is fraught with difficulties due to some patients' innate or acquired resistance to these drugs, in addition to pharmacokinetic drawbacks. [21,22] Recent efforts in pharmaceutical research are increasingly focused on designing platinum-based anticancer agents with improved pharmacokinetic profiles, aiming to enhance cytotoxicity while reducing side effects. [21]

The great potential of platinum-based drugs as anticancer agents is attributed to the unique coordination chemistry of transition metals, which opens up intriguing possibilities for designing better analogs in terms of pharmacokinetics. The ability to modify ligands and alter coordination geometry, such as transitioning from Pt(II)'s square planar to Pt(IV)'s octahedral configuration, significantly impacts properties such as solubility, lipophilicity, stability, and reduction potential. [20] Moreover, Pt(IV) complexes are kinetically more inert than Pt(II) complexes, reducing their reactivity with off-target biomolecules. They are also prodrugs, which must be reduced under the hypoxic conditions of tumor tissues before becoming active, thereby increasing their selectivity [26]. The Pt(IV) complex satraplatin, for instance, with its additional axial acetato ligands, stands out as the first orally formulated platinum drug evaluated in clinical trials. [27] Finally, the strategic design of ligands in these complexes allows for the targeting of specific tumor sites or incorporation of additional bioactive components, making them promising for innovative and tailored drug design. [20]

Predicting the solubility of Pt complexes based on their chemical structures is a crucial step for designing the required platinum-based anticancer candidates with improved pharmacokinetics. However, the sparse experimental solubility data for Pt complexes and the inherent challenges in their computational modelling hinder the development of robust machine learning-based methods for this purpose. In this study,

we aim to address this existing gap by evaluating the effectiveness of both classic and current state-of-the-art algorithms for predicting solubility of platinum complexes. Building upon our previously developed methods for lipophilicity prediction of platinum complexes, [27–30] as well as existing models for predicting the solubility for organic molecules, [31] we adopt two approaches: (1) modelling solubility of platinum complexes as a single endpoint and (2) developing a multitask model that concurrently predicts solubility and lipophilicity as two endpoints. Using molecular descriptor-based and representation learning-based methods, we provide a comprehensive set of insights that will be highly valuable for addressing salient challenges in early-stage drug design–in particular, for identifying optimal ligands and functional groups that enhance solubility, streamlining the reverse-engineering of new platinum complexes with improved pharmacokinetics profiles.

2. Data and methods

2.1. Data collection

The data used to develop the ML models for this study were curated from about 80 literature sources collected up until November 2024, spanning more than a century of research, as shown in Fig. 1. The data were further augmented by solubility values for $n = 18$ [27,32–37] complexes reported for the first time in this study (Table S1). Detailed procedures for solubility determination are provided in the Supplementary Material (see section S1: Experimental Solubility Determination). Molecular structures were digitally sketched in the On-line Chemical Modelling environment (OCHEM). [38] Solubility values measured in different units (e.g., μg/mL, mg/mL, mg/L) were uploaded as reported in the literature. These values were automatically converted to log(mol/L) values by OCHEM. Storing data with their original units facilitates tracking and makes it easier to detect and correct mistakes in chemical structures.

It should also be noted that dichloridoplatinum(II) complexes, such as cisplatin, can undergo aquation (exchange of chlorido ligands with water), resulting in higher reported water solubilities depending on sample preparation and measurements protocol. This can, at least in part, explain the large variation in the reported solubility data. Also we excluded one compound (see Section 2.2), which the reporting authors described as being unstable in water.

This comprehensive data collection reflects the evolving research trends in platinum complexes. Fig. 1 demonstrates the growing interest in designing new platinum complexes, particularly around the approval year of key cisplatin analogues. Carboplatin was approved for clinical use in 1986 in America [39]; oxaliplatin in 1996 in Europe [39] and in America in 2002 [40]; and nedaplatin in 1995 in Japan. [41] This does indeed emphasize the actively ongoing research in this area, with scientists continuously developing novel complexes or optimizing existing derivatives, aiming at improving pharmacokinetics, enhancing cytotoxicity and reducing side-effects. [42]

2.2. Data cleaning and handling

We initially compiled a dataset of more than 400 Pt complexes, each with reported solubility values. After a preliminary review, we excluded compounds co-containing additional transition metals (e.g., Pd, Cr, Rh, etc.), or alkali metals (e.g., Rb^+, Cs^+) with the exception of K^+ and Na^+, which are counterions for several anionic Pt complexes. We also excluded duplicated entries, one compound that was reported to be unstable in water, and measurements taken at high temperatures (80 °C). This left us with a refined and cleaned dataset of 392 complexes.

The majority of the solubility values in this refined dataset were measured at 25 °C. However, there were a few exceptions, with 20 records measured at 20 °C and 10 at 37 °C. According to Balakin et al., [1] temperature variation below $\Delta T = 30$ °C does not significantly increase the error in solubility prediction, therefore we also kept those compounds. The refined dataset included 155 entries for Pt(IV), 236 for Pt (II), and one entry for a Pt(0) complex. The solubility values varied, ranging from −5.70 to 0.54 on the logS scale (Table S2), reflecting the large variability in the experimental solubility data of platinum complexes. For example, fewer data points fall into the highly soluble or moderately soluble categories (Fig. 2) which may introduce challenges or limitations for the models' ability to learn the complex structure-

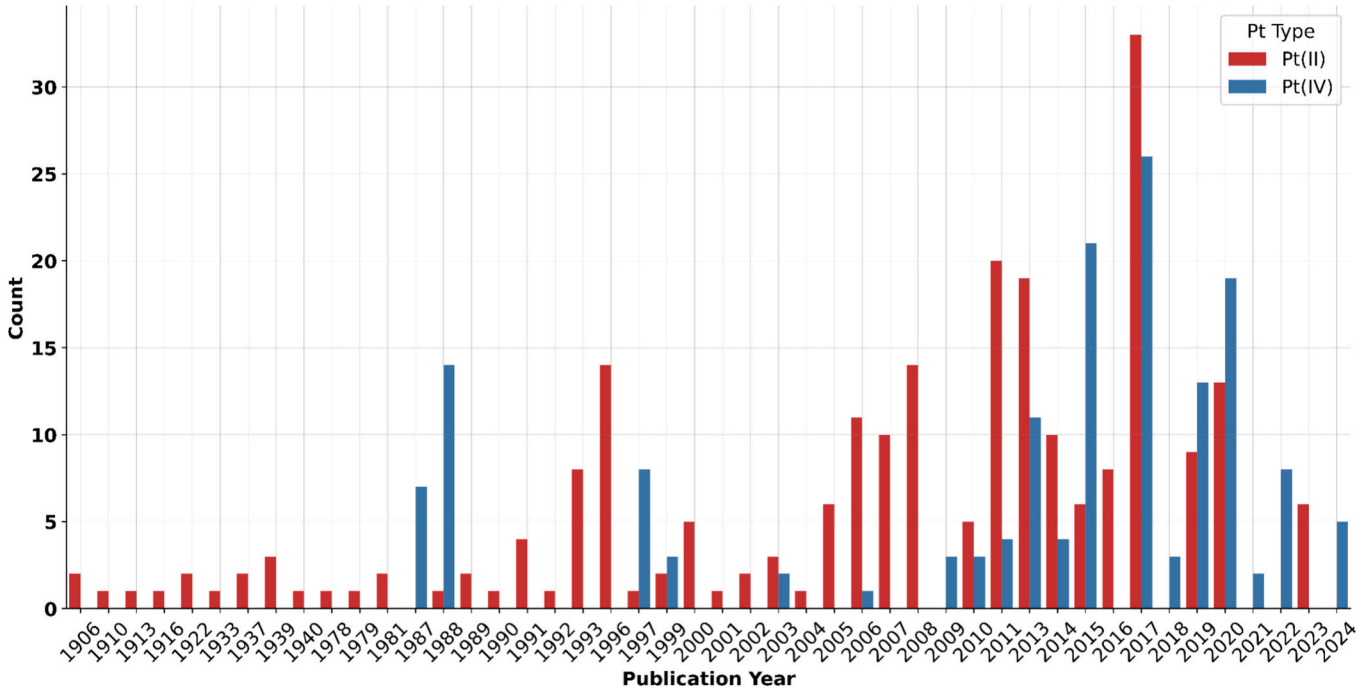


Fig. 1. Time span of collected solubility data. Starting in 2013, the annual data indicates a growing interest in designing new Pt-complexes, particularly Pt (IV) complexes.

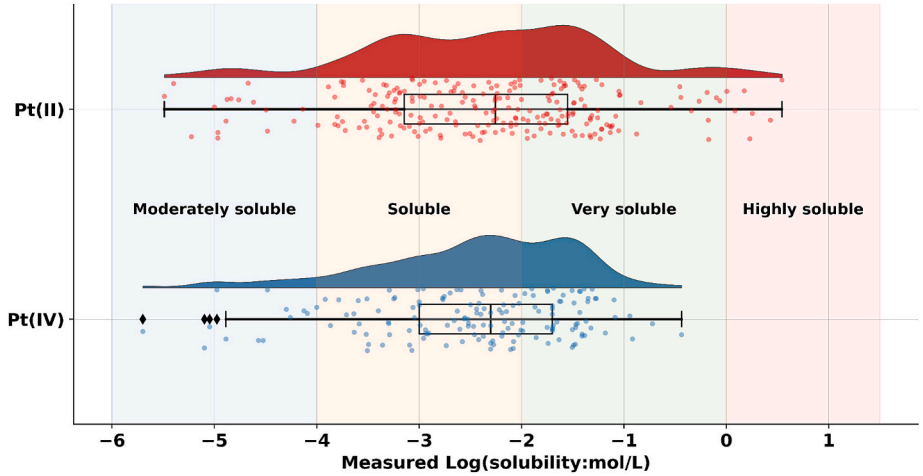


Fig. 2. Distribution of Solubility data for Pt(II) and Pt(IV) complexes. Pt(II) exhibited a wide solubility range with a mean solubility of −2.3 log(mol/L) and a SD of 1.2. In contrast, Pt(IV) complexes had a narrower range with a mean of −2.5 and a SD of 1.0. The data suggest that Pt(II) complexes have higher variability in solubility, while Pt(IV) complexes are slightly skewed towards lower solubility. Solubility classes adapted from SwissADME. [43].

property relationships for these molecules.

2.2.1. Handling of intervals and ranges

Four solubility values were reported as censored values indicating solubility either below a lower threshold or above an upper threshold rather than specific values, e.g., "Water solubility > 25.0 mg/mL" and "Water solubility < 0.1 mg/mL". To handle these records effectively, boundary values were used for the "greater" and "less" ranges (e.g., Water solubility = 25.0 mg/mL and "Water solubility = 0.1 mg/mL", ensuring accurate representation and appropriate handling of the data.

2.2.2. Handling of uncertainty margins

Additionally, 76 values were reported with uncertainty margins, exemplified by entries such as "Water solubility = 0.33 ± 0.02 mM." The central value (e.g., 0.33 mM) was used as the representative solubility value in these cases. The accompanying uncertainty margin (e.g., ± 0.02 mM) indicates the precision of the measurement but was not directly incorporated into the analysis. Instead, these margins were noted for reference, as they provide insights into the reliability and variability of the measurements.

