AGAPE (comput**A**tional **G**-quadruplex **A**ffinitiy **P**r**E**diction): The first AI In-silico workflow for G-quadruplex binding affinity prediction.

AUTHORS

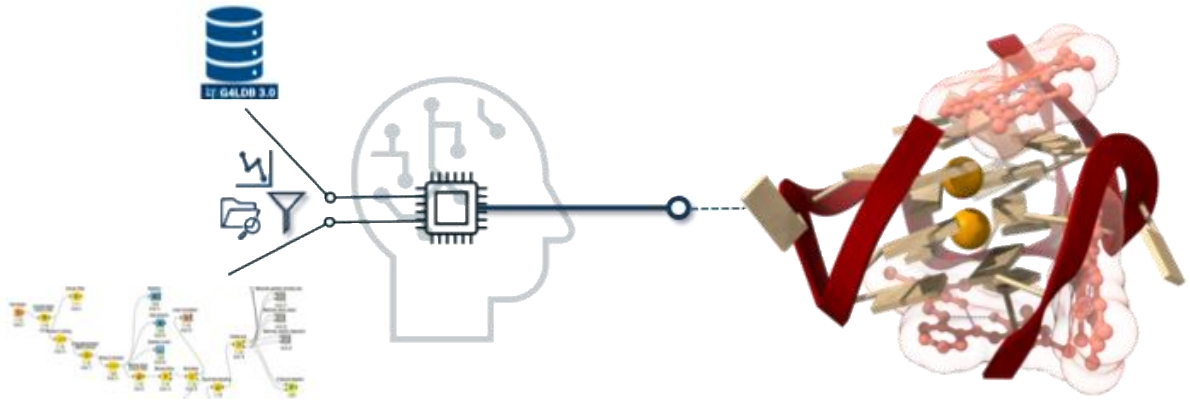
Luisa D'anna, [*1] Ugo Perricone, [*2] Giada De Simone,[2] Antonio Monari,[3] Giampaolo Barone,[1] Alessio Terenzi[1]

[1] Department of Biological, Chemical and Pharmaceutical Sciences, University of Palermo, Viale delle Scienze, Ed. 17, 90128, Palermo, Italy. E-mail: luisa.danna@unipa.it

[2] Fondazione Ri.MED, Corso Calatafimi 414, 90100, Palermo, Italy. E-mail: uperricone@fondazionerimed.com

[3] Université Paris Cité and CNRS, ITODYS, F-75006, Paris France

GRAPHICAL ABSTRACT



ABSTRACT

AGAPE (computAtional G-quadruplex Affinitiy PrEdiction) is an innovative machine learning (ML)-based workflow developed to predict the potential stabilization of small molecules by G-quadruplex (G4) structures. G4s, especially prevalent in telomeres and oncogene promoters, represent promising therapeutic targets, yet designing selective binders remains challenging. This study, thus, addresses this gap by implementing an ML framework in KNIME, specifically designed for ease of use across the scientific community. The **AGAPE** workflow integrates 5666 classical and quantum chemical (QC) descriptors for G4 ligands, enabling a comprehensive representation of binding-relevant molecular features. Using data from G4LDB and in-house collections, we created a robust dataset of 1217 compounds categorized by Förster Resonance Energy Transfer (FRET) assays as ACTIVE or INACTIVE based on their G4 stabilization capacity. Feature selection algorithms and ML models, particularly XGBoost with Naïve Bayesian and Random Forest classifiers, were employed to achieve an optimized prediction model with an accuracy close to 90%. A consensus voting system among top-performing models further

improved classification reliability. **AGAPE** efficiently predicts G4 stabilization, offers interpretability of key chemical interactions, and provides a scalable, accessible tool for G4-focused drug discovery. This workflow lays the foundation for targeted therapeutic development, enhancing ligand selectivity, and expanding the application of AI in chemoinformatics

INTRODUCTION

Guanine quadruplexes (G4s) are nucleic acid structures that have garnered increasing attention in recent years due to their significant roles in diverse biological processes. G4 structures originate within Guanine-rich sequences, in both DNA and RNA, when four guanines interact via Hoogsteen hydrogen bonds, forming a planar tetrad. Stacking multiple tetrads results in a stable arrangement, further stabilized by a monovalent cation (usually Na^+ or K^+) occupying the central channel and compensating for the negative density due to the presence of the external phosphate groups (Figure 1)(1). Despite their rigid core, G4s exhibit structural polymorphism, with variations in strand orientation, number of stacked tetrads, and loop features such as their position, length, and type. Additionally, the different allowed orientations of the glycosidic bond lead to the coexistence of parallel, antiparallel, and hybrid topologies, whose relative stability is also influenced by loop flexibility (2, 3).

