Journal of Cheminformatics

RESEARCH Open Access

Prediction of Pt, Ir, Ru, and Rh complexes light absorption in the therapeutic window for phototherapy using machine learning

V. Vigna[1*], T. F. G. G. Cova[2], A. A. C. C. Pais[2] and E. Sicilia[1]

Abstract

Effective light-based cancer treatments, such as photodynamic therapy (PDT) and photoactivated chemotherapy (PACT), rely on compounds that are activated by light efficiently, and absorb within the therapeutic window (600–850 nm). Traditional prediction methods for these light absorption properties, including Time-Dependent Density Functional Theory (TDDFT), are often computationally intensive and time-consuming. In this study, we explore a machine learning (ML) approach to predict the light absorption in the region of the therapeutic window of platinum, iridium, ruthenium, and rhodium complexes, aiming at streamlining the screening of potential photoactivatable prodrugs. By compiling a dataset of 9775 complexes from the Reaxys database, we trained six classification models, including random forests, support vector machines, and neural networks, utilizing various molecular descriptors. Our findings indicate that the Extreme Gradient Boosting Classifier (XGBC) paired with AtomPairs2D descriptors delivers the highest predictive accuracy and robustness. This ML-based method significantly accelerates the identification of suitable compounds, providing a valuable tool for the early-stage design and development of phototherapy drugs. The method also allows to change relevant structural characteristics of a base molecule using information from the supervised approach.

Scientific Contribution: The proposed machine learning (ML) approach predicts the ability of transition metal-based complexes to absorb light in the UV–vis therapeutic window, a key trait for phototherapeutic agents. While ML models have been used to predict UV–vis properties of organic molecules, applying this to metal complexes is novel. The model is efficient, fast, and resource-light, using decision tree-based algorithms that provide interpretable results. This interpretability helps to understand classification rules and facilitates targeted structural modifications to convert inactive complexes into potentially active ones.

Keywords Photodynamic therapy, Classification, Photoactivated chemotherapy, UV–vis, Machine learning

Introduction

Cancer is currently one of the deadliest diseases, causing millions of deaths each year [1]. Although established standard chemotherapy is a reasonably successful strategy for fighting cancer, challenges persist related to inherent or acquired cell resistance and adverse side effects [2, 3]. In order to address these challenges, it is essential to develop more efficient and less invasive alternative approaches compared to currently used therapies [4]. Cancer treatment is one of the fields of medicine that

*Correspondence:
V. Vigna
vincenzo.vigna@unical.it
[1] PROMOCS Laboratory, Department of Chemistry and Chemical Technologies, University of Calabria, Arcavacata di Rende (CS), Italy
[2] Coimbra Chemistry Centre, Department of Chemistry, Institute of Molecular Sciences (IMS), Faculty of Sciences and Technology, University of Coimbra, Coimbra, Portugal

has significantly benefited from the use of therapeutic effects of light, and the employing of light-based strategies is becoming increasingly important in this field, exploiting advancements in photonics and optical technologies to enhance therapeutic and diagnostic applications. Photodynamic therapy (PDT) and photoactivated chemotherapy (PACT) are both therapies relying on the use of light for activating drugs that are relatively non-toxic in dark [5, 6]. Light irradiation enables precise spatial and temporal control over drug activation, enhancing selectivity and reducing side-effects. Transition metal complexes represent promising candidates as photoactivatable prodrugs due to their favorable photophysical and photochemical properties, which can be finely tuned through subtle modifications of ligands, metal center and oxidation state. The mechanism of action of clinically approved PDT is based on the interaction of three key components: a photosensitizer (PS), light, and molecular oxygen [7]. PACT, instead, is an oxygen-independent phototherapy, which involves chemical changes of prodrugs upon irradiation [8–10]. For both techniques, the wavelength of activating light ideally falls within the therapeutic window, a spectral range where light optimally penetrates biological tissues up to 1 cm without causing damage from high-energy irradiation [10]. This window definition is influenced by tissue optical properties and constituent concentrations, making it challenging to establish a precise definition. The most common spectral region identifying the therapeutic window for PDT and PACT extends typically from 600 to 850 nm [11, 12], as PDT photosensitizers activated beyond 800 nm are usually not efficient in promoting molecular oxygen from the triplet to the singlet state. Therefore, being the absorption of light in the correct near-infrared spectral region, the first prerequisite for a compound to be considered a promising PDT and/or PACT agent, prediction of the maximum absorption wavelength in the former stages of the drug design process can significantly accelerate the selection of suitable compounds, even before their synthesis. The time dependent version of Density Functional Theory, TDDFT, is the method of choice for computing spectral properties due to its reliable outcomes at an affordable computational cost [12]. However, the accuracy of TDDFT results strongly relies on the specific computational protocol tailored for each type of system, requiring an accurate preliminary benchmark [14, 15].

Machine learning (ML) has opened a new frontier in theoretical and computational chemistry by allowing to conjugate accuracy and efficiency and also the accelerated ML discovery and design of transition-metal complexes, despite some of additional challenges due to their peculiar properties, with desired characteristics is showing rapid progress across a range of applications [16–18].

This study integrates ML techniques for an efficient screening of numerous metal complexes, aiming at extracting crucial information to guide the design and development of potential agents for PDT and/or PACT applications. Decision Tree was trained on the same dataset solely for the purpose of extrapolating information regarding the classification, rather than as a predictive model. The model facilitated the identification of an inactive compound with respect to its therapeutic window uptake. Consequently, a structural modification was proposed, based on the features considered and extrapolated decisions.

ML models have been recently used to predict a wide array of UV–vis spectral properties for organic molecules [18–20], but the use of this approach is new for the category of metal complexes.

Methods

Data collection

Molecular structures of metal complexes were obtained from the Reaxys database (http://www.reaxys.com) with associated UV–vis absorption wavelength (λ). The initial dataset consisted in 9775 Pt, Rh, Ir and Ru complexes. The main classes of molecules in the dataset were described by generating Murcko Scaffold and Skeleton [21]. For each compound, the Simplified Molecular Input Line Entry System (SMILES) [22] was collected after filtering with some shrinkage: only one-fragment molecules, with UV–vis spectroscopy data available and only data extracted from scientific papers and reviews. Compounds were labeled as "active" (classified as 1) if they had at least one UV–vis absorption wavelength between 500 and 850 nm, chosen as the extremes of the therapeutic window. The lower limit of the window was selected on the basis of the values of the maximum absorption wavelength reported for many metal complexes considered suitable as PDT and PACT agents. All other complexes were labeled as "inactive" (and classified as 0).

Chemical representation

The structures of the metal complexes were represented by six types of molecular descriptors: Extended Connectivity Fingerprints (ECFP), 2-Dimensional Chemically Advanced Template Search (CATS2D), 2-Dimensional Atom Pairs (AtomPairs), Functional Groups Count (FGroup), Molecular ACCess System Key (MACCS) and Walk and Path Counts (WPC). All descriptors were calculated using the Alvadesc [23] software from the SMILES string of each molecule. In order to ensure the consistent representation of a molecule regardless of the original representation, Alvadesc employs a series of standardization steps on molecular structures. These include the standardization of nitro groups,

aromatization, and the addition of implicit hydrogen. In addition, some types of descriptors were combined in order to merge information from different classes of descriptors. A detailed explanation of the molecular descriptors explored in this study can be found in the Supporting Information (Additional file, Listing 1).