2.2.3. Consistent and reliable data representations

The structures uploaded to the OCHEM platform were internally stored as Structure-Data Files (SDF). During data upload, where required, all structures were represented using de-aromatized Kekulé forms. Coordination bonds were denoted as bond type #8 provided by the JSME editor [44]. For Transformer CNN models, [45] as well as Quantitative Name Property Relationship (QNPR) and ISIDA Fragmentor [46] descriptors this bond type was automatically substituted by a non-specific bond "~". However, the coordination bonds posed challenges for other descriptor calculation programs. To address this issue, an automatic conversion was implemented, replacing them with standard single bonds for calculation of descriptors. While this conversion introduced some trade-offs in chemical specificity, it was consistent across all structures and was appropriately accounted for by the data mining algorithms. For model development, all structures were standardized. This process involved the transformation of molecules following a predefined set of "SMILES-Arbitrary-Target Specification" (SMARTS) templates, primarily aimed at standardizing chemical groups. Subsequently, a neutralization step was implemented. Some Pt-complexes were inherently charged and were represented together with their counterions (see subsection *Standardized Representation of Pt Complexes for Accurate Model Predictions*). Despite the availability of automated preprocessing in OCHEM as outlined in its Preprocessing Manual, [47] we manually checked all compounds due to the manageable size of the dataset.

2.3. Classical machine learning methods

We used the Random Forest regression (RF) [48] and Associative Neural Network [49] (ASNN) as classic descriptor-based methods and compared their performance to the current state-of-the-art Transformer-CNN [45] - the representation learning model in our case.

2.3.1. Overview of descriptor-based machine learning methods

Random Forest [48] is an ensemble learning method. It aggregates predictions of multiple individual trees, each of which is built independently one from another by randomly selecting a subset of descriptors (features). The training data used to develop each tree is sampled with replacement (a "bootstrap" sample) from the original training set. The samples that were not used in the development of the respective tree can be used to estimate performance of the model (so called "out-of-bag" (OOB) error).

Associate Neural Networks [49] (ASNN) is an ensemble method that trains individual neural networks models and corrects their bias using k nearest neighbors. This approach was inspired by studies of thalamo-cortical organization of the brain [50] and was shown to improve accuracy of the neural network ensemble.

For the RF and ASNN methods, molecules were used with seven sets of molecular descriptors and fingerprints to generate features as input for the models. We employed a combination of two-dimensional structural fingerprints, and 2D topological descriptors to capture and quantify global and local chemical information of the molecules as described in the next subsection. We also explored whether 3D descriptors could provide better results.

2.3.2. Chemical descriptors

Below is a brief overview of the analyzed descriptors. Detailed descriptions of each set of descriptors can be found on the OCHEM platform. [38] Moreover, the utilized descriptor packages, with the exception of alvaDesc, are also available as part of open source GitHub version of openOchem https://github.com/openochem/openochem.

***E*-state** [51] involves calculating atom-type E-state indices and molecular bond E-state indices, which are 2D electro-topological descriptors derived from a compound's molecular graph. These indices represent the electronic (charge) state of each bonded atom within a molecule, capturing information about its topological (shape) nature. In this study, we employed the counts and indices of atoms and bonds.

MOLD2 [52] is a software tool developed by the FDA that calculates 777 descriptors ranging from 1D to 2D.

MODRED2 [53] is a software application based on the invariance of adjacency relationships. It is capable of generating over 1800 descriptors of both 2D and 3D characteristics at a high speed. Mordred consists of two main classes: "Descriptor" and "Calculator." They automatically preprocess molecules (add or remove hydrogen atoms depending on calculated descriptors, perform Kekulization, and detect molecular aromaticity).

QNPR [54] Quantitative Name Property Relationship (QNPR) are 1D descriptors directly based on the SMILES string. They are calculated by splitting the respective string into fixed-length substrings. In this work, substrings with a length one to three characters were used.

Fragmentor ISIDA fragments are 2D descriptors. [46] Molecules are divided into substructural molecular fragments (SMFs), which have lengths 2–4 in this work. The frequencies of occurrence of SMFs in a molecule were used as descriptors.

Various **RDKIT and ECFP4** [55] descriptors [56] as well as circular topological fingerprints [55] generated by the RDKit chemoinformatics library. ECFP4 represents the variant with a radius of two bonds. These fingerprints are designed to capture the chemical environment of each atom in a molecule and are particularly useful for structure-property modelling.

StructuralAlerts descriptors (also known as Extended Functional Groups, or EFGs) [57], featured by the ToxAlerts tool, [58] are uniquely identified by SMARTS templates encoding the presence or absence of 583 heterocyclic compound classes and periodic table groups (e.g., hydrocarbons, ketones, dioxanes, etc.) in molecules.

AlvaDesc [59] **(v. 2.0.16)** which has 33 blocks that calculates 5666 descriptors such as constitutional, topological, and pharmacophore descriptors. It includes ETA and Atom-type *E*-state indices, functional groups, and fragment counts. Additionally, AlvaDesc implements an extensive set of 3-dimensional descriptors, including 3D-autocorrelation, Weighted Holistic Invariant Molecular Descriptors (WHIM), and GETAWAY. AlvaDesc also provides the calculation of several model-based physicochemical properties, such as the aqueous solubility model (ESOL) and includes geometrical and chirality descriptors.

For descriptor packages requiring 3D information, 3D molecule structures were generated and optimized using ULYSSES. [60] The ULYSSES was selected since this quantum-chemistry package supports optimization of metal-complexes.

2.3.3. Unsupervised filtering of descriptors

A series of unsupervised filters were applied when selecting features to retain only informative descriptors. Descriptors with fewer than 2 unique values were eliminated for lack of discriminatory power and those with a variance below 0.01 were also excluded to ensure meaningful variability. Pairwise correlations among descriptors were assessed using Pearson's correlation coefficient, and those with a correlation coefficient exceeding 0.95 were grouped together to minimize redundancy.

2.4. Representation learning

Transformer-CNN [45] used a learned representation based on augmented SMILES notation. The model deployed in this study leveraged pre-training on SMILES representations of 1.7 M organic compounds from ChEMBL database used in the original study [45] enriched with approximately 600 platinum complexes, curated from the literature as part of our ongoing research into logS and logP predictions. For this step, which aimed to train neural networks to learn SMILES canonization, no experimental data were used, only SMILES. The curated dataset was augmented by generating 10 non-canonical SMILES representations for each compound as explained elsewhere. [45] This allowed the model to learn the numerical representation of the organic as well as platinum complexes, which is the key to transferring the learning to our smaller dataset. Additionally, stereochemical features such as (/,\, @, @@) were incorporated within the vocabulary of 66 symbols used by the model, enriching the representation of molecular structures with stereochemical information.

2.5. Consensus model

The consensus model approach was applied to leverage the complementary strengths of different modelling methods. Predictions of individual base models were averaged as shown in Eq. (1).

$$\mathbf{y} = \mathrm{sum}(\mathbf{y}_i)/\mathrm{n} \tag{1}$$

In this equation, the sum is done over all $i = 1, \ldots, n$ is selected models, and $\mathbf{y}_i$ is the vector of prediction of the selected model. Consensus modelling enhances the robustness and accuracy of individual models. [61]

2.6. Optimization of hyperparameters of methods

The models developed in this study use a pre-set of hyperparameters, that were optimized across several datasets of different properties. Namely 512 trees and 1/3 numbers of all features were used to develop each model for the RF, and 64 single hidden layer models with 3 hidden neurons for the ASNN. For the ASNN the settings were the same as used in logP models for Pt compounds. [27,28] These hyperparameters were also used by us consistently across multiple previous studies, including those predicting the lipophilicity of Pt complexes. [27,28] Transformer CNN was used to develop models using the same hyperparameters as described elsewhere. [45] The model was trained for a maximum of 25 epochs. Optimizing hyperparameters specifically for the relatively small solubility dataset, such as the one analyzed in this study, could easily result in overfitting. [17] Therefore, to maintain model generalizability and prevent overfitting, we did not perform hyperparameter optimization in this study.

2.7. Model evaluation

The performance of each representation and model approach was evaluated using the Root Mean Squared Error (RMSE) as the error metric. This statistical coefficient measures the average differences between the values predicted by a model, y_p^i, and the measured values, y_e^i, as shown in Eq. (2).

$$\mathrm{RMSE} = \mathrm{sqrt}\left(\mathrm{sum}\left(\mathrm{y}^{\mathrm{i}}{}_{\mathrm{p}} - \mathrm{y}^{\mathrm{i}}{}_{\mathrm{e}}\right)^2 \Big/ \mathrm{N}\right) \tag{2}$$

where sum is over all $i = 1, \ldots, N$ samples in the dataset.

2.8. Model validation

Since the main goal of this study was to develop models capable of predicting the solubility of new platinum complexes, we estimated the predictive performance of models using two rigorous validation protocols: a time-based split and five-fold cross-validation (5CV) approach.

2.8.1. Time-based Split for prospective validation of models

To simulate real-world conditions, where historical data guide model training and newer data serve as a prospective test, we split the dataset based on publication year of the source literature. Specifically, 284 complexes published before 2018 were designated as historical data for training, while 90 complexes published in 2018 or later formed the test set. The test set was further augmented with 18 complexes, whose solubility was measured for this study (Table S1). The test set was not used for any aspect of model selection or tuning, reflecting practical applications and ensuring unbiased assessment of predictive performance.

2.8.2. Five-fold cross-validation (5CV)

The 5CV was done to estimate prediction ability of models developed with respective dataset. The dataset was randomly split on five folds using 14 first characters of InChi hash codes, which account only for the connectivity information. This ensured that any racemates or isomers of the same compound were confined to a single fold – whether for the training fold or the validation fold – thereby preventing accidental overlap/leakage between training and validation subsets (i.e., the compound was never in training and validation sets simultaneously). Models were trained on four folds and validated on the remaining fifth fold, which was not used at any point for model selection or tuning. This process was repeated five times so that each fold served once as the validation subset. The final model was then developed with the entire dataset following exactly the same procedure as per individual model during the cross-validation process.

3. Results and discussion

3.1. Chemical space analysis

The post-2017 test set had marginally lower mean solubility (−2.44 log(mol/L) compared to the earlier (historical) training set (−2.36 log (mol/L), as shown in Table S2. While this might seem counterintuitive if one expects newer compounds to be more soluble, the discrepancy could reflect efforts targeting other properties rather than purely optimizing for solubility. A notable example is satraplatin, which was developed with a relatively low aqueous solubility (0.4 mg/mL or − 3.1 log(mol/L)), however it has higher lipophilicity and in vivo stability and less reactivity, leading the new generation of oral platinum agents. Although efforts to enhance pharmacokinetics continue through modifications of existing ligands and the introduction of more hydrophilic ligands, these strategies may not necessarily or uniformly boost solubility across all newly designed Pt(II) and Pt(IV) complexes.

Upon analyzing the distribution of Pt(II) and Pt(IV) within the training time split set, we found that it included only 74 Pt(IV) complexes (26 %), whereas the test set was predominantly composed of Pt (II) complexes (70 %). This imbalance is critical, especially considering our previous analysis of lipophilicity prediction, which revealed that models developed with Pt(II) logP values provided low accuracy for Pt (IV) complexes. [27] Consequently, the underrepresentation of Pt(IV) complexes in the training time split is anticipated to decrease the accuracy of the model for our time-split analysis for these Pt(IV) types of complexes in the test set.