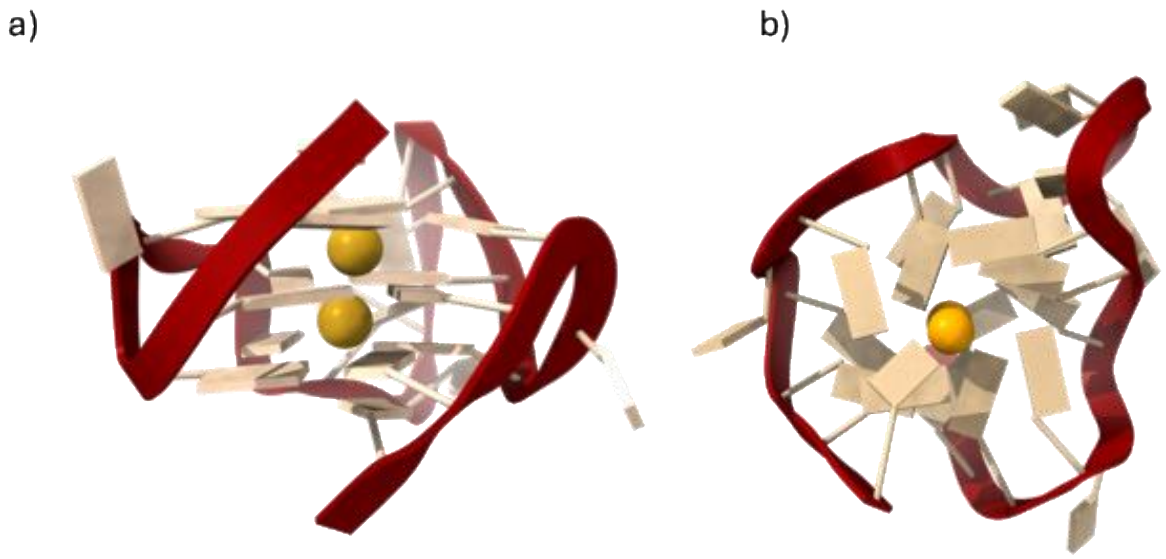


Figure 1. Side (a) and top (b) view of a typical parallel G-quadruplex motif.

The biological relevance of these motifs is connected to their location in highly conserved sequences and in significant position within the eukaryotic genome, such as telomeres and oncogene promoters, where they play crucial roles in regulating gene expression, replication, and chromosomal stability (4, 5). The alteration in the G4 landscape structure and stability has been implicated in the pathogenesis of various disease, such as cancer and neurodegenerative disorders (6).

The design of small molecules interacting with, and stabilizing the G4 secondary structure typically involves extended flat aromatic polycyclic moieties, which promote stacking with the external G-quartets, along with positively charged substituents to enhance electrostatic interactions with the nucleic acid backbone. Additional structural features, potentially leading to further interactions with the loops, grooves, phosphate backbone, and extruded nucleic acids bases, should also be considered in developing new active compounds (3). So far, different organic moieties as well as metal complexes have been synthesized and have demonstrated significant binding affinity to DNA/RNA G4s (7–9). Salphen metal complexes, in particular, have attracted significant interest due to their strong and selective binding to G4 structures over canonical double helical nucleic acids. Thus, many transition metal complexes involving salphen-like ligands have been investigated to elucidate the different role of the central metal ion and lateral substituents in optimizing the interactions with G4 structures (10–13).

Despite their promising potential, the clinical and even preclinical development of G4-targeting therapies remains constrained by the challenge in identifying molecules that can selectively recognize G4 structures. Many known G4 binders also show affinity for double-helical DNA, besides lacking specificity for particular G4 sequences. An ideal binding molecule should be able of select specific G4 structures implicated in disease progression, thereby modulating only the biological process or signalling pathways in which the structures are involved.

To address these challenges, a deeper understanding of the intricate structural relationships between G4 structures and potential small-molecule drug candidates is essential. The development of selective G4 binders and stabilizers is clearly a complex task, often tackled using tailored combination of computational and experimental approaches. These often include time consuming multiscale modelling, such as all-atom molecular dynamics

simulations, and physical-chemical measurements, including calorimetry and spectroscopic titration.