In order to visualize the molecules in three-dimensional space, Principal Component Analysis (PCA) was applied to the descriptors to reduce their size. This approach can be useful for identifying molecules that are structurally very similar and, consequently, can be eliminated as they do not add any relevant information to the model. Specifically, PCA was applied to all the descriptors listed above and the variance explained by the first three Principal Components (PCs) was calculated. It was, then, necessary to identify the descriptor that best fits the problem. In particular, the type of descriptor was sought whose first three PCs expressed the highest possible variance. In other words, the descriptors whose first three PCs contained the largest possible percentage of the information given by the descriptor before transformation. Of all the listed descriptors, the first three PCs of the WPC descriptor explained 93% of the variance. In this context, WPC was used exclusively to represent molecules in three-dimensional space, to define applicability domain and to perform subsampling as explained below.

Classification models

In order to find the best classifier for predicting the activity of the compounds, a benchmark of six different classification models was employed. Each model was trained with different types of descriptors, using the default hyperparameters provided for each model for a preliminary assessment. The classifiers selected for the benchmark are: Logistic Regressor (LR) [24], Decision Tree Classifier (DTC) [25], Random Forest Classifier (RFC) [26], Support Vector Classifier (SVC) [27], Gradient Boosting Classifier (GBC) [28], all provided by scikit-learn python package [29] and Extreme Boosting Classifier (XGBC) [30].

Cross validation and model evaluation

A total of 9775 samples was collected in literature. After cleaning, a dataset with 4640 different molecule was obtained, which has been divided into training set and final validation set using an 80:20 ratio [31] (3712 for train set and 928 for validation set). The predictive capacity of the model on the training set was evaluated by internal Stratified k-Fold Cross Validation (SkFoldCV) provided by scikit-learn with three different splitting values k = 5 (S5FoldCV), k = 10 (S10FoldCV) and k = 15 (S15FoldCV), in order to assess the robustness of the model, the ability to generalize and to prevent overfitting [32]. The performance of classification models was evaluated by the following metrics: Accuracy (A), Sensitivity (SE), Specificity (SP), Precision (P), F1-score (F1) and AUC-ROC [33]. These metrics were determined by calculating previously false positives (FP), false negatives (FN), true positives (TP) and true negatives (TN) that were considered for the construction of the confusion matrix as well. In addition, true positive rate (TPR) and false positive rate (FPR) were also calculated for the construction of the ROC-curve (ROC) and to calculate the Area Under the ROC-curve (AUC-ROC). For a detailed description of all evaluation metrics see the Additional file, Listing 2.

Model optimization

Permutation Features Importance (PFI) and Recursive Features Elimination (RFE), both provided by scikit-learn, were performed to assess the predictive importance of each feature and eliminate less informative ones. In fact, with these combined strategies, it is possible to evaluate the minimum number of features the model needs for achieving the best results. Hyperparameter optimization was performed by using AUC-ROC value as a metric to maximize the probability of achieving the best hyperparameter combination. The search was carried out by exploring the hyperparameter space using a Random Grid Search (RGS) methodology for a systematic investigation.

Results and discussion

In what follows, the evaluation of undersampling techniques addresses the challenge posed by class imbalance in the dataset. This evaluation emphasizes the distribution of λ values across the initial dataset consisting of 4640 complexes, illustrating the effective rebalancing achieved through undersampling. The study explores the intricacies of feature engineering, detailing the methodologies applied to refine descriptor sets and optimize computational efficiency. The analysis also presents a comprehensive evaluation of classification models, emphasizing the performance metrics derived from various descriptors paired with six distinct classifiers. Finally, the approach to model optimization is outlined, with a focus on feature selection and hyperparameter tuning of the selected XGB_AP2D model, demonstrating its robustness and efficacy across diverse evaluation criteria.

Undersampling

The histogram in Fig. 1 shows the distribution of λ_{MAX} values for the complete dataset (Fig. 1a) composed of 9775 complexes, and pie charts showing the ratio of active to inactive complexes (Fig. 1b, c. As might be expected, the number of active compounds is by far

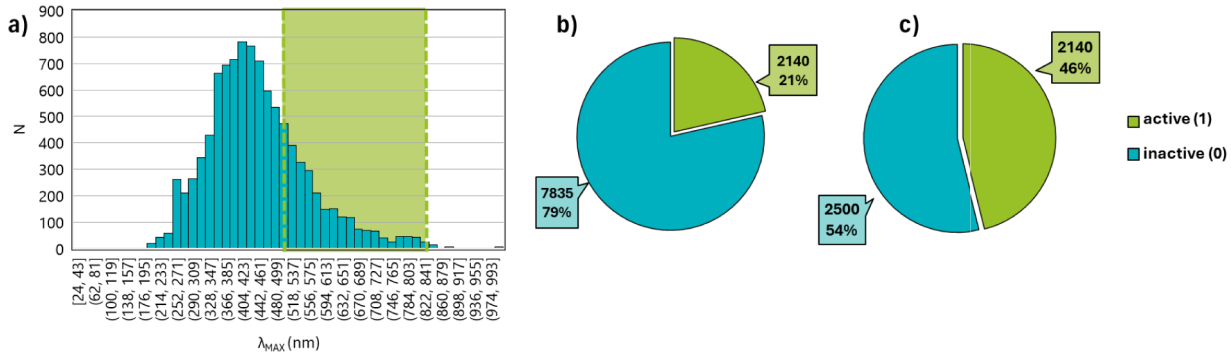


Fig. 1 **a** Distribution of the λ_{MAX} values for the complete initial dataset with therapeutic window highlighted. mean: 449 nm; **b** proportion of the number of samples belonging to the class 'active' and 'inactive' before sampling; **c** proportion of the number of samples belonging to the class 'active' and 'inactive' after undersampling

fewer than inactive ones (Fig. 1a), as the λ_{MAX} of most compounds falls in the region from 300 to 500 nm. Only about 27.3% of the complexes are assigned to the "active" class for a total of 2140 complexes. Unfortunately, class imbalance is a common problem that significantly worsens the performance of ML models [34]. Possible solutions to this problem lie in the use of special ML models that are unaffected by class imbalance, such as XGBC. On the other hand, sampling techniques can also be adopted to improve the performance of the model. In this work, the Majority Class Undersampling technique was used to overcome the imbalance between classes. With this approach, only some of the samples in the predominant class are selected, until the ratio between the two classes reaches approximately 50%. After subsampling, the final dataset was obtained with a total of 4640 complexes of which 2140 were active (46.1%) and 2500 inactive (53.9%) (Fig. 1c).

Undersampling was performed taking into consideration the distribution of all samples in the three-dimensional chemical space generated by applying PCA with WPC descriptors. A more detailed explanation of the undersampling technique can be found in Additional file, Listing 3. With this method, it was possible to switch from an unbalanced dataset (Fig. 1b, 9775 compounds) to a balanced dataset (Fig. 1c, 4640 compounds) while maintaining the same chemical space distribution for the subsampled class (Additional file, Figures S1-S2).

Applicability domain

In order to define model applicability domain, the most frequent structures in the entire final dataset are analyzed. Thus, an investigation was conducted on the structures from the perspective of the nature of the core metal and the binders. This analysis involved the generation of Murcko Scaffold and Murcko Skeleton. The Murcko Scaffold retains the core structural features of a molecule, including its specific atom types and bonds. In contrast, the Murcko Skeleton abstracts further, representing just the connectivity, thereby making it even more generalizable. The relative percentage of complexes in the dataset, based on the central metal, is shown in Fig. 2.