The chemical space distribution of the test set in respect to the training set is shown in Fig. 3. The figure highlights examples that offer insights into the diversity and underrepresentation of certain chemical groups within the test and training sets. As per literature, this includes: the novel design of a series of eight Pt(IV) complexes incorporating halogenated phenylacetic acid derivatives [62] derived from Pt(II) bearing heterocyclic ligand phenanthroline (FPA-phen). This group of phen-containing compounds is unique to the test set; a series of seven satraplatin analogues with the replacement of satraplatin's equatorial

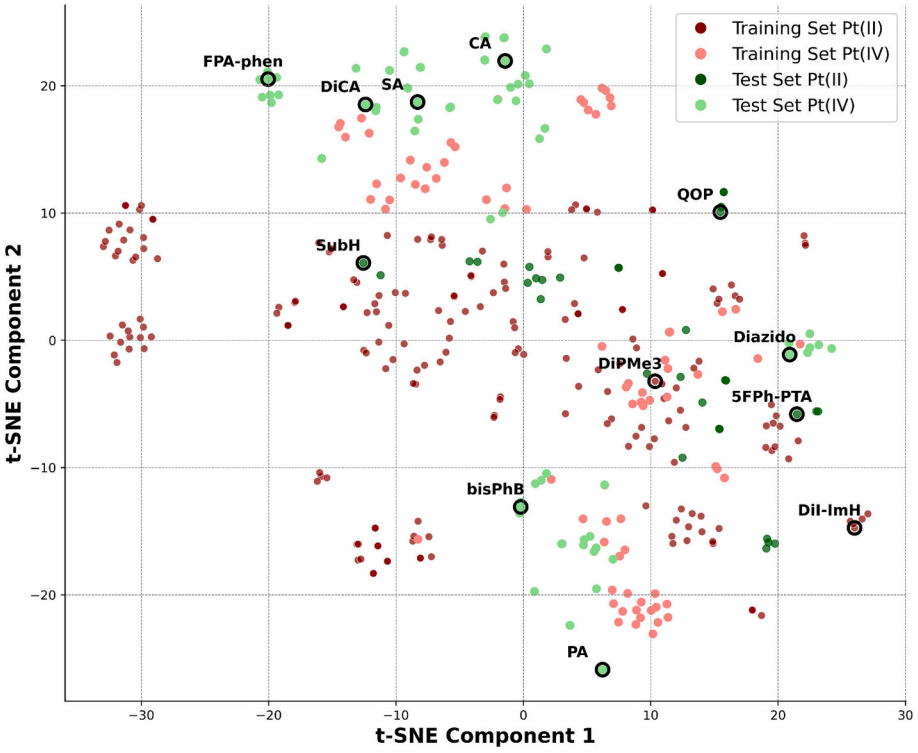


Fig. 3. t-SNE visualization illustrating the 2D distribution of the post-2017 test set with respect to the training set chemical space. The distribution of Pt types (II, IV) is also demonstrated within each data set, with Pt(IV) colored with lighter shades. Molecules were encoded using ECFP4 fingerprints: radius 2 and 2048 Bits. Distinct clusters of both datasets indicate the innovative strategies recently adopted to design novel compounds. Each circled and labeled compound is an example of its respective derivative cluster, with labels indicating the names of ligands coordinated to the platinum (Pt) or characterizing these compounds. The figure showcases the diversity of chemical properties and highlights the specific areas within the chemical space where the predictive model based on the training set may have low accuracy since the test set will be out of the applicability domain of the model. *Here, "FPA-phen" represents phenanthroline-containing derivatives with halogenated (fluorine) phenyl acetate.* [62] *"DiCA" a series of asymmetric Pt(IV) dicarboxylato derivatives.* [63] *"SA" is oxaliplatin-based Pt(IV) complex bearing succinic acid.* [64] *"CA" a series of unsymmetric Pt(IV) complexes bearing carboxylate or carbamate ligands.* [65] *"SubH" is the suberoyl-bis-hydroxamic acid ligand. "QOP" a series of cationic Pt(II) complexes bearing quinolinate and phosphine ligands.* [66] *"diazido" a series of trans-di-(N-heterocyclic)imine dihydroxido diazido Pt(IV) complexes.* [67] *"5FPh-PTA" a pentafluorophenyl Pt(II) complex of PTA (1,3,5-triaza-7-phosphaadamantane).* [68] *"bisPhB" denotes bis phenylbutyric acid.* [26] *"PA" two Pt(IV) conjugates containing one or two molecules of perillic acid.* [69]. *From the training set, "DiI-ImH" diiodido derivative of a series of Pt(II) containing bis (imidazole).* [70] *"DiPMe3" phosphine Pt(II) derivatives.* [71]

chlorides with acetates and two axial phenylbutyric acid groups (bisPhB); the development of Pt(IV) complexes with axial carbamate ligands;

3.2. Effect of molecular representations and descriptors on model performance

We assessed model performances using a wide range of structure representations, and employing RMSE as the evaluation metric. Preliminary analyses revealed that RF provided the highest accuracy for the training set using 5-fold CV compared to other descriptor-based methods in predicting solubility. Therefore, the RF method was selected for modelling. The performance results for each RF model are given in Table 1 for both time-split sets as well as combined sets.

For the time-split dataset the best performance (RMSE = 0.64) was achieved by the RF using alvaDesc [59] followed by Estate, [51] and RDKIT (2D) descriptors. Two other sets of descriptors, namely ISIDA Fragments and MOLD2 calculated RMSEs which were close to the top-performing descriptors. Transformer CNN performance was similar to the best RF models. Furthermore, applying the consensus approach by combining the Transformer CNN with seven 2D descriptor-based models resulted in the highest prediction accuracy across all datasets as reported in Table 1.

The comparison between 2D and 3D descriptors showed that both contributed models with similar RMSE. This observation is intriguing given the complex nature of solubility as a property influenced by both molecular packing in crystal structure and interaction with the solvent. However, the same result was also observed and explained with respect to solubility of general organic compounds by Balakin et al., [1] who pointed to difficulties in identifying the conformational minimum of compounds in crystal structures which could contribute to this problem. The generation of atomic coordinates by the ULYSSES package [60] also failed for two structures – large complexes for which no suitable 3D structures were generated. Therefore, we decided to not use 3D descriptors in this study.

The narrow variability in 5CV accuracy across 2D descriptors emphasizes the robustness of model performances notwithstanding the used molecular representations. Across all 2D descriptors, the RMSE was higher in the prospective time-split test set which inherently introduced new chemical classes which were not found in the earlier complexes from the training set, or which were underrepresented there (Fig. 3). This expansion of the chemical space, particularly if the training set is small, can significantly impact model performance. Due to the small sizes of training and test sets, the confidence intervals for all sets are rather wide. The lowest RMSE across all descriptors was calculated using alvaDesc and Estate indices.

Table 1
RMSE values across analyzed molecular representations and models developed using RF method.

Descriptor set	Time Split Training 5CV (n = 284)	Time Split Test Set (n = 108)	Combined Set 5CV (n = 392)
AlvaDesc (2D)*	0.64 ± 0.04	0.90 ± 0.06	0.66 ± 0.03
Estate*	0.65 ± 0.04	0.91 ± 0.07	0.66 ± 0.03
RDKIT (2D)*	0.65 ± 0.04	0.81 ± 0.06	0.65 ± 0.03
MOLD2*	0.66 ± 0.04	0.86 ± 0.06	0.66 ± 0.03
ISIDA Fragmentor (length: 2–4)*	0.66 ± 0.04	0.85 ± 0.06	0.65 ± 0.03
MORDRED (2D)*	0.69 ± 0.04	0.87 ± 0.06	0.65 ± 0.03
QNPR (length: 1–3*)	0.7 ± 0.05	0.97 ± 0.07	0.7 ± 0.03
RDKIT (ECFP4)	0.72 ± 0.04	0.91 ± 0.06	0.74 ± 0.03
EFGs	0.79 ± 0.04	0.84 ± 0.06	0.74 ± 0.03
MORDRED (3D)	0.67 ± 0.04	0.88 ± 0.06	0.67 ± 0.04
AlvaDesc (3D)	0.66 ± 0.04	0.93 ± 0.06	0.71 ± 0.04
RDKIT (3D)	0.67 ± 0.04	0.87 ± 0.05	0.72 ± 0.04
PyDescriptor (3D)	0.71 ± 0.04	0.97 ± 0.07	0.75 ± 0.03
Learned Representation (LR)			
Transformer CNN*	0.65 ± 0.04	0.92 ± 0.06	0.63 ± 0.03
Consensus model (eight models)			
1 (LR) + 7 (2D) base models	0.62 ± 0.04	0.86 ± 0.06	0.62 ± 0.03

* indicates base models used to build consensus models.

3.3. Analysis of consensus models

3.3.1. Time split sets

The consensus model was built as an average of eight models with the lowest RMSE, developed with RF and Transformer-CNN. Its performance on the time split training set, evaluated through 5CV, yielded an RMSE of 0.64 while an RMSE of 0.86 was obtained for the test set of compounds measured after 2017 (Table 1, Fig. 4a). It should be mentioned that the model based on RDKIT fragments provided a lower RMSE = 0.80 for this test set. However, we could not know in advance which set of descriptors would contribute the lowest RMSE for unseen data. Also, selection of models based on their test set performances increases the risk of chance correlation, in comparison to selection based on the results of 5CV on the training set. [72] Indeed, confidence intervals for the test set performances are larger than those for training test set, which could introduce bias into the selection and result in overfitting. [72]

This discrepancy between the training and test sets highlights the challenges associated with model generalization to unseen data. In general, the increase in RMSE for the test set indicates two possibilities: overfitting, or presence of molecular structures in the test set that fall outside the model's applicability domain as defined by the training data. The former issue is excluded since performance for 5-fold CV was calculated using exactly the same protocol as for time-split set. This issue is exacerbated especially if the dataset is small.

As previously noted, Pt(IV) complexes were underrepresented in the time-split training set (26 %), but overrepresented in the test set (70 %). The model predicted Pt(II) and Pt(IV) complexes with RMSE of 0.70 ± 0.08 and 0.91 ± 0.07 respectively. Thus, the model had difficulty with generalizing solubility prediction for Pt(IV) complexes, which were scarce in the historical training set but prevalent in the newer compounds contained in the test set. This was visually and statistically evident in our case, as shown by the distinct clusters of compounds from the test set depicted in Fig. 3, indicating their higher structural diversity as compared to compounds from the training data. For example, RMSE = 1.3 was for 8 phenanthroline-contacting compounds from study of Aputen et al. [62] (FPA-phen group) as well as for 12 compounds from the DiCA group. [63] These sets of compounds were on the border of the t-SNE plot as shown in Fig. 3 and thus were more structurally diverse than the compounds in the training set. However, once the model was redeveloped using all compounds, RMSEs of 0.34 ± 0.08 and 0.7 ± 0.2 were obtained for the FPA-phen and DiCA groups, respectively. The higher accuracy for FPA-phen can be explained by the lower diversity of molecules within this group as compared to DiCA group, as evidenced by the sizes of the clusters for both groups in Fig. 3.

This comparison underscores the benefits of using a larger and more diverse dataset for training in enhancing model reliability and accuracy. Extension of the training set with compounds from the test set allowed the model to account for structural diversity within this set.