Conversely, the integration of artificial intelligence (AI) techniques in biomedical research is becoming increasingly prominent, particularly in rationalizing and enhancing the discovery and development of new therapeutics(13–15). Machine Learning (ML) enables the efficient analysis of large chemical and biological datasets, facilitating the identification of potential drug candidates and predicting complex molecular interactions(17, 18). ML approaches are nowadays ripe enough to offer quantitative models predicting biological activities, chemical-physical and toxicological properties, or algorithms to predict protein folding and dynamics (18–21). The success of deep learning approaches in predicting protein structure has been widely recognized, as exemplified by the awarding of the 2024 Nobel Prize in Chemistry for the advances in computational protein design, and particularly the AlphaFold model. Alongside these groundbreaking achievements, numerous scientific studies have explored various methodologies, including classification and linear regression applications, to further enhance design strategies. AI-based approaches could also have great potential in the design of specific G4 binders. Indeed, by applying AI-driven methods, it is possible to effectively explore and design ligands that selectively target G4s, thereby advancing therapeutic strategies aimed at modulating these distinctive and functionally important structures.
Extensive literature research indicates that the application of AI in designing small molecules as G4 stabilizers has not been explored yet, and our investigation found no evidence of classification models being used to predict the potential stabilization properties of small molecules towards G4 conformations. Although, some AI tools are available, they are mainly focused on the prediction of G4 structures and folding, starting from a specific sequence. These models are based on the evaluation of the G4 conformation ratio and the relative stability of the structure (22–26), while they overlook the affinity of ligands and the potential stabilization induced by external binders.

In this work, we present the first ligand-based ML framework to predict the potential stabilization of G4 by small molecules. Beyond prediction, our goal is to rationalize and identify the key chemical features that correlate with effective G4 binding and stabilization. Explainable AI approaches are particularly valuable in chemoinformatics, offering insights into the mechanisms behind chemical interactions; however, they are not always straightforward to

apply (27). Although ML approaches are promising, the current dataset on G4 binders is relatively limited compared to other biologically relevant targets, such as kinase inhibitors(28, 29). This relative scarcity of data and specialized tools underscores the need for collecting and organizing information that can contribute to building an initial framework. Such a tool could guide scientists in identifying chemical features that may enhance G4 binding and stabilization.

Our approach involves utilizing molecular descriptors (30, 31) as molecular embeddings to build a binary classification model (ACTIVE/INACTIVE) trained on several annotated small molecules. The use of molecular descriptors in ML techniques proves to be highly valuable across various fields(32–34), as it facilitates the development of more interpretable models compared to utilizing fingerprints for molecular structures representation (32,34,35). Considering the complex molecular and electronic structure of some G4 binders, such as metal complexes, we decided to include Quantum Chemical (QC) properties as additional feature to characterise the different molecules, i.e. QC molecular descriptors. The advantage in using QC properties for ML has been recently demonstrated and the combination of QC descriptors (37–40) and ML has been widely used to predict various physical-chemical properties as well as molecular reactivity(41), regio-selectivity of substitution (42), or for chemical space exploration (43). Furthermore, the combination of QC and ML is also promising in medicinal chemistry for drug design and leads discovery and optimization (44). However, this approach is not yet a standard common practice in drug design, probably because of the heavy calculation overload needed.

Machine learning (ML) approaches generally demand advanced coding skills, which can limit accessibility for scientists who may lack experience with programming tools for model development. To address this, we developed **AGAPE**, the first comprehensive ML workflow for G4 binding prediction, fully implemented in KNIME—an open-source, user-friendly block programming platform suitable even for non-programmers(45). The **AGAPE** workflow is openly accessible and designed for ease of use by the entire scientific community. It guides users through all key ML steps, starting from raw data, including data cleaning and preparation, feature selection algorithms, ML model development, and a final voting system. Users can apply this workflow to their own datasets, for which molecular descriptors have already been calculated, to implement the ML approach.

We also show that our model is able to efficiently predict the stabilization capacity of diverse G4 binders against G4, while providing explainability of the results in terms of chemical properties. Hence, it offers a reliable and useful tool for the development of the next generation, selective, binders.