Subsequently, Murcko Scaffolds and Murcko Skeletons (Fig. 3) were constructed for each complex. Figure 3a illustrates the scaffolds that are most prevalent in the dataset.

A scaffold diversity of 65.7%, as calculated in accordance with the specifications outlined in Eq. 1, and a Gini index of 0.29 were determined for this group. In contrast, Fig. 3b illustrates the most prevalent Murko Skeletons, for which the skeleton diversity, as calculated in accordance with the specifications outlined in Eq. 2, is 57.4%, with a Gini index of 0.35.

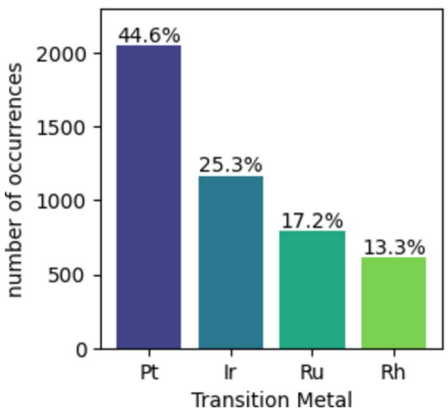


Fig. 2 Bar plot represents the percentage of compounds in the dataset divided according to the nature of the central atom

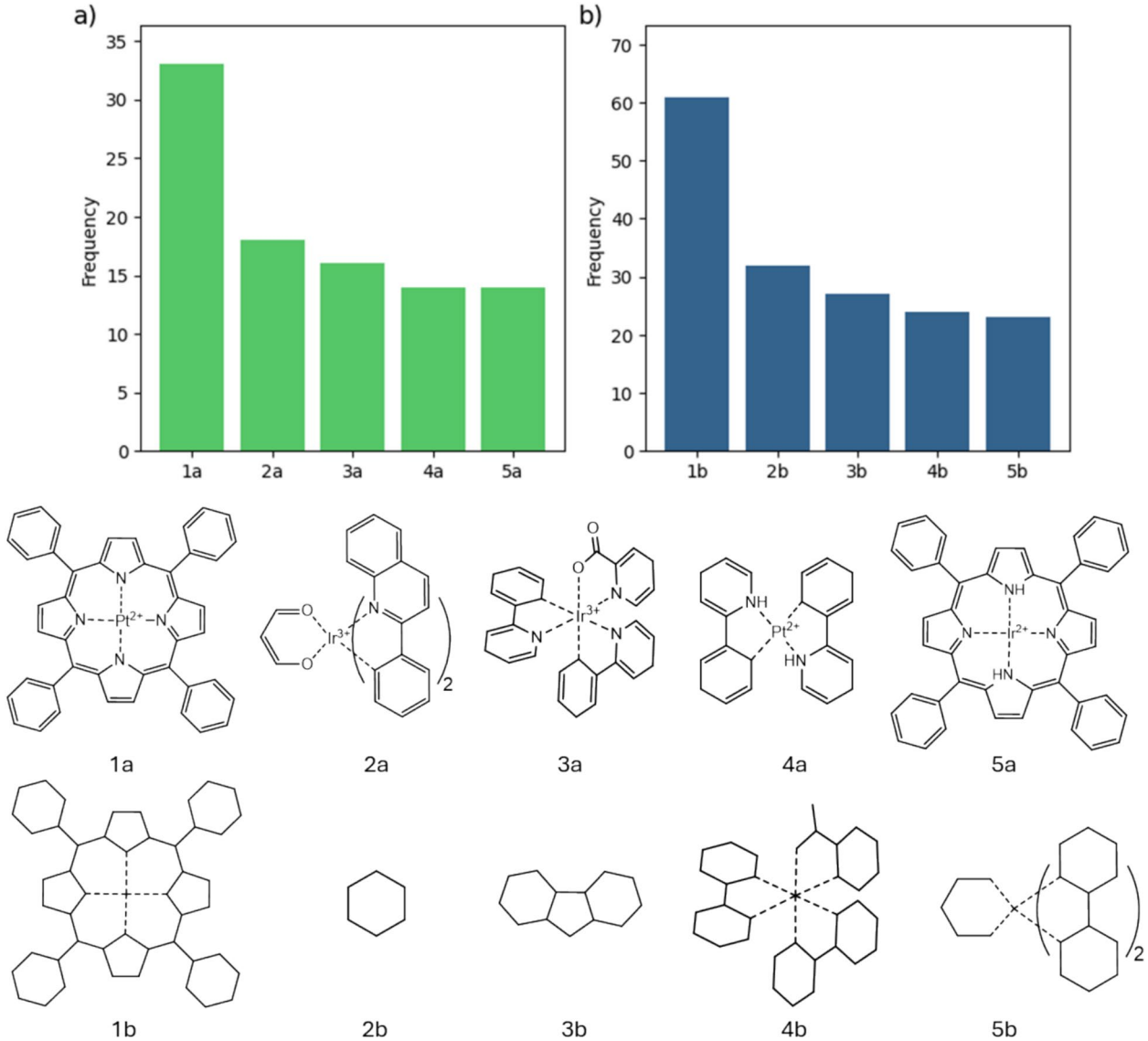


Fig. 3 **a** Murcko Scaffold frequency barplot with the five most frequent scaffold; **b** Murcko Skeleton frequency bar plot with the five most frequent skeletons

$$Scaffold\ diversity(\%) = \frac{Unique\ scaffolds}{Total\ N\ of\ scaffolds} \cdot 100\% \quad (1)$$

$$Skeleton\ diversity(\%) = \frac{Unique\ skeletons}{Total\ N\ of\ skeletons} \cdot 100\% \quad (2)$$

The diversity values and Gini index of the Scaffold and Skeleton are comparable. A diversity of approximately 50–60% is indicative of a favorable degree of structural variability. Data indicates that tetraphenyl porphyrins are the most common skeletons, which are divided between Pt (1a) and Ir (1e) complexes. In contrast, Skeletons 2b and 3b suggest a notable prevalence of complexes with relatively simple structures. The Gini index value of approximately 0.3 for both scaffolds and skeletons suggests that the dataset is relatively homogeneous, with each unique scaffold/skeleton appearing with a similar frequency as the others, and no structures being particularly prevalent. Chemical space was represented by applying PCA to the WPC descriptors calculated on unique scaffolds (Fig. 4). As mentioned above, the WPC descriptors were selected according to the highest variance explained by the first two PCs for better representation.