3.3.2. Combined set

When evaluated on the combined dataset using 5CV, the RMSE of the ensemble model did not change 0.62 ± 0.03. This result again indicates that the time split test set had around the same experimental accuracy as the other data, and their low accuracy was mainly due to differences in chemical diversity of the test set compounds. The broader range of solubility values in the combined dataset enabled the model to better capture the complexity of solubility prediction across a wider array of molecular structures. This improvement highlights the potential benefits

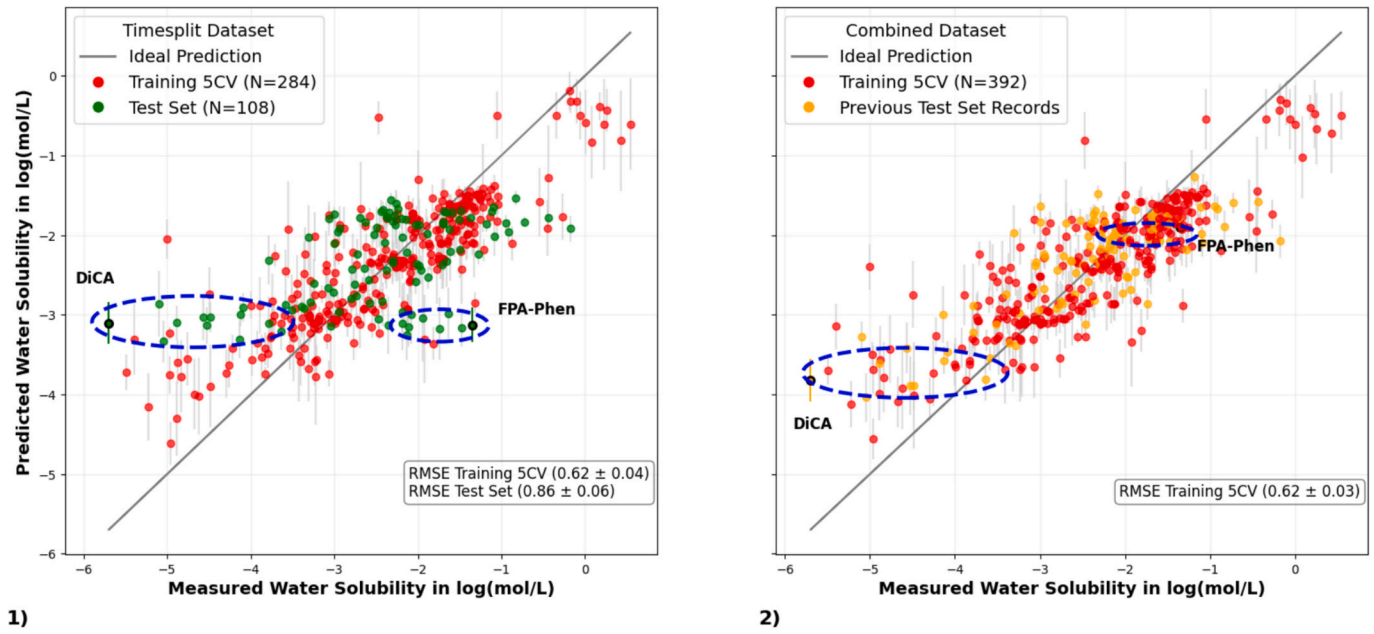


Fig. 4. Performances of Consensus Models. Measured values (x-axis) are plotted vs predicted for consensus models built using 1) time split and 2) combined datasets. Standard deviations for individual predictions are shown as error bars. Eight and twelve molecules from the test set (FPA-phen and DiCa, see Fig. 3) are encircled. For both these sets the model calculated RMSE = 1.3 in the time split dataset and RMSE = 0.34 (FPA-phen) and 0.7 (DiCA) in the combined dataset.

of incorporating diverse training data to enhance model performance and generalizability in solubility prediction tasks.

3.4. Analysis of outlying compounds

The outliers in the test set could be mainly attributed to the limited chemical space coverage in the training set, as exemplified by the "FPA-phen" and "DiCA" series discussed in the previous subsection. Once the model was retrained on the combined dataset to include these respective series of compounds, they were no longer identified as outliers.

Several compounds were also identified as significant absolute outliers within the training sets. The largest outlier in both sets was *cis*-dichloridobis(trimethylphosphine)platinum(II) (Fig. 5). The compound exhibited a solubility of −5 log(mol/L), yet was predicted with higher solubility value of −2.05 log(mol/L) when using the time-split training set, and − 2.4 log(mol/L) when using the full training set, respectively. This compound was rather unique within the training set, and its nearest neighbor compounds had low Tanimoto similarity (<0.5). Additionally, these nearest neighbors were much more soluble, which might contribute to the prediction discrepancy.

The figure highlights *cis*-dichloridobis(trimethylphosphine)platinum (II), the largest outlier in our dataset, alongside its nearest neighbors from the training set. This compound has a substantial lower solubility compared to its most similar neighbors, which do not contain any phosphorus atoms. Note that the model is unable to distinguish geometric (cis-trans) isomerism; therefore, representing compounds as either cis or trans isomers has no impact.

The second largest outlying molecule in both training sets was a di-iodido bis(imidazole) derivative synthesized and measured by the co-authors. [70] The compound is depicted on Fig. 6 alongside its two analogues. In general, di-chlorido and di-hydroxido platinum complexes exhibit higher solubility compared to di-iodido complexes due to a combination of factors including higher electronegativity, reduced steric hindrance, lower polarizability leading to decreasing hydrophobicity, and enhanced hydrogen bonding capabilities. In addition, iodido ligands do not undergo aquation as readily as chlorido ligands. Chlorido ligands are exchangeable with water molecules through aquation within the biological environment of human cells. Although the aquation reaction is not thermodynamically favorable under standard conditions, cellular environment drives the formation of more soluble aqua complexes as demonstrated by the classical example of cisplatin.

The solubility dataset contained 132 compounds with Pt—Cl bonds, 12 with Pt—Br, 8 with Pt—F bonds, and only 4 compounds with Pt—I bonds. Complexes containing Pt—Cl, Pt—Br, or Pt—F bonds had RMSE of 0.64, 0.5, and 0.3 in the combined set, respectively, indicating good predictive performance. Additionally, the dataset contained six compounds with iodine bonded to benzene rings or CH_2 groups instead of directly coordinated to the Pt center (i.e., not Pt—I bond), for which the model did not have any difficulty and predicted them with RMSE = 0.6. However RMSE = 1.3 was obtained for complexes with Pt—I bonds. This suggested that a limited number of only 4 compounds was insufficient to learn the influence of Pt—I bond on the solubility of compounds. Thus, more solubility measurements for compounds featuring Pt—I bonds are required to accurately learn the influence of this bond type on the solubility of platinum complexes.

3.5. Analysis of overrepresented groups in low/high soluble compounds

To identify structural features that are overrepresented in lower-solubility compounds, we divided the dataset into lower and higher soluble groups using the median logS value −2.73 log(mol/L) as the threshold. We used the SetCompare tool [73] in OCHEM, which utilizes a hypergeometric distribution with Bonferroni correction, to determine the prevalence of certain Extended Functional Groups (EFGs) [57] within each class/set. Our analysis showed that the most significant EFGs associated with lower solubility were halogens, aromatic compounds, and arenes (Fig. 7). These groups also featured in the linear equation model (see the next subsection, Table 2) with negative coefficients, indicating their contribution to reduced solubility and thereby confirming the statistical findings of the SetCompare tool.

Arenes and aromatic compounds are characterized by their stable, conjugated ring systems that tend to stack. This contributes to their hydrophobic nature, limited hydrogen bonding capability, and structural rigidity. These features disrupt the hydrogen bond network of water, making the solvation process less energetically favorable. In contrast, compounds containing any cations, saturated six-membered

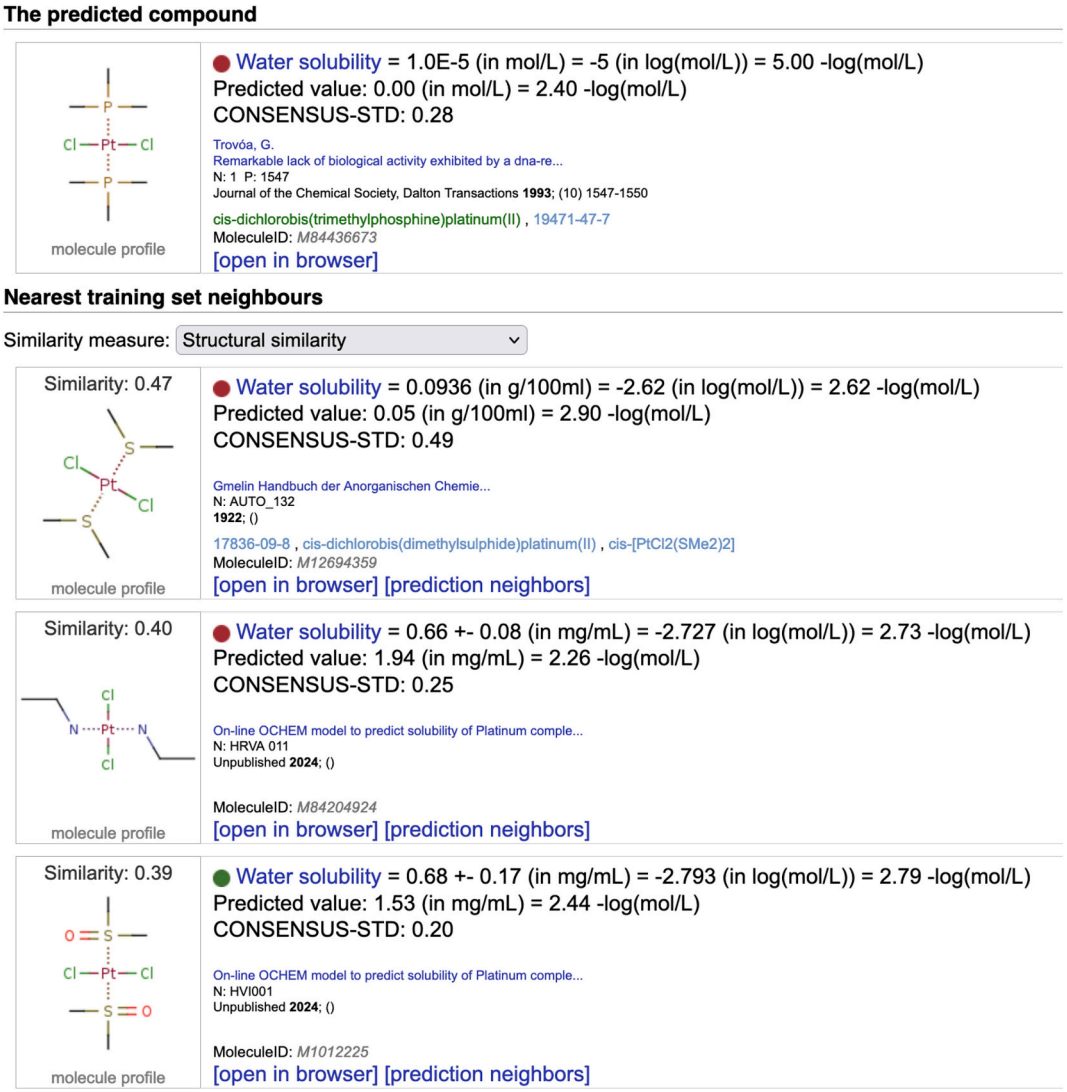


Fig. 5. *Cis*-dichloridobis(trimethylphosphine)platinum(II) compound ("DiPMe3", as shown in Fig. 3) and its nearest neighbors in the training set.

heterocycles with three heteroatoms (Fig. S1), and dialkyl-ethers were associated with higher solubility.

3.6. Interpretable model based on logP and melting point

The Yalkowsky General Solubility Equation (GSE) [74] links compound solubility to their lipophilicity and melting point.

$$\log S = 0.5–0.01(MP - 25) - \log P \quad (3)$$

where logS is the predicted solubility in log(mol/L), MP is the melting point in Celsius (°C) and logP is the lipophilicity of compounds in log units.

For this study, logP values were predicted using our previously developed model [27] (which was trained on a dataset of $n = 233$ compounds). Melting points were predicted using a Transformer CNN model, initially trained on a 275 k dataset [54] and extended to include melting point data for a subset of 55 Pt complexes.

The Transformer CNN method used in this study calculated RMSE = 33 °C for the training set, which was within the same accuracy as the consensus model used in the original study. [54] However, for the subset of Pt-complexes, a much higher RMSE = 49 °C was obtained. It is known that Pt complexes can decompose before melting, which could contribute to the low accuracy of the model for these chemicals.

An attempt to directly predict logS applying the Yalkowsky equation with the predicted MP and logP values resulted in an RMSE = 1.6. However, we redeveloped the model using linear regression with the same properties. The linear regression model had an RMSE = 0.98. Notably, the melting point did not contribute to the regression model and was excluded from the final equation. This potentially was due to the low accuracy of the melting point model. At the same time our analysis validated the significant correlation between logP and logS, highlighting the predominant influence of logP over melting point in determining the solubility of platinum complexes.