The applicability domain, generated on the basis of the unique scaffolds of the complexes in the dataset used, demonstrates a considerable breadth of coverage within

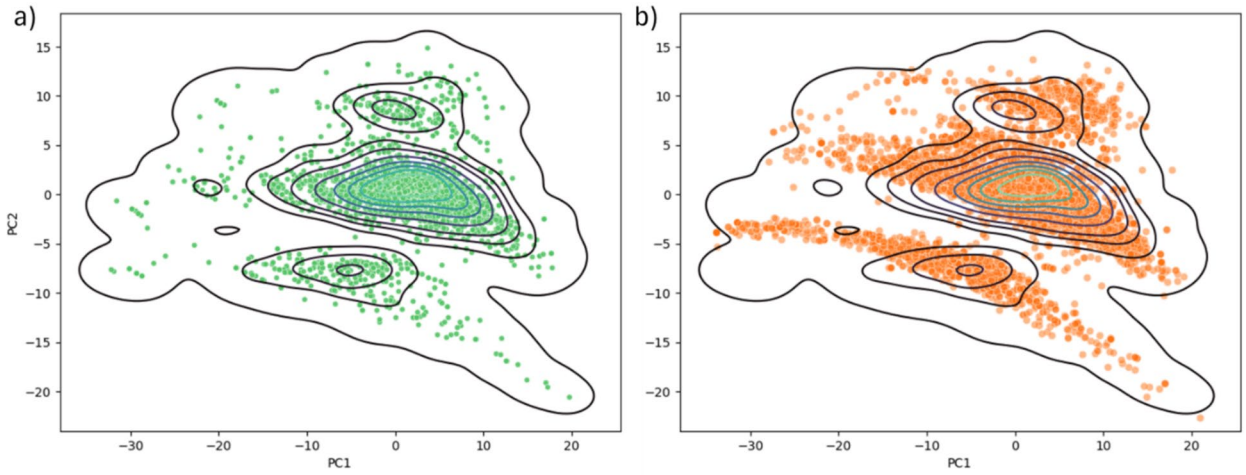


Fig. 4 **a** Chemical space of WPC descriptors calculations on unique molecular scaffolds. The density lines indicate the applicability domain of the classifier. **b** The applicability domain defined by the scaffolds is superimposed on the complete dataset of 9775 complexes showing that it covers a wide range of structures of Ru, Rh, Ir and Pt complexes

the generated chemical space. Furthermore, this result corroborates the hypothesis that undersampling the majority class results in the removal of complexes while maintaining the diversity of the dataset and its chemical space. The principal component analysis (PCA) of these descriptors revealed the emergence of three clusters. A more detailed examination demonstrated that the clusters are primarily distinguished by varying values of the SRW07 and TWC descriptors, which increase from the cluster on the bottom to the one on the top. Consequently, clusters comprise molecules of decreasing simplicity at the bottom, and those of increasing complexity at the top.

Feature engineering

The descriptors used to train the six types of selected classification models selected are listed in Table 1, together with their respective initial length. However, transformations were applied to reduce the length of all descriptors in order to enhance model performance by eliminating noise and selecting smaller, more informative feature sets. By selecting smaller sets of descriptors and reducing noise, improvements were achieved in the models in terms of performance and calculation time. The feature transformations involved several steps for removing constant features, quasi-constant features (with threshold = 0.1), duplicate features and highly correlated

Table 1 Descriptors used for the dataset with their initial and final dimensions after feature engineering

Name	Type	Initial length	Final length
ECFP1024_4	Fingerprint	1024	105
ECFP1024_6	Fingerprint	1024	128
ECFP2048_4	Fingerprint	2048	85
ECFP2048_6	Fingerprint	2048	93
ECFP4096_4	Fingerprint	4096	80
ECFP4096_6	Fingerprint	4096	82
CATS2D	Categorical	150	118
2DAtomPairs	Categorical	1596	387
FGroup	Categorical	153	35
MACCS	Fingerprint	167	86
CATS2D + 2DAtomPairs	Combined	1746	503
CATS2D + 2DAtomPairs + ECFP4096_4	Combined	5842	682
FGroup + CATS2D	Combined	304	151
FGroup + ECFP1024_6	Combined	1178	165
2DAtomPairs + ECFP4096_4	Combined	5693	549

features (with threshold ≥ 0.9). Table 1 summarizes the results of feature engineering applied to all generated descriptors.

Classification models

Based on fifteen types of descriptors summarized in Table 1 and the six types of classification models, a total of 90 classification models were explored. Each set of compounds was divided into a training set and a validation set with an 80:20 ratio (3712 training compounds and 928 external validation compounds). In addition, model performance validation metrics on the training set were calculated by further dividing the set according to the SkFoldCV methodology [29]. As a first approach, the models were trained with the default hyperparameters and using the S10FoldCV. The complete tables showing the results obtained for all possible model-descriptor combinations are given in Additional file, Table S1. Table 2 shows only the results for the AtomPairs2D (AP2D) descriptor, which were selected as best descriptors in terms of overall performance.

In general, several types of descriptors exhibit similar performance, particularly those that are combined. However, the AtomPairs2D descriptor was selected due to its exceptional performance, simplicity in interpretation, and minimal number of features. Although combined descriptors achieve the highest performance, they demonstrate comparable predictive capability to individual descriptors. This suggests that combining descriptors mainly increases the number of features, without substantially improving performance.

As can be seen from Table 2, the best results in terms of all the examined metrics are from the combinations RFC + AtomPairs2D (RFC_AP2D) and XGBC + AtomPairs2D (XGB_AP2D). Considering these two classifiers, it was determined that XGB_AP2D would be an optimal choice for further investigation. In fact, since all conditions were equal and performance was comparable, it was deemed preferable to utilize the XGBC model, which is considerably faster to train and apply than the RFC [35].

It is notable that other studies in the literature have identified the same target for other types of compounds [36, 37]. These studies include the construction of consensus models by combining the most effective models identified. Similarly, two models were constructed by combining the RFC and XGBC models, which were identified as the most effective for AP2D descriptors. The initial consensus model was constructed through the integration of the RFC and XGBC models employing the voting method (RFXGBC_vot). In contrast, the second model was built up by combining the same models with the stacking method, utilizing an LR as a meta-learner (RFXGBC_StackLR). Nevertheless, the results indicate that there is no significant enhancement in performance (Additional File, Table S2). Indeed, the outcomes yielded by the consensus models are indistinguishable from those obtained using the individual models.

In order to verify the classifier robustness, the training process was repeated with different k-values of the SkFoldCV. The results obtained are summarized in Table 3, in which it is observed that the values of all evaluation metrics do not vary significantly.

Evaluation metrics on validation set are in line with the values found for the training set for all split values of SkFoldCV (Table 3). From the results obtained, XGB_AP2D shows good generalization capabilities and performs well on data not provided during the training phase. To validate the effectiveness of the undersampling technique applied to the dataset, the model was also tested on the unbalanced dataset. The results are summarized in the Additional file, Table S3.

Although undersampling reduces the amount of available data for training, it helps balance the class distribution, allowing the model to focus more effectively on the minority class. In an imbalanced dataset, the minority class often receives insufficient representation during training. By removing a portion of the majority class samples, undersampling prevents the model from being overwhelmed by the majority class, improving its ability to distinguish between the two classes.

Table 2 Performance of the six classification models on the training of AtomPairs2D descriptors with S10FoldCV with mean and standard deviation

Descriptors	Model	A	SE	SP	P	F1
	LR	0.81 ± 0.01	0.79 ± 0.03	0.82 ± 0.03	0.79 ± 0.02	0.79 ± 0.02
AtomPairs2D	DTC	0.85 ± 0.01	0.85 ± 0.02	0.85 ± 0.02	0.83 ± 0.02	0.84 ± 0.02
AP2D	RFC	0.91 ± 0.01	0.89 ± 0.02	0.94 ± 0.01	0.92 ± 0.01	0.90 ± 0.01
118 features	SVC	0.69 ± 0.03	0.55 ± 0.05	0.81 ± 0.03	0.71 ± 0.04	0.62 ± 0.04
	GBC	0.85 ± 0.02	0.79 ± 0.02	0.90 ± 0.03	0.87 ± 0.03	0.83 ± 0.02
	XGBC	0.91 ± 0.01	0.90 ± 0.01	0.92 ± 0.02	0.91 ± 0.02	0.91 ± 0.01

Table 3 Performance of XGB_AP2D classifier with different SkFoldCV split values on train set with standard deviation across folds and the same metrics evaluated with an external validation set

Internal validation	A	SE	SP	P	F1
S5FoldCV	0.91 ± 0.01	0.90 ± 0.02	0.91 ± 0.01	0.90 ± 0.01	0.90 ± 0.01
S10FoldCV	0.91 ± 0.01	0.90 ± 0.02	0.92 ± 0.02	0.93 ± 0.02	0.91 ± 0.02
S15FoldCV	0.91 ± 0.02	0.90 ± 0.03	0.92 ± 0.02	0.91 ± 0.03	0.91 ± 0.03
External validation	0.91	0.88	0.93	0.91	0.90

Furthermore, undersampling can mitigate the risk of overfitting to a particularly dominant class, which could otherwise skew the model's learning. Without the oversaturation of the majority class, the model is more likely to learn more general and discriminative features, improving its generalization capacity, even on an external validation set.