3.7. Development of interpretable model based on lipophilicity and EFGs

In addition to being instrumental for the identification of molecular features associated with low- or high-soluble compounds, the EFGs contributed to one of the best individual models for the test set in the time-split validation study for RF (Table 1). The advantage of using EFGs is that models based on such descriptors are inherently interpretable due to the clear chemical rationale behind the descriptors: presence or absence of specific functional groups.

For the time-split validation a linear model based on EFGs calculated an RMSE of 1.05 ± 0.05 and 1.2 ± 0.1 for training and test sets, respectively, and had an overall RMSE of 0.97 ± 0.04 for the combined

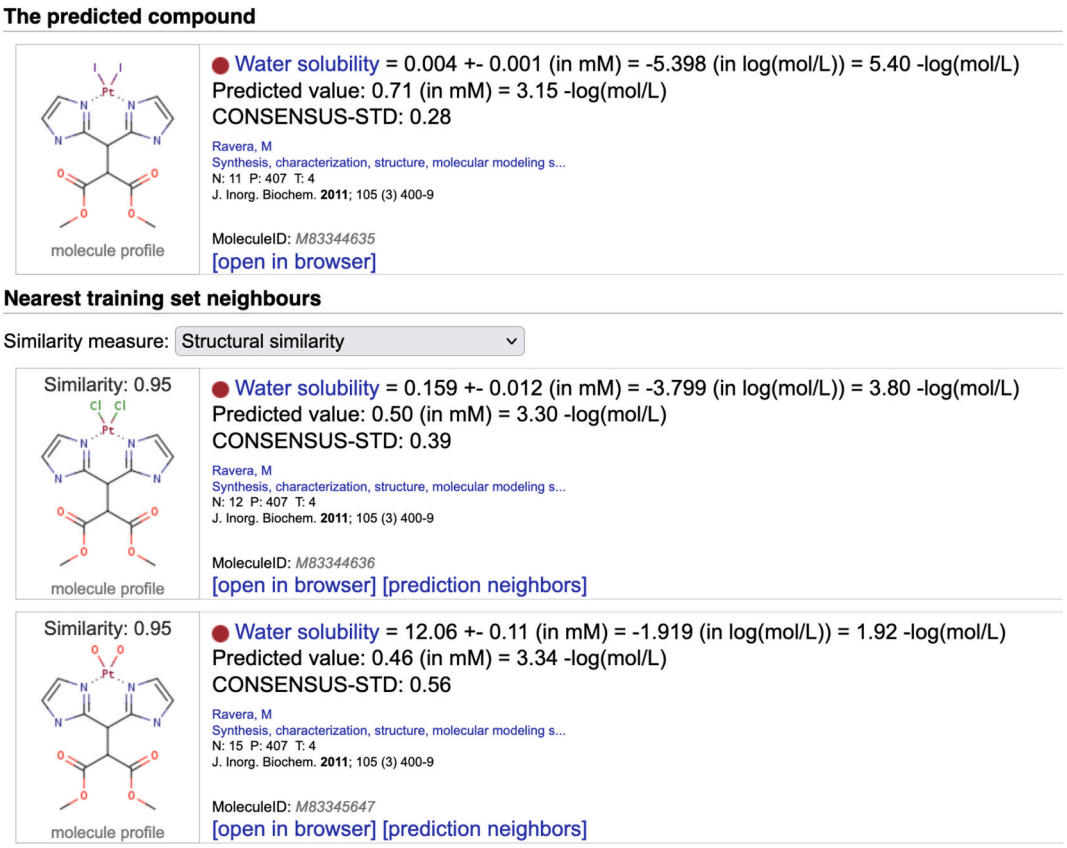


Fig. 6. Di-iodido bis(imidazole) derivative ("DiI-ImH", as shown in Fig. 3), the second largest outlying compound, and its two nearest analogues in the training set.

set. Note two large compounds were excluded (Fig. S2), as they were large outliers for the regression model and 3D conversion of these compounds failed due to their size. In addition to EFGs we also used predicted logP [27] as a descriptor which significantly improved the model, yielding RMSE of 0.89 ± 0.04 and 1.03 ± 0.08 for training and test sets respectively, and 0.83 ± 0.03 for the combined set, respectively. The functional groups with regression coefficients selected by the model for the combined set are listed in Table 2.

Positive coefficients indicate that the appearance of the respective functional group increases solubility, while groups with negative coefficients decrease it. For example, functional groups with positive coefficients, such as cations or hydroxy compounds, or esters typically increase solubility due to their ability to engage in hydrogen bonding or ion-dipole interactions with water molecules, which enhances aqueous compatibility. In contrast, groups with negative coefficients, such as halogens decrease it.

Interestingly, the regression equation includes three separate terms for halogen effects, which allow for a nuanced "fine-tuning" of their contributions depending on the specific functional group and molecular environment. In general, the presence of halogens is associated with decreasing solubility. This is confirmed by their overrepresentation among insoluble compounds in Fig. 7 and their negative coefficient in Table 2. However, their effect varies depending on the types of atoms to which they are bonded within the molecule– the molecular context. Hence, Table 2 includes two additional terms: 'Aryl halides,' which have a large negative coefficient, and another for 'halogen derivatives including (alkyl, alkenyl, aryl)' which contribute positively to solubility. These distinctions arise due to the overlapping effects among the functional groups (cross-dependencies), and the linear model attempts to partition their contributions across multiple coefficients or terms. This analysis highlights the complexity of developing and interpreting even linear models when using seemingly interpretable groups.

3.8. Limitations of models: Accounting for geometric isomerism and stereochemistry

Diastereomers lacking explicit chiral centers annotations would be challenging for 2D descriptors to distinguish, as they are chemically identical but have different spatial arrangements. For example [erythro-1,2-diamino-1-(4-fluorophenyl)propan-1-ol]dichloridoplatinum(II) and [threo-1,2-diamino-1-(4-fluorophenyl)propan-1-ol]dichloridoplatinum (II) have measured solubilities of −3.22 and − 3.99 log(mol/L), respectively. [75] However, both compounds were predicted to have a solubility of −3.66 log(mol/L), which could be attributed to the absence of explicit chiral centers annotation in the source paper. The model is also unable to distinguish cis/trans configurations. E.g., for cis/trans forms of {[2-(ethylsulfanyl)acetyl]oxy}platino2-(ethylsulfanyl)acetate [76] which had measured solubilities of −1.50 and − 2.94 log(mol/L), respectively. The model predicted a value of −1.95 log(mol/L) for each of them. This limitation of the current method should be addressed in the future with the inclusion of 3D descriptors, following the correct prediction of molecular crystal structures.

3.9. Model for simultaneous prediction of solubility and lipophilicity

Given the correlation between logP and water solubility, we decided to develop a multitask model to predict them simultaneously. The RF implementation available in OCHEM does not support the simultaneous development of models for several properties. Therefore we used the ASNN as well as Transformer CNN, which inherently support multitask learning. The combined dataset included 233 logP values from the previous study [27] as well as solubility values from this study (n = 392). The consensus model developed using the same 2D descriptor sets had an RMSE = 0.62 for solubility and an RMSE = 0.44 for logP (which was similar to RMSE = 0.40 reported for logP elsewhere [27]).

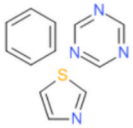
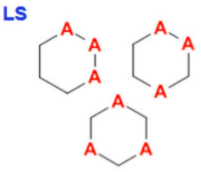
Descriptor		In set 1 (188 unique molecules)	In set 2 (184 unique molecules)	Enrichment factor	p-Value*
Halogens		125 (66.5%)	67 (36.4%)	1.8	2.0E-7
Aromatic compounds		91 (48.4%)	44 (23.9%)	2.0	3.0E-5
Arene		87 (46.3%)	41 (22.3%)	2.1	4.0E-5
Any cations		30 (16.0%)	58 (31.5%)	2.0	-0.01
Saturated six-membered heterocycles with three heteroatoms		32 (17.0%)	60 (32.6%)	1.9	-0.02
Dialkylethers		3 (1.6%)	18 (9.8%)	6.1	-0.02

Fig. 7. Overrepresented EFGs between lower and higher soluble complexes. Positive and negative *p*-values indicate that the group is overrepresented in the set of lower or highly soluble compounds, respectively.

Considering that the multitask model was trained to fit two properties simultaneously and leveraged a larger dataset size (625 compounds), this multitask model is therefore more robust and is recommended to predict solubility and lipophilicity of Pt complexes. Indeed, this multitask model provided improved predictions for compounds with Pt—I bonds, an RMSE = 0.8 compared to an RMSE = 1.3 obtained with the model based on logS data alone.

3.10. Applicability domain

The applicability domains (AD) of the consensus models were based on standard deviation of the individual models. [78]. The AD was defined as standard deviation covering 95 % of data in the training set. Each prediction of the models is accompanied by confidence intervals. The AD can signal to the users that predictions are not reliable. Of course, chemical space is huge and such a situation can easily arise when analyzing new data. For example, application of a multitask model to logP data of $n = 9$ alizarin derivatives [77] indicated that all predictions were outside of the AD of the model (Fig. 8). RMSE = 1.2 was calculated for logP prediction. However, despite the RMSE being high, the model correctly captured trends within this series and a Pearson correlation coefficient $R^2 = 0.7$ between predicted and experimental values was calculated.

3.11. Extension of models with new data

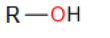
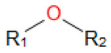
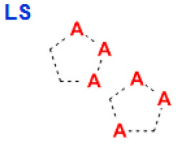
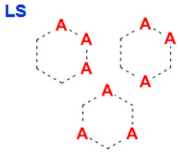
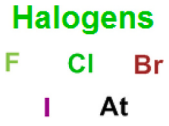
OCHEM allows users to extend the training set with new data and recalculate models. The individual steps involved, including the upload of new data, correction of the representation and the redevelopment of models, are described in the Supplementary Materials. For the previously mentioned dataset of alizarin derivatives [77] the redevelopment of the consensus model after incorporation of these compounds led to the RMSE being reduced to 0.31. Therefore, after retraining the model could accurately predict compounds from this new series. Users are strongly advised to consider predictions within the AD of models, check the predicted and experimental values and extend the model with new data, if required.

3.12. Standardized representation of Pt complexes for accurate model predictions

The correct use of the developed model requires that the platinum complexes collected by the user are represented consistently with the representation methodology used for model development.

Coordination bonds to the Pt center should be explicitly depicted (see Fig. 9 and Fig. S3, S4, S5). Notably, depicting single bonds instead of coordination bonds provided very similar results for the consensus model predictions since the coordination bond is used mainly for Transformer CNN. This similarity in results arises because, for most descriptor calculation programs, coordination bonds are internally

Table 2
Linear regression model based on EFG to predict water solubility of Pt-compounds. The model distinguishes between high specificity (HS) and low specificity (LS) patterns. **HS patterns** only match compounds containing the exact heterocyclic moiety, such as pyrrole (N), furan (O), and thiophene (S) rings. In contrast, **LS patterns** provide a broader classification, generalizing these groups to include any "aromatic" atom except carbon.

Functional group/term	OCHEM representative image	Regression coefficient
constant term		−2.664
predicted logP using model from ref. [27]		- 0.3514
Increasing solubility		
Cations		+0.4066
Hydroxy compounds: alcohols or phenols		+0.268
Ethers		+0.3705
Halogen derivatives (alkyl, alkenyl, aryl)		+0.1932
Five-membered heterocycles with three heteroatoms (LS)		+0.3765
Six-membered heterocycles with three heteroatoms (LS)		+0.1706
Decreasing solubility		
Halogens		−0.1239
Five-membered heterocycles (LS)		−0.4321
Aryl halides		−0.4001

converted to single bonds thus making both these representations equivalent. Exceptions would be the fragmentor ISIDA and QNPR.