The results obtained in this study suggest that, for this case, undersampling had a positive impact on the model's performance. Key metrics such as precision, recall, F1-score, and accuracy all improved significantly. This approach helped to reduce the effects of class imbalance, enhancing the model's ability to recognize both the majority and minority classes. These findings demonstrate how balancing techniques, such as undersampling, can optimize the performance of machine learning models when dealing with imbalanced datasets.

Model optimization and evaluation

The selected classifier was optimized by focusing on highly informative features and optimizing hyperparameters. The relative importance was assigned to each feature, based on the PFI method [38]. Subsequently, the features sorted by descending importance were phased out using the method known as RFE [39]. Given the obtained results, represented in Additional File, Figure S3, the model retains the same classification metrics using only the 53 most relevant features. When only a few features are considered, the risk of overfitting is minimized, the classifier is less computationally expensive, the hyperparameter optimization process becomes faster and, most importantly, results are more easily interpretable.

After selecting the XGB_AP2D model with 53 features as the best classifier, the optimization of the most important model hyperparameters was carried out to try to achieve an improvement in prediction performance. The optimization of six hyperparameters was then executed by random search in the hyperparameter space shown in Table 4.

Once the best hyperparameters combination was determined, the threshold value of the classification was optimized with steps of 0.1 in the range of values 0–1. At the end, a threshold value of 0.22 was found to be the best for the classification. The results for the evaluation of the XGB_AP2D model are shown in Table 5.

Table 4 Hyperparameter space explored with default values for XGBC and optimal values for XGB_AP2D

Hyperparameter	Range explored	Default value	Optimal value
n_estimator	100 to 500, step: 100	100	400
max_depth	3 to 15, step: 3	3	8
learning_rate	0.01, 0.05, 0.1, 0.2, 0.3	0.1	0.31
subsample	0.6 to 1.0, step: 0.1	1.0	1.0
colsample_bytree	0.6 to 1.0, step: 0.1	1	1
gamma	0 to 0.4, step: 0.1	0	0

Table 5 Evaluation metrics for the optimized XGB_AP2D classifier. The internal evaluation was calculated with S15FoldCV

evaluation metric	Internal evaluation	External evaluation
A	0.92	0.91
SE	0.90	0.90
SP	0.93	0.93
P	0.92	0.92
F1	0.91	0.90

As shown in Table 5, the final model XGB_AP2D correctly classifies a high percentage of the complexes. The evaluation metrics for all split values of the SkFoldCV are comparable to those obtained for the external evaluation set. However, after optimization, a deterioration in the evaluation metrics was observed. Despite this, the model demonstrates a low propensity to overfit and performs well with a relatively small number of features.

Finally, the confusion matrix and the ROC curve were generated (Fig. 5) on the results of the validation test. The ROC curve, together with AUC-ROC suggest that the classification model in question has excellent performance, being able to correctly distinguish between positive and negative classes with a high degree of accuracy.

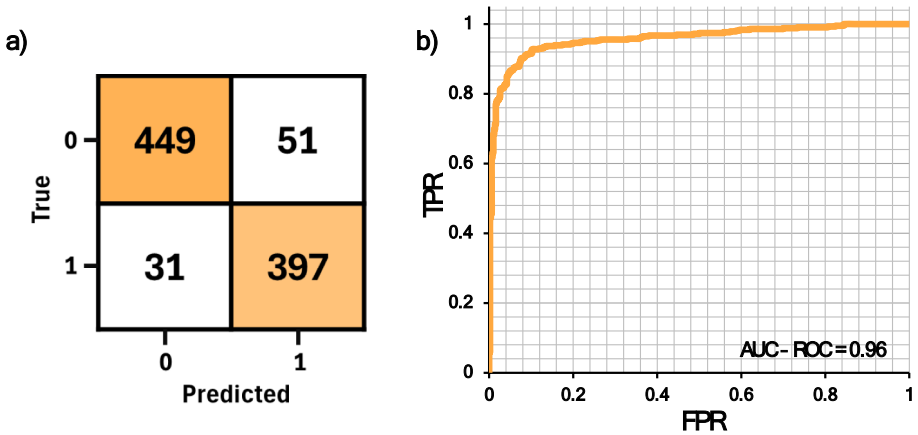


Fig. 5 **a** Confusion matrix on validation set prediction results; **b** ROC curve and AUC-ROC value for validation set prediction

Decision tree analysis

Since the XGBClassifier is an ensemble model, it consists of several decision trees trained in series to improve prediction performance. While XGBC_AP2D is a model capable of predicting the target with high precision and accuracy, the single decision tree can be used to interpret the obtained results.

Looking at Additional file, Table S1, one can see that prediction performance of the single decision tree (DTC) is lower than that of the XGBC_AP2D model, but it remains acceptably high. Furthermore, by applying the same feature cleaning protocol used for XGBC_AP2D, it appears that the performance of DTC remains almost unchanged by reducing the number of AP2D features to 26.

XGBC and DTC models' performance was evaluated both by applying and not applying the RFE method, the results obtained are shown in the Supporting file, Table S4.

The results show that reducing the number of features through feature selection does not lead to a significant change in the evaluation metrics of the models. Specifically, both the XGBC and DTC models achieved similar performance in terms of key metrics such as F1-score, accuracy, precision, and recall, with fewer features compared to the full set. For XGBC, performance with 53 selected features was comparable to that with 118 features, while for DTC, 26 features sufficed to achieve the same results as with all 118.

This suggests that many of the original features were either irrelevant or redundant, offering little to no added predictive value. The feature selection process, therefore, not only reduces the complexity of the models but also helps to improve interpretability, making the models easier to understand without sacrificing performance. Moreover, using fewer features reduces the risk of overfitting and enhances computational efficiency, as simpler models tend to be more stable and generalized.

Moreover, it was possible to extrapolate information on the structure of DTC, with a focus on the leaf nodes. This approach aims to stop not only at the features importance for prediction but to explore the single tree structure for a more in-depth analysis of its impact. This is just one example of how DTCs can be analyzed to extract key information about it and the dataset. More detailed information can be found in the reference repository. Analysis starts with the first information which is about feature importances listed in Table 6.

The feature assigned to have the highest importance is B01[C-X]. The transition metal is indicated by the letter X in the AP2D descriptor. Thus, this feature indicates the presence or absence of a direct bond between a carbon atom and the transition metal.

Looking at the tree separation rules, feature B01[C-X] results in the separation criterion present in node 780 generating the two leaves 781 and 782 as depicted in Fig. 6.

Table 6 List of the 10 most important features ordered by decreasing importance

Feature	Importance
B01(C-X)	0.031573
F02(C-N)	0.011746
F03(C-X)	0.010668
F02(C-X)	0.010237
F01(C-C)	0.008190
F05(C-X)	0.007974
F04(F-X)	0.006466
F03(N-S)	0.006034
F03(N-N)	0.005280
F03(C-N)	0.004957

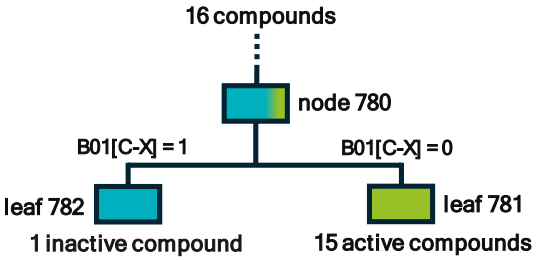


Fig. 6 Decision tree structure around node 780

It can be seen from Fig. 6 that node 780 contains 16 complexes of which 1 is inactive and 15 are active. Complexes that travel the same path in the decision tree and reach the same node together are complexes whose features obey the same rules. This means that all complexes arriving at node 780 are structurally similar enough to obey the same decision rules in the tree. However, the most important separation occurs at the last node, when the model separates the inactive complex from the active ones using the descriptor B01[C–X] as a criterion. This means that the inactive complex differs from the active ones because it contains the C–X bond. The structure of the inactive complex, named **inactPt**, is shown in Scheme 1.

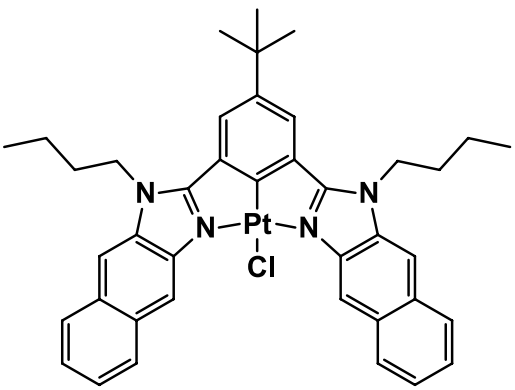
By searching back through the training dataset, the inactive complex turns out to be a cyclometalated platinum complex synthesized and characterized by Chan et al. [40] whose λ_{MAX} is 441 nm.

The presence of the direct bond between carbon and the metal, in this case platinum, is the reason why the model assigns this complex to the inactive class. Therefore, the C–Pt bond was replaced with a N–Pt one to see the effect caused on λ_{MAX}. The structural modification, however, was carried out without changing the rest of the structure, ensuring that the decision rules of all previous nodes remain equally respected by the new complex. Scheme 2 shows the modified complex, named **actPt**, according to the decision rules of the tree.

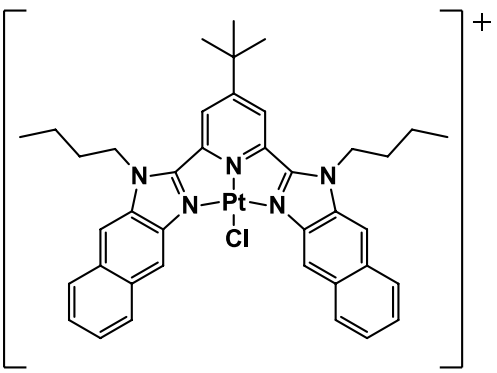
The new complex shown in Scheme 2 has an N–X bond instead of a C–X bond. The feature B01[C-X] of this compound is in fact equal to 0. This structural change makes the complex positively charged, but for the present analysis only the cationic portion of the complex was taken into account, neglecting the counter anion.

The XGBC_AP2D model was therefore used to predict whether this modified complex could absorb in the therapeutic window or not. In fact, according to the model, the newly modified complex absorbs in the therapeutic window with a high probability of assignment to the active class of 0.84. Further confirmation of the results was obtained by comparing the spectra of both complexes simulated using TD-DFT. The comparison is shown in Fig. 7. All molecular geometry optimizations have been carried out with the Gaussian16 software package [41], at the density functional level of theory, employing the B3LYP functional [42, 43]. For the Pt atom, the relativistic compact Stuttgart–Dresden effective core potential [44] has been used in conjunction with the corresponding split valence basis set. The standard double-ζ 6-31G* basis sets have been used for all the rest of the atoms. The electronic spectra were obtained, within the nonequilibrium time-dependent TDDFT approach, as vertical electronic excitation on the ground-state structure, by using the M06 [45] functional.

Such computational protocol has been selected on the basis of the outcomes of a preliminary benchmark study, considering the performance of several exchange–correlation functionals, in reproducing the



Scheme 1 Structure of the **inactPt** complex showing the presence of a direct C-Pt bond



Scheme 2 The structure of the new complex, **actPt**, obtained by structurally modifying the inactive complex, according to the rules of the decision tree

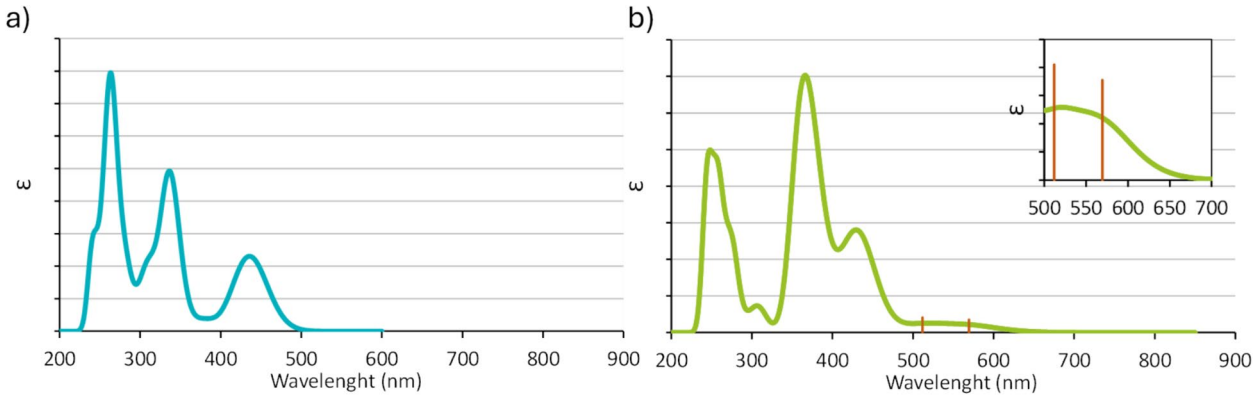


Fig. 7 TD-DFT theoretical UV–vis spectra simulated by using the M06 functional of **a inactPt** and **b actPt** complexes

absorption spectrum of the complex experimentally detected. Further details can be found in Additional File, Table S3, where the benchmark outcomes are reported for the best performing functionals.

Figure 7a shows the theoretical spectrum of the **inactPt** complex, that reproduces very well that experimentally detected reported in the literature (theoretical $\lambda_{MAX}=441$ nm, experimental $\lambda_{MAX}=441$ nm). On the other hand, the calculated spectrum of (Fig. 7b) the new structurally modified complex, actPt, shows an absorption band in a region of the spectrum inside the therapeutic window, as suggested by the DTC and XGBC_AP2D models, with two absorptions peaks at $\lambda_{MAX}=594.4$ nm and $\lambda=512$ nm. It is worth mentioning, however, that the model can only predict the presence of one or more transitions in the correct region of the spectrum, but not their intensity.