For ionic complexes, the counterions (second coordination sphere) must be explicitly included to ensure charge neutrality. The charge within the complex ion (first coordination sphere) should be attributed to the platinum atom itself (see Fig. 9a,b).

Furthermore, representing Pt complexes in an ionized form with a set of separate ligands not bonded or coordinated to the Pt center, rather than depicting explicit coordination or single bonds (Fig. S4) will result in incorrect predictions. The users of the model must adhere to the same representation scheme used in this study to ensure correct solubility predictions.

We note that, in particular with transition metal complexes, many representations and interpretations of the chemical bonds are possible, each with its own advantages and limitations. The chemical interpretation of Pt—N bonds as depicted in Fig. 9a focuses on the bond dissociation limit. Formally, the platinum atom is seen with a charge +2, whereas, e.g., the 3,5-dimethyl pyridine is neutral. In other words, a vacant ***d*** orbital from the Pt center will accept a nitrogen's lone pair, as depicted in Fig. 9c. Though useful for rationalizing reactivity, this representation is in disagreement with quantum mechanical calculations. We performed a Density Functional Theory benchmark on a similar complex, and the calculations reveal a charge on platinum far from the formal value of +2 (Fig. 9d). It is also noteworthy that the chlorine's partial charge differs quite significantly from the formal value of −1. Further, it is seen in Fig. 9d that, although Cl is involved in a single bond (Mayer Valence = 0.94), the Pt atom has an unconventional pattern between 3 and 4 bonds. This shows that any classical interpretation of bonds is limited and cannot cover all chemical aspects of the molecule. As long as the representation models are consistent and systematic, the resulting model should not be limited by any specific representation. To make matters more complicated, the quantum mechanical analysis of molecules is not restricted to a single methodology. Though still used in applied computational chemistry, Mulliken and Mayer population analyses are deprecated by the theoretical chemistry community, e.g., to Hirschfeld analysis. The purpose of this evaluation was, however, to illustrate that many bond descriptors are possible, and that they will most likely disagree with one another, even if only slightly. Translating these insights to our ML applications, this illustrates that it is not possible to obtain a unique bond descriptor with which to train ML models. The most important consideration is that the cheminformatics representation of atoms and bonds is consistent, as this is the information that the model learns.

Predictions of water solubility and lipophilicity of Pt-complexes are certainly not sufficient to explain the biological activity of compounds. Understanding what the body does to the drug (pharmacokinetics) is also extremely important. To a large degree, the bioavailability of a drug is dependent on its solubility, lipophilicity, and ability to cross membranes, bind to plasma proteins, etc., which are part of the ADMET parameters widely used in drug development. [1,2,5] Naturally, it should be even more important to consider the real milieu and to determine the concentration of the drug at the site of action. For example, the measurement of solubility in serum (or artificial complex environments simulating serum) could provide better, more physiologically relevant estimates of the bioavailability of compounds. However, this is almost impossible, especially in the early stages of drug discovery, where very large numbers of compounds are screened. Such measurements could only be made for a small, self-consistent study on a limited number of samples. Therefore, simpler and more standardized pharmacologically relevant physicochemical properties, such as water solubility and lipophilicity, are widely used by pharma companies as approximations of drug bioavailability.

4. Conclusions

The primary aim of this study is to evaluate the current state-of-the-art algorithms in accurately predicting the solubility of platinum complexes. In this respect, we collected the largest database of Pt-complex solubilities to this date (Nov. 2024). The database was further augmented by solubility values ($n = 18$) which are reported publicly for the first time in this study (Table S1).

In order to ensure the use of the best informative features for our predictive task, we initially analyzed the performance of RF and Transformer CNN models. Training RF models using a wide range of 2D and 3D descriptors, we found that the spatial information generated by the 3D descriptors did not provide significant improvement to the RMSE values. In addition, generating 3D structures failed for few large compounds, therefore we decided to use only the best performing 2D descriptors to develop our consensus model.

Further analyses highlighted the comparative effectiveness of classical (RF) and modern (Transformer CNN) ML models in solubility prediction of Pt complexes. Consensus models that leveraged the strengths of descriptor- and learning-representation-based highlighted the potential benefits of incorporating diverse methods to enhance model performance and generalizability in solubility prediction tasks. We have developed three different models, of increasing complexity and

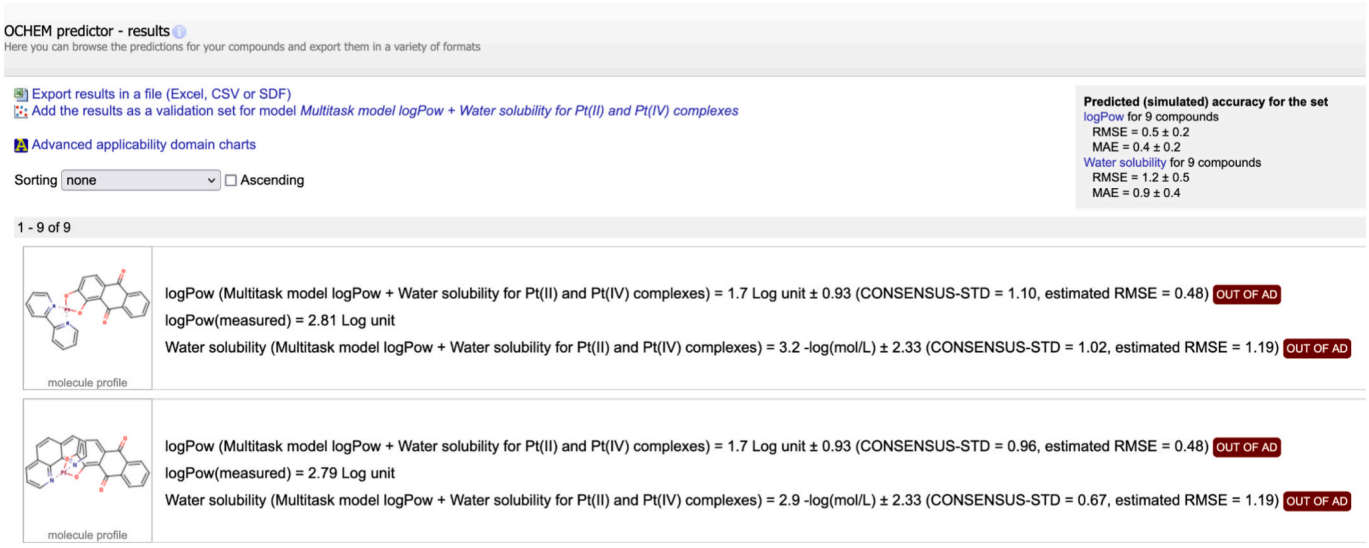


Fig. 8. The application of the multitask consensus model to alizarin derivatives [77]. The model reports that the analyzed compounds are outside of its applicability domain and thus its predictions of their properties are unreliable.

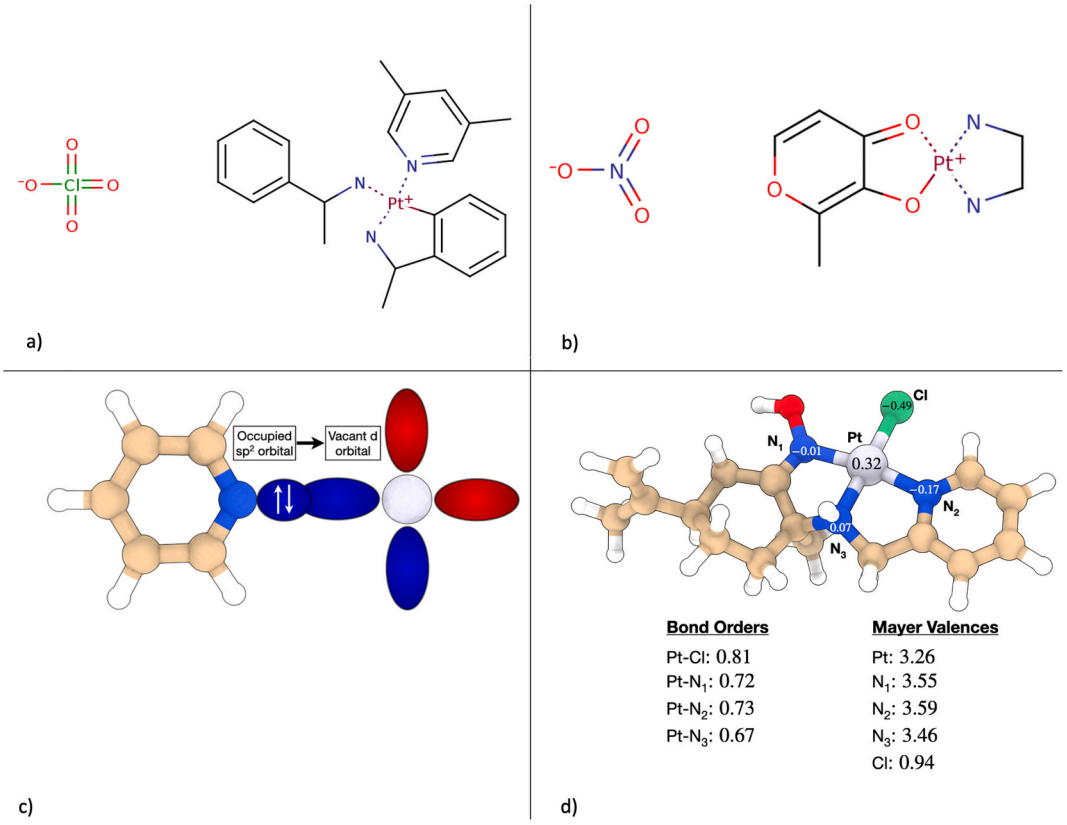


Fig. 9. Representation of (ionic) complexes used for model development. Dotted lines are coordination N—Pt bonds. a-b) Examples of cationic complexes. c) Schematic representation of the bonding in a Pt complex with a pyridine ligand. A metal's vacant d orbital receives two electrons from the nitrogen's lone pair. d) Calculated chemical descriptors for the Pt complex. On top of each atom, the Mulliken partial charge is reported, evidencing the deviations with respect to the expected formal charges. Mayer bond orders and valences are also given. The method and software used for the Density Functional Calculations are detailed in the supplementary material (see S6: Density Functional Calculations).

number of data points: (1) a model developed using temporally split dataset (training set = 284, test set = 108), (2) a model developed using a combined dataset (n = 392), and (3) a multitask model which simultaneously predicted solubility and lipophilicity of Pt complexes as two endpoints (n = 625). Interestingly all three models had the same RMSE = 0.62 log(mol/L). Given the fact that the multitask model developed using the largest number of experimental data, it should exhibit the highest robustness and is recommended for practical applications. Notably, this study presents the first publicly available model for predicting the solubility of Pt complexes, providing a valuable tool for the

scientific community, and addressing the existing gap in research in this area.

The outlier analysis with respect to the chemical space converge analysis revealed that the underrepresentation of certain structural features in the training set was the primary cause of high-prediction errors in both the 5CV sets and the test set. This finding was emphasized by the accurate predictions of the previously-seen outliers in the test set when the model was trained on the combined dataset and further on an expanded training set with the multitask model. This confirms that integrating a more diverse array of molecules does indeed enhance the accuracy and generalizability of predictive tasks of machine learning models within drug discovery. Given the fact that the research into platinum complexes is still actively ongoing, with scientists exploring new chemical classes or optimizing the existing ones, our model's predictive power is expected to improve with further expansion and diversification of available datasets.

Analysis of the chemical features associated with lower solubility of platinum complexes outlined that the presence of aromatics, halogens or, in particular, five-membered heterocycles, unsaturated six-membered heterocycles with heteroatoms, aryl halides and nonmetals (Table 2 and Fig. 7) can decrease the solubility of Pt-complexes. Achieving optimal solubility without compromising lipophilicity and, consequently, therapeutic efficacy requires a careful selection of ligands and counterions that can enhance water interactions while maintaining the desired biological activity. In this regard, the interpretable EFG-based linear model presented in this study offers valuable insights for future structural feature selection or modification strategies aiming at improving the pharmacokinetic properties of platinum complexes.

CRediT authorship contribution statement

Nesma Mousa: Writing – review & editing, Writing – original draft, Visualization, Software, Investigation, Formal analysis, Data curation. **Hristo P. Varbanov:** Writing – review & editing, Data curation. **Vidya Kaipanchery:** Writing – original draft, Investigation, Data curation. **Elisabetta Gabano:** Writing – review & editing, Data curation. **Mauro Ravera:** Writing – review & editing. **Andrey A. Toropov:** Investigation. **Larisa Charochkina:** Data curation. **Filipe Menezes:** Writing – review & editing, Visualization, Formal analysis. **Guillaume Godin:** Software, Investigation. **Igor V. Tetko:** Writing – review & editing, Supervision, Software, Investigation, Formal analysis, Data curation, Conceptualization.

Declaration of competing interest

Igor V. Tetko is CEO of BIGCHEM GmbH (https://bigchem.de) which licenses OCHEM (https://ochem.eu) software. The other authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

We acknowledge Tanzeem Haque for collecting the initial solubility data, Mathias Klose for his assistance with ICP-MS measurements for some solubility determinations, and Katya Ahmad for her comments and remarks to the manuscript. The study was partially supported with Marie Sklodowska-Curie Innovative Training Network European Industrial Doctorate grant agreement No. 956832 "Advanced machine learning for Innovative Drug Discovery" (AIDD) https://ai-dd.eu. The costs for open access publication were covered through CRUI Consortium (Prof. M. Ravera).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jinorgbio.2025.112890.

Data availability

The data are publicly available at https://ochem.eu/article/31.

References

[1] K.V. Balakin, N.P. Savchuk, I.V. Tetko, In silico approaches to prediction of aqueous and DMSO solubility of drug-like compounds: trends, problems and solutions, Curr. Med. Chem. 13 (2) (2006) 223–241, 16472214.
[2] I.V. Tetko, P. Bruneau, H.-W. Mewes, D.C. Rohrer, G.I. Poda, Can we estimate the accuracy of ADME-Tox predictions? Drug Discov. Today 11 (15–16) (2006 Aug) 700–707. S1359-6446(06)00230-3.
[3] L. Kumari, et al., Advancement in Solubilization approaches: a step towards bioavailability enhancement of poorly soluble drugs, Life 13 (5) (2023), https://doi.org/10.3390/life13051099.
[4] D.S. Palmer, J.B.O. Mitchell, Is experimental data quality the limiting factor in predicting the aqueous solubility of Druglike molecules? Mol. Pharm. 11 (8) (2014 Aug) 2962–2972, https://doi.org/10.1021/mp500103r.
[5] I.V. Tetko, A. Yan, J. Gasteiger, Prediction of physicochemical properties of compounds, Appl. Chemoinform. (2018) 53–81, https://doi.org/10.1002/9783527806539.ch3.
[6] A. Avdeef, Prediction of aqueous intrinsic solubility of druglike molecules using random Forest regression trained with wiki-pS0 database, ADMET DMPK 8 (1) (2020 Mar) 29–77, https://doi.org/10.5599/admet.766.
[7] W.L. Jorgensen, E.M. Duffy, Prediction of drug solubility from structure, Comput. Methods Predict. ADME Toxic. 54 (3) (2002 Mar) 355–366, https://doi.org/10.1016/S0169-409X(02)00008-X.
[8] G. Panapitiya, et al., Evaluation of deep learning architectures for aqueous solubility prediction, ACS Omega 7 (18) (2022 May) 15695–15710, https://doi.org/10.1021/acsomega.2c00642.
[9] T. Deng, G. Jia, Prediction of aqueous solubility of compounds based on neural network, Mol. Phys. 118 (2) (2020 Jan) e1600754, https://doi.org/10.1080/00268976.2019.1600754.
[10] Z. Xiong, et al., Pushing the boundaries of molecular representation for drug discovery with the graph attention mechanism, J. Med. Chem. 63 (16) (2020 Aug) 8749–8760, https://doi.org/10.1021/acs.jmedchem.9b00959.
[11] Q. Cui, et al., Improved prediction of aqueous solubility of novel compounds by going deeper with deep learning, Front. Oncol. 10 (2020), https://doi.org/10.3389/fonc.2020.00121 [Online]. Available.
[12] C.W. Coley, R. Barzilay, W.H. Green, T.S. Jaakkola, K.F. Jensen, Convolutional embedding of attributed molecular graphs for physical property prediction, J. Chem. Inf. Model. 57 (8) (2017 Aug) 1757–1772, https://doi.org/10.1021/acs.jcim.6b00601.
[13] M. Withnall, E. Lindelöf, O. Engkvist, H. Chen, Building attention and edge message passing neural networks for bioactivity and physical–chemical property prediction, J. Chemother. 12 (1) (2020 Jan) 1, https://doi.org/10.1186/s13321-019-0407-y.
[14] B. Tang, S.T. Kramer, M. Fang, Y. Qiu, Z. Wu, D. Xu, A self-attention based message passing neural network for predicting molecular lipophilicity and aqueous solubility, J. Chemother. 12 (1) (2020 Feb) 15, https://doi.org/10.1186/s13321-020-0414-z.
[15] Z. Xia, P. Karpov, G. Popowicz, I.V. Tetko, Focused library generator: case of Mdmx inhibitors, J. Comput. Aided Mol. Des. 34 (7) (2020 Jul) 769–782, https://doi.org/10.1007/s10822-019-00242-8.
[16] M.C. Sorkun, A. Khetan, S. Er, AqSolDB, a curated reference set of aqueous solubility and 2D descriptors for a diverse set of compounds, Sci. Data 6 (1) (2019 Aug) 143, https://doi.org/10.1038/s41597-019-0151-1.
[17] I.V. Tetko, R. van Deursen, G. Godin, Be aware of overfitting by hyperparameter optimization!, J. Chemother. 16 (1) (2024 Dec) 139, https://doi.org/10.1186/s13321-024-00934-w.
[18] A. Hunklinger, P. Hartog, M. Šícho, G. Godin, I.V. Tetko, The openOCHEM consensus model is the best-performing open-source predictive model in the first EUOS/SLAS joint compound solubility challenge, SLAS Discov. 29 (2) (2024 Mar) 100144, https://doi.org/10.1016/j.slasd.2024.01.005.
[19] A. Yüksel, E. Ulusoy, A. Ünlü, T. Doğan, SELFormer: molecular representation learning via SELFIES language models, Mach. Learn. Sci. Technol. 4 (2) (2023 Jun) 025035, https://doi.org/10.1088/2632-2153/acdb30.
[20] U. Jungwirth, C.R. Kowol, B.K. Keppler, C.G. Hartinger, W. Berger, P. Heffeter, Anticancer activity of metal complexes: involvement of redox processes, Antioxid. Redox Signal. 15 (4) (2011 Aug) 1085–1127, https://doi.org/10.1089/ars.2010.3663.
[21] C. Jia, et al., Synthesis, characterization, and biological activity of new mixed ammine/amine platinum(IV) complexes, Appl. Organomet. Chem. 34 (8) (2020) e5680, https://doi.org/10.1002/aoc.5680.
[22] I. Zanellato, et al., Biological activity of a series of cisplatin-based aliphatic bis (carboxylato) Pt(IV) prodrugs: how long the organic chain should be? J. Inorg. Biochem. 140 (2014 Nov) 219–227, https://doi.org/10.1016/j.jinorgbio.2014.07.018.
[23] U. Ndagi, N. Mhlongo, M.E. Soliman, Metal complexes in cancer therapy – an update from drug design perspective, Drug Des. Devel. Ther. 11 (2017 Mar) 599–616, https://doi.org/10.2147/DDDT.S119488.

[24] M.J. Cleare, J.D. Hoeschele, Studies on the antitumor activity of group VIII transition metal complexes. Part I. Platinum (II) complexes, Bioinorg. Chem. 2 (3) (1973 Jan) 187–210, https://doi.org/10.1016/S0006-3061(00)80249-5.

[25] E. Armstrong-Gordon, et al., Patterns of platinum drug use in an acute care setting: a retrospective study, J. Cancer Res. Clin. Oncol. 144 (8) (2018 Aug) 1561–1568, https://doi.org/10.1007/s00432-018-2669-6.

[26] S. Karmakar, I. Poetsch, C.R. Kowol, P. Heffeter, D. Gibson, Synthesis and cytotoxicity of water-soluble dual- and triple-action Satraplatin derivatives: replacement of equatorial chlorides of Satraplatin by acetates, Inorg. Chem. 58 (24) (2019 Dec) 16676–16688, https://doi.org/10.1021/acs.inorgchem.9b02796.

[27] I.V. Tetko, et al., Prediction of logP for Pt(II) and Pt(IV) complexes: comparison of statistical and quantum-chemistry based approaches, J. Inorg. Biochem. 156 (2016 Mar) 1–13, https://doi.org/10.1016/j.jinorgbio.2015.12.006.

[28] I.V. Tetko, I. Jaroszewicz, J.A. Platts, J. Kuduk-Jaworska, Calculation of lipophilicity for Pt(II) complexes: experimental comparison of several methods, J. Inorg. Biochem. 102 (7) (2008 Jul) 1424–1437. S0162-0134(08)00015-9.

[29] A.A. Toropov, A.P. Toropova, Application of the Monte Carlo method for building up models for octanol-water partition coefficient of platinum complexes, Chem. Phys. Lett. 701 (2018 Jun) 137–146, https://doi.org/10.1016/j.cplett.2018.04.012.

[30] A.A. Toropov, A.P. Toropova, P.G.R. Achary, Prediction of n-octanol–water partition coefficient of platinum (IV) complexes using correlation weights of fragments of local symmetry, Struct. Chem. 34 (4) (2023 Aug) 1517–1526, https://doi.org/10.1007/s11224-023-02197-x.

[31] I.V. Tetko, V.Y. Tanchuk, T.N. Kasheva, A.E. Villa, Estimation of aqueous solubility of chemical compounds using E-state indices, J. Chem. Inf. Comput. Sci. 41 (6) (2001) 1488–1493, 11749573.

[32] P. Kapitza, et al., Benzimidazole-based NHC metal complexes as anticancer drug candidates: gold(I) vs. platinum(II), Inorganics 11 (7) (2023), https://doi.org/10.3390/inorganics11070293.

[33] H.P. Varbanov, S.M. Valiahdi, C.R. Kowol, M.A. Jakupec, M.S. Galanski, B. K. Keppler, Novel tetracarboxylatoplatinum(iv) complexes as carboplatin prodrugs, Dalton Trans. 41 (47) (2012) 14404–14415, https://doi.org/10.1039/C2DT31366A.

[34] H. Varbanov, et al., Synthesis and characterization of novel bis(carboxylato) dichloridobis(ethylamine)platinum(IV) complexes with higher cytotoxicity than cisplatin, Eur. J. Med. Chem. 46 (11) (2011 Nov) 5456–5464, https://doi.org/10.1016/j.ejmech.2011.09.006.

[35] S. Göschl, H.P. Varbanov, S. Theiner, M.A. Jakupec, M.S. Galanski, B.K. Keppler, The role of the equatorial ligands for the redox behavior, mode of cellular accumulation and cytotoxicity of platinum(IV) prodrugs, J. Inorg. Biochem. 160 (2016 Jul) 264–274, https://doi.org/10.1016/j.jinorgbio.2016.03.005.