Conclusions

The absorption of light within the near-infrared spectral range known as the therapeutic window (600–850 nm), is essential for evaluating the potential of PDT and/or PACT agents. This study introduces an Extreme Boosting Classifier model trained via supervised learning to reliably predict whether Pt, Ir, Ru, and Rh metal complexes absorb electromagnetic radiation within this wavelength range of the therapeutic window. The investigated range, spanning from 500 to 850 nm, was chosen based on reported maximum absorption wavelengths of PDT and PACT relevant metal complexes. The lower limit was chosen on the basis of the values of the maximum absorption wavelength reported for many metal complexes considered suitable as PDT and PACT agents, while the choice of the highest extreme depends on the low efficiency of PDT photosensitizers activated beyond 800 nm in promoting molecular oxygen from the triplet to the singlet state. The dataset, albeit relatively small and imbalanced between active and inactive compounds regarding therapeutic window absorption, supports robust predictions without overfitting. The descriptors used count very few features that are crucial for an accurate classification, and are very easy to interpret and straightforwardly calculated with licensed or free software. Consequently, the model is highly efficient, fast and requires limited computational resources. Machine learning models, particularly decision tree-based algorithms, offer interpretable results, enabling insight into classification rules. This characteristic, coupled with the simplicity of the algorithms based on decision trees, allowed us to explore the nodes of a single tree to understand the rules learnt during the training step to correctly classify metal complexes into their respective classes. It also facilitates targeted structural modifications to transform initially inactive complexes into potentially active ones or, more generally, it allows to modify the structural characteristics of a base molecule resorting to results from the supervised approach.

One of the main limitations of the model is the narrow chemical space on which it is trained, so it may have insufficient information for a correct classification of complexes that are structurally very different from those in the training dataset. Moreover, the model does not quantify absorption peak intensities, which may be very low.

In summary, this study proposes a machine learning model capable of reliably predicting light absorption by Pt, Ir, Ru, and Rh complexes within the therapeutic window, guiding subsequent experimental and theoretical studies. Active complexes identified by using the approach proposed here require further investigation to meet additional criteria that compounds have to fulfill to be considered suitable PDT and PACT agents. Additionally, the study suggests a strategy for designing new promising complexes by introducing structural

modifications based on the features identified by the model during supervised training.

Accelerated ML discovery and design of transition-metal complexes with desired characteristics, in spite of additional challenges due to their peculiar properties, is rapidly progressing across a wide range of applications. However, even if ML models have been recently used to predict a wide array of UV–vis spectral properties of organic molecules, this approach is new for the category of metal complexes. The approach proposed here is the first step towards a complete, efficient and fast selection of compounds as PDT and PACT agents.

Abbreviations

PDT	Photodynamic therapy
PACT	Photoactivated chemotherapy
TD-DFT	Time dependent-density functional theory
ML	Machine learning
PS	Photosensitizer
PCA	Principal Component Analysis
PCs	Principal Components
SkFoldCV	Stratified K-Fold Cross Validation
S5FoldCV	Stratified 5 Fold Cross Validation
S10FoldCV	Stratified 10 Fold Cross Validation
S15FoldCV	Stratified 15 Fold Cross Validation
RGS	Random Grid Search
PFI	Permutation Features Importance
RFE	Recursive Features Elimination
A	Accuracy
SE	Sensitivity
SP	Specificity
P	Precision
F1	F1-score
AUC-ROC:	Area Under Curve-Receiver Operating Characteristic
FP	False positives
FN	False negatives
TP	True positives
TN	True negatives
TPR	True positive rate
FPR	False positive rate
LR	Logistic Regressor
DTC	Decision Tree Classifier
RFC	Random Forest Classifier
SVC	Support Vector Classifier
GBC	Gradient Boosting Classifier
XGBC	Extreme Boosting Classifier
ECFP	Extended Connectivity Fingerprints
CATS2D	2-Dimensional Chemically Advanced Template Search
AtomPairs	2-Dimensional Atom Pairs
FGroup	Functional Groups Count
MACCS	Molecular ACCess System Key
WPC	Walk and Path Counts

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13321-024-00939-5.

Additional file 1. Used Molecular Descriptor, used Evaluation Metrics, Majority Class Undersampling Technique, PCA plot with WPC descriptors before undersampling, PCA plot with WPC descriptors after undersampling, Benchmark, Recursive Features Elimination, TD-DFT Benchmark.

Acknowledgements
The authors acknowledge Fundação para a Ciência e a Tecnologia (FCT), the Portuguese Agency for Scientific Research for the financial support through projects UIDP/00313/2020. Tânia Cova acknowledges the Junior Researcher Grant CEECIND/00915/2018 assigned by FCT. This research has been supported by MUR, Autorità di Gestione PON "Ricerca e Innovazione" 2014–2020 (CCI 2014IT16M2OP005) and Università della Calabria.

Author contributions
V.V.: data collection, investigation, writing, review. T.F.: investigation, editing and review, A.P.: investigation, supervision and review. E.S.: writing, supervision and review. All the authors have read and approved the final manuscript.

Funding
Not applicable.

Availability of data and materials
Data is provided within the github repository at: https://github.com/vorsamaqoy/MetalComplexClassifier/tree/main.

Declarations

Competing interests
The authors declare no competing interests.

Received: 22 August 2024 Accepted: 8 December 2024
Published online: 05 January 2025

References

1. Siegel RL, Giaquinto AN, Jemal A (2024) Cancer statistics, 2024. CA Cancer J Clin 74:12–49. https://doi.org/10.3322/caac.21820
2. Altun İ, Sonkaya A (2018) The most common side effects experienced by patients were receiving first cycle of chemotherapy. Iran J Public Health 47:1218–1219
3. Eslami M, Memarsadeghi O, Davarpanah A et al (2024) Overcoming chemotherapy resistance in metastatic cancer: a comprehensive review. Biomedicines 12:183. https://doi.org/10.3390/biomedicines12010183
4. Debela DT, Muzazu SG, Heraro KD et al (2021) New approaches and procedures for cancer treatment: current perspectives. SAGE open Med 9:20503121211034370. https://doi.org/10.1177/20503121211034366
5. Wainwright M (1998) Photodynamic antimicrobial chemotherapy (PACT). J Antimicrob Chemother 42:13–28. https://doi.org/10.1093/jac/42.1.13
6. Anas A, Sobhanan J, Sulfiya KM et al (2021) Advances in photodynamic antimicrobial chemotherapy. J Photochem Photobiol C Photochem Rev 49:100452. https://doi.org/10.1016/j.jphotochemrev.2021.100452
7. Plaetzer K, Krammer B, Berlanda J et al (2009) Photophysics and photochemistry of photodynamic therapy: fundamental aspects. Lasers Med Sci 24:259–268. https://doi.org/10.1007/s10103-008-0539-1
8. Farrer NJ, Salassa L, Sadler PJ (2009) Photoactivated chemotherapy (PACT): the potential of excited-state d-block metals in medicine. Dalt Trans. https://doi.org/10.1039/b917753a
9. Bonnet S (2018) Why develop photoactivated chemotherapy? Dalt Trans 47:10330–10343. https://doi.org/10.1039/C8DT01585F
10. Qiao L, Liu J, Han Y et al (2021) Rational design of a lysosome-targeting and near-infrared absorbing Ru (ii)–BODIPY conjugate for photodynamic therapy. Chem Commun 57(14):1790–1793. https://doi.org/10.1039/D0CC06926D
11. Bonnet S (2015) Shifting the light activation of metallodrugs to the red and near-infrared region in anticancer phototherapy. Comments Inorg Chem 35:179–213. https://doi.org/10.1080/02603594.2014.979286
12. Adamo C, Jacquemin D (2013) The calculations of excited-state properties with time-dependent density functional theory. Chem Soc Rev 42:845–856. https://doi.org/10.1039/C2CS35394F
13. Barretta P, Mazzone G (2023) Mechanism of action of an Ir (iii) complex bearing a boronic acid active as a H_2O_2-responsive photosensitizer: ROS generation and quinone methide release for GSH scavenging. Inorg Chem Front 10(12):3686–3698. https://doi.org/10.1039/D3QI00203A
14. Barretta P, Scoditti S, Belletto D et al (2024) Ruthenium complexes bearing nile red chromophore and one of its derivative: theoretical evaluation

of PDT-related properties. J Comput Chem. https://doi.org/10.1002/jcc.27392
15. Kulik HJ (2020) Making machine learning a useful tool in the accelerated discovery of transition metal complexes. WIREs Comput Mol Sci. https://doi.org/10.1002/wcms.1439
16. Nandy A, Duan C, Taylor MG et al (2021) Computational discovery of transition-metal complexes: from high-throughput screening to machine learning. Chem Rev 121:9927–10000. https://doi.org/10.1021/acs.chemrev.1c00347
17. Vigna V, Cova TFGG, Nunes SCC et al (2024) Machine learning-based prediction of reduction potentials for Pt IV complexes. J Chem Inf Model 64:3733–3743. https://doi.org/10.1021/acs.jcim.4c00315
18. Shao J, Liu Y, Yan J et al (2022) Prediction of maximum absorption wavelength using deep neural networks. J Chem Inf Model 62:1368–1375. https://doi.org/10.1021/acs.jcim.1c01449
19. Mamede R, Pereira F, Aires-de-Sousa J (2021) Machine learning prediction of UV–Vis spectra features of organic compounds related to photoreactive potential. Sci Rep 11:23720. https://doi.org/10.1038/s41598-021-03070-9
20. McNaughton AD, Joshi RP, Knutson CR et al (2023) Machine learning models for predicting molecular UV–Vis spectra with quantum mechanical properties. J Chem Inf Model 63:1462–1471. https://doi.org/10.1021/acs.jcim.2c01662
21. Bemis GW, Murcko MA (1996) The properties of known drugs. 1. Molecular frameworks. J Med Chem 39:2887–2893. https://doi.org/10.1021/jm9602928
22. Weininger D (1988) SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. J Chem Inf Comput Sci 28:31–36. https://doi.org/10.1021/ci00057a005
23. Mauri A (2020) alvaDesc: a tool to calculate and analyze molecular descriptors and fingerprints. Methods Pharmacol Toxicol. 801–820. https://doi.org/10.1007/978-1-0716-0150-1_32
24. Sperandei S (2014) Understanding logistic regression analysis. Biochem Medica. https://doi.org/10.11613/BM.2014.003
25. Rokach L, Maimon O. Decision Trees. In: Data Mining and Knowledge Discovery Handbook. Springer-Verlag, New York, pp 165–192
26. Breiman L (2001) Random Forest. Mach Learn 45:5–32
27. Evgeniou T, Pontil M (2001) Support vector machines: theory and applications. pp 249–257
28. Bentéjac C, Csörgő A, Martínez-Muñoz G (2021) A comparative analysis of gradient boosting algorithms. Artif Intell Rev 54:1937–1967. https://doi.org/10.1007/s10462-020-09896-5
29. Pedregosa F, Varoquaux G, Gramfort A et al (2011) Scikit-learn: machine learning in Python. J Mach Learn Res 12:2825–2830
30. Chen T, Guestrin C (2016) XGBoost. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, New York, NY, USA, pp 785–794
31. Rácz A, Bajusz D, Héberger K (2021) Effect of dataset size and train/test split ratios in QSAR/QSPR multiclass classification. Molecules 26:1111. https://doi.org/10.3390/molecules26041111
32. Browne MW (2000) Cross-validation methods. J Math Psychol 44:108–132. https://doi.org/10.1006/jmps.1999.1279
33. Vujovic ŽĐ (2021) Classification model evaluation metrics. Int J Adv Comput Sci Appl. https://doi.org/10.14569/IJACSA.2021.0120670
34. Zakharov AV, Peach ML, Sitzmann M, Nicklaus MC (2014) QSAR modeling of imbalanced high-throughput screening data in PubChem. J Chem Inf Model 54:705–712. https://doi.org/10.1021/ci400737s
35. Cha G-W, Moon H-J, Kim Y-C (2021) Comparison of random forest and gradient boosting machine models for predicting demolition waste based on small datasets and categorical variables. Int J Environ Res Public Health 18:8530. https://doi.org/10.3390/ijerph18168530
36. Ksenofontov AA, Lukanov MM, Bocharov PS et al (2022) Deep neural network model for highly accurate prediction of BODIPYs absorption. Spectrochim Acta Part A Mol Biomol Spectrosc 267:120577. https://doi.org/10.1016/j.saa.2021.120577
37. Rusanov AI, Dmitrieva OA, Mamardashvili NZ, Tetko IV (2022) More is not always better: local models provide accurate predictions of spectral properties of porphyrins. Int J Mol Sci 23:1201. https://doi.org/10.3390/ijms23031201
38. Kaneko H (2022) Cross-validated permutation feature importance considering correlation between features. Anal Sci Adv 3:278–287. https://doi.org/10.1002/ansa.202200018
39. Priyatno AM, Widiyaningtyas T (2024) A systematic literature review: recursive feature elimination algorithms. JITK Jurnal Ilmu Pengetah dan Teknol Komputer. 9:196–207. https://doi.org/10.33480/jitk.v9i2.5015
40. Chan AK, Lam ES, Tam AY et al (2013) Synthesis and characterization of luminescent cyclometalated platinum(II) complexes of 1,3-Bis-heteroazolylbenzenes with tunable color for applications in organic light-emitting devices through extension of π conjugation by variation of the heteroatom. Chem A Eur J 19:13910–13924. https://doi.org/10.1002/chem.201301586
41. Frisch MJ, Trucks GW, Schlegel HB, et al (2016) G16_C01. Gaussian 16, Revision C.01, Gaussian, Inc., Wallin
42. Becke AD (1998) Density-functional thermochemistry. III. The role of exact exchange. J Chem Phys 98:5648. https://doi.org/10.1063/1.464913
43. Lee C, Yang W, Parr RG (1988) Development of the Colic-Salvetti correlation-energy formula into a functional of the electron density. Phys Rev B 37:785
44. Andrae D, Häußermann U, Dolg M et al (1990) Energy-adjusted ab initio pseudopotentials for the second and third row transition elements. Theor Chim acta 772(77):123–141. https://doi.org/10.1007/BF01114537
45. Zhao Y, Truhlar DG, Zhao Y, Truhlar DG (2008) The M06 suite of density functionals for main group thermochemistry, thermochemical kinetics, noncovalent interactions, excited states, and transition elements: two new functionals and systematic testing of four M06-class functionals and 12 other function. Theor Chem Acc 120:215–241. https://doi.org/10.1007/s00214-007-0310-x

Publisher's Note