[36] H.P. Varbanov, et al., A novel class of Bis- and Tris-chelate Diam(m)inebis (dicarboxylato)platinum(IV) complexes as potential anticancer prodrugs, J. Med. Chem. 57 (15) (2014 Aug) 6751–6764, https://doi.org/10.1021/jm500791c.

[37] M. Ravera, et al., Functional fluorescent nonporous silica nanoparticles as carriers for Pt(IV) anticancer prodrugs, J. Inorg. Biochem. 151 (2015 Oct) 132–142, https://doi.org/10.1016/j.jinorgbio.2015.08.001.

[38] I. Sushko, 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 (6) (2011 Jun) 533–554, https://doi.org/10.1007/s10822-011-9440-2.

[39] C. Zhang, C. Xu, X. Gao, Q. Yao, Platinum-based drugs for cancer therapy and anti-tumor strategies, Theranostics 12 (5) (2022) 2115–2132, https://doi.org/10.7150/thno.69424.

[40] A. Ibrahim, S. Hirschfeld, M.H. Cohen, D.J. Griebel, G.A. Williams, R. Pazdur, FDA drug approval summaries: oxaliplatin, Oncologist 9 (1) (2004 Feb) 8–12, https://doi.org/10.1634/theoncologist.9-1-8.

[41] Z. Lin, W.Z. Lv, S.Y. Wang, J.L. Zou, Y.Y. Con, Z.H. Wang, M. Xiao, P.J. Peng, Efficacy and safety of pemetrexed and nedaplatin followed by pemetrexed maintenance therapy in advanced lung adenocarcinoma, Cancer Manag. Res. 20 (9) (2025) 671–677, https://doi.org/10.2147/CMAR.S150975.

[42] S. Dilruba, G.V. Kalayda, Platinum-based drugs: past, present and future, Cancer Chemother. Pharmacol. 77 (6) (2016 Jun) 1103–1124, https://doi.org/10.1007/s00280-016-2976-z.

[43] A. Daina, O. Michielin, V. Zoete, SwissADME: a free web tool to evaluate pharmacokinetics, drug-likeness and medicinal chemistry friendliness of small molecules, Sci. Rep. 7 (1) (2017 Mar) 42717, https://doi.org/10.1038/srep42717.

[44] B. Bienfait, P. Ertl, JSME: a free molecule editor in JavaScript, J. Chemother. 5 (1) (2013 May) 24, https://doi.org/10.1186/1758-2946-5-24.

[45] P. Karpov, G. Godin, I.V. Tetko, Transformer-CNN: Swiss knife for QSAR modeling and interpretation, J. Chemother. 12 (1) (2020 Mar) 17, https://doi.org/10.1186/s13321-020-00423-w.

[46] A. Varnek, et al., ISIDA - platform for virtual screening based on fragment and Pharmacophoric descriptors, Curr. Comput. - Aided Drug Des. 4 (2008 Sep) 191–198, https://doi.org/10.2174/157340908785747465.

[47] Molecule Preprocessing - OCHEM User's Manual - OCHEM Docs, Accessed: Jan. 10, 2024. [Online]. Available, https://docs.ochem.eu/display/MAN/Molecule+preprocessing.html, 2024.

[48] L. Breiman, Random forests, Mach. Learn. 45 (1) (2001 Oct) 5–32, https://doi.org/10.1023/A:1010933404324.

[49] I.V. Tetko, Associative neural network, in: D.J. Livingstone (Ed.), Artificial Neural Networks: Methods and Applications, Humana Press, Totowa, NJ, 2009, pp. 180–197, https://doi.org/10.1007/978-1-60327-101-1_10.

[50] A.E. Villa, I.V. Tetko, P. Dutoit, Y. De Ribaupierre, F. De Ribaupierre, Corticofugal modulation of functional connectivity within the auditory thalamus of rat, guinea pig and cat revealed by cooling deactivation, J. Neurosci. Methods 86 (2) (1999 Jan) 161–178, 10065984.

[51] L.H. Hall, L.B. Kier, Electrotopological state indices for atom types: a novel combination of electronic, topological, and valence state information, J. Chem. Inf. Comput. Sci. 35 (6) (1995 Nov) 1039–1045, https://doi.org/10.1021/ci00028a014.

[52] H. Hong, et al., Mold2, molecular descriptors from 2D structures for Chemoinformatics and Toxicoinformatics, J. Chem. Inf. Model. 48 (7) (2008 Jul) 1337–1344, https://doi.org/10.1021/ci800038f.

[53] H. Moriwaki, Y.-S. Tian, N. Kawashita, T. Takagi, Mordred: a molecular descriptor calculator, J. Chemother. 10 (1) (2018 Feb) 4, https://doi.org/10.1186/s13321-018-0258-y.

[54] I.V. Tetko, D.M. Lowe, A.J. Williams, The development of models to predict melting and pyrolysis point data associated with several hundred thousand compounds mined from PATENTS, J. Chemother. 8 (1) (2016 Jan) 2, https://doi.org/10.1186/s13321-016-0113-y.

[55] D. Rogers, M. Hahn, Extended-connectivity fingerprints, J. Chem. Inf. Model. 50 (5) (2010 May) 742–754, https://doi.org/10.1021/ci100050t.

[56] G. Landrum, RDKit: Open-Source Cheminformatics, 2006.

[57] E.S. Salmina, N. Haider, I.V. Tetko, Extended functional groups (EFG): an efficient set for chemical characterization and structure-activity relationship studies of chemical compounds, Molecules 21 (1) (2016), https://doi.org/10.3390/molecules21010001.

[58] I. Sushko, E. Salmina, V.A. Potemkin, G. Poda, I.V. Tetko, ToxAlerts: a web server of structural alerts for toxic chemicals and compounds with potential adverse reactions, J. Chem. Inf. Model. 52 (8) (2012 Aug) 2310–2316, https://doi.org/10.1021/ci300245q.

[59] A. Mauri, alvaDesc: a tool to calculate and analyze molecular descriptors and fingerprints, in: K. Roy (Ed.), Ecotoxicological QSARs, Springer US, New York, NY, 2020, pp. 801–820, https://doi.org/10.1007/978-1-0716-0150-1_32.

[60] F. Menezes, G.M. Popowicz, ULYSSES: an efficient and easy to use Semiempirical library for C++, J. Chem. Inf. Model. 62 (16) (2022 Aug) 3685–3694, https://doi.org/10.1021/acs.jcim.2c00757.

[61] H. Zhu, et al., Combinatorial QSAR modeling of chemical toxicants tested against *Tetrahymena pyriformis*, J. Chem. Inf. Model. 48 (4) (2008 Apr) 766–784, https://doi.org/10.1021/ci700443v.

[62] A.D. Aputen, et al., Bioactive platinum(IV) complexes incorporating halogenated Phenylacetates, Molecules 27 (20) (2022), https://doi.org/10.3390/molecules27207120.

[63] E. Gabano, et al., Synthesis and characterization of cyclohexane-1R,2R-diamine-based Pt(iv) dicarboxylato anticancer prodrugs: their selective activity against human colon cancer cell lines, Dalton Trans. 48 (2) (2019) 435–445, https://doi.org/10.1039/C8DT03950J.

[64] Z. Xu, et al., Synthesis, structure, and cytotoxicity of Oxaliplatin-based platinum (IV) anticancer prodrugs bearing one axial fluoride, Inorg. Chem. 57 (14) (2018 Jul) 8227–8235, https://doi.org/10.1021/acs.inorgchem.8b00706.

[65] S. Chen, H. Yao, Q. Zhou, M.-K. Tse, Y.F. Gunawan, G. Zhu, Stability, reduction, and cytotoxicity of platinum(IV) anticancer prodrugs bearing carbamate axial ligands: comparison with their carboxylate analogues, Inorg. Chem. 59 (16) (2020 Aug) 11676–11687, https://doi.org/10.1021/acs.inorgchem.0c01541.

[66] P. Domingo-Legarda, A. Casado-Sánchez, L. Marzo, J. Alemán, S. Cabrera, Photocatalytic water-soluble cationic platinum(II) complexes bearing Quinolinate and phosphine ligands, Inorg. Chem. 59 (19) (2020 Oct) 13845–13857, https://doi.org/10.1021/acs.inorgchem.0c01326.

[67] E. Shaili, L. Salassa, J.A. Woods, G. Clarkson, P.J. Sadler, N.J. Farrer, Platinum(iv) dihydroxido diazido N-(heterocyclic)imine complexes are potently photocytotoxic when irradiated with visible light, Chem. Sci. 10 (37) (2019) 8610–8617, https://doi.org/10.1039/C9SC02644D.

[68] P. Sgarbossa, et al., Pentafluorophenyl platinum(II) complexes of PTA and its N-allyl and N-benzyl derivatives: synthesis, characterization and biological activity, Materials 12 (23) (2019), https://doi.org/10.3390/ma12233907.

[69] M. Ravera, et al., Cis,cis,trans-[PtIVCl2(NH3)2(perillato)2], a dual-action prodrug with excellent cytotoxic and antimetastatic activity, Dalton Trans. 50 (9) (2021) 3161–3177, https://doi.org/10.1039/D0DT04051G.

[70] M. Ravera, et al., Synthesis, characterization, structure, molecular modeling studies and biological activity of sterically crowded Pt(II) complexes containing bis (imidazole) ligands, J. Inorg. Biochem. 105 (3) (2011 Mar) 400–409, https://doi.org/10.1016/j.jinorgbio.2010.12.002.

[71] G. Trovóa, et al., Remarkable lack of biological activity exhibited by a dna-reactive and water-soluble cis-bis(phosphino) platinum(II) complex, J. Chem. Soc. Dalton Trans. 10 (1993) 1547–1550, https://doi.org/10.1039/DT9930001547.

[72] S. Novotarskyi, A. Abdelaziz, Y. Sushko, R. Körner, J. Vogt, I.V. Tetko, ToxCast EPA in vitro to in vivo challenge: insight into the rank-I model, Chem. Res. Toxicol. 29 (5) (2016 May) 768–775, https://doi.org/10.1021/acs.chemrestox.5b00481.

[73] S. Vorberg, I.V. Tetko, Modeling the biodegradability of chemical compounds using the online CHEmical modeling environment (OCHEM), Mol. Inform. 33 (1) (2014 Jan) 73–85, https://doi.org/10.1002/minf.201300030.

[74] Y. Ran, S.H. Yalkowsky, Prediction of drug solubility by the general solubility equation (GSE), J. Chem. Inf. Comput. Sci. 41 (2) (2001 Mar) 354–357, https://doi.org/10.1021/ci000338c.

[75] I. Würtenberger, B. Angermaier, B. Kircher, R. Gust, Synthesis and in vitro pharmacological behavior of platinum(II) complexes containing 1,2-Diamino-1-(4-fluorophenyl)-2-alkanol ligands, J. Med. Chem. 56 (20) (2013 Oct) 7951–7964, https://doi.org/10.1021/jm400967z.

[76] L. Ramberg, Notiz über die Umlagerung eines inneren Komplexsalzes durch Belichtung, Ber. Dtsch. Chem. Ges. 43 (1) (1910 Jan) 580–584, https://doi.org/10.1002/cber.19100430198.
[77] R. Caligiuri, et al., Cytotoxic Pt(ii) complexes containing alizarin: a selective carrier for DNA metalation, Dalton Trans. 53 (6) (2024) 2602–2618, https://doi.org/10.1039/D3DT03889K.
[78] I.V. Tetko, 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 (9) (2008 Sep) 1733–1746, https://doi.org/10.1021/ci800151m.