MATERIAL AND METHODS

Data collection and dataset creation

Data were mainly retrieved from the public database G4LDB (v. 2.2)(46), a collection aimed to explore ligands targeting all kinds of G-quadruplexes (DNA and RNA), as well as from compounds from literature not included in G4LDB, as well as from our in-house compounds library. To classify molecules in our dataset as ACTIVE or INACTIVE, we utilized results from Förster Resonance Energy Transfer (FRET) melting assays, using thresholds derived from benchmark compounds in the G4LDB database and supporting literature data (47). Specifically, we designated molecules with $\Delta T \leq 8°C$ as "INACTIVE" and those with $\Delta T \geq 15°C$ as "ACTIVE" (46, 48, 49). Although a ΔT of approximately 15°C is generally accepted in the literature as an indicator of activity, we chose a lower threshold to more reliably identify INACTIVE compounds and minimize the risk of false positives. The QC geometry optimization of the selected molecules was performed using Jaguar tool from Schrödinger suite (Schrödinger Release 2023-2: Jaguar, Schrödinger, LLC, New York, NY, 2024)(53). Density Functional Theory (DFT) calculations, performed with B3LYP as exchange correlation functional and LANLD2Z basis set, was used to optimize the molecular geometry and calculate relevant QC properties. QC is, indeed, particularly useful to obtain the optimized geometry of small molecules and especially transition metal complexes. This is particularly true in our case since the parameterization of classical force fields for transition metal complexes is not straightforward and force field parameters for some of the metals in our dataset are either missing or not standardized. In addition, resorting to QC approach allows to explicitly include features directly related to the electronic structure of the ligand obtained from an *ab initio* approach, i.e. without relying on any empirical assumptions. Thus, on top of the QC-optimized structure of each ligand, we calculated QC properties and retain them as additional features for our model creation. The selected QC descriptors were mainly focused on the determination of point charges

distribution, polarizability, electrostatic potentials, and surface size. From the QC-optimized structures we also calculated classical molecular descriptors using alvaDesc Software(50, 51).

Data cleaning and preparation

Dataset has been processed into 3 consecutive steps, namely *data cleaning, data transformation and dimensionality reduction*.

Data Cleaning: The dataset was organized into columns representing individual variables. Columns with constant values, entirely missing values, or more than 20% missing values were removed to improve data quality.

Data Transformation: Missing values in numeric columns containing integer have been replaced with the median value, while those in columns containing double-precision floating data points were substituted with the mean value. Furthermore, the dataset was normalized using a linear transformation, scaling all the entries such that the minimum and maximum values in each column fell within the interval [0;1].

Dimensionality Reduction: We calculated the Pearson correlation coefficient (52) for each pair of columns, as a measure of the correlation between the two variables. Variables with a high degree of correlation (threshold set at 0.9) were identified and subsequently removed, allowing to significantly reduce redundancy in the dataset.

After these three steps of cleaning and preparation, the residual dataset consisted of 1723 features, which were then used as descriptors for building our ML model.

Feature selection algorithms

As it is customary in ML approach, we tested different Feature Selection (FS) algorithms to extract the most significant descriptors to classify our molecules as ACTIVE and INACTIVE, with respect to G4 stabilization. Such selection methods have then been tested using performance metrics, and notably by comparing Forward (FFS) and Backward feature selection (BFS). When comparing the two approaches, BFS, although more computationally extensive, did not demonstrate significantly superior performance (best accuracy 0.85 with 1720 features). Even though BFS is usually preferred for achieving high accuracy, in our models, FFS outperformed BFS, yielding better results with fewer features selected. Particularly, we compared four

different types of FFS each one coupled with different ML models: Decision Tree, Naïve Bayesian, Random Forest and Boruta algorithms (53–55). For each feature selection method, but Boruta, four different configurations were set to filter maximum significant features: 1:5, 1:10, square root (sqrt) of total dataset entries (N) and finally 1:5 of ACTIVE compounds within the dataset.

The output of the feature selection model has been then submitted to ML predictor and then to the score calculation for the final evaluation. While all the methods used were compared, only the best performing feature selection algorithm has been chosen for further ML model training.

ML models adopted (Training-test and validation subset creation)

Before testing different ML models combined with the feature selection algorithms previously described, our dataset was divided into Training set (80% of the whole dataset), validation set (10% of the whole dataset) and test set (10% of the whole dataset). The partitioning was done using shuffling data and randomly selecting the partitions. In order to strengthen the testing part of our model, we also considered an internal dataset of 28 molecules, synthesized in our laboratory, as a secondary test set. To maximise the prediction accuracy, different ML approaches have been compared; namely Logistic regression, SVM, XGBoost, Gradient Boost Tree and Tree ensemble RF(57–59). In Supplementary Data (Table S1) we report all the performance parameters retrieved for the best 9 model. As already proved by our previous experience and from literature (60–62), XGboost demonstrated to be the most accurate model to correctly assign binary classes to our dataset. The XGboost settings were: **Regularisation** (Eta=0.3; Lambda=1; Alpha=1; Gamma and Maximum delta step=0) **Tree construction** (Booster Tree with max depth =6 and minimum child weight=1) **Tree method** (Auto with sketch epsilon =0.03 and scale positive weight =1) Grow policy (Depth Wise with maximum number of leaves =0 and maximum number of bins =256).

Voting System

Based on the results obtained with all the different combination of FS and ML methods, we decided to adopt a voting system as a final score to classify molecules as ACTIVE/INACTIVE. This system was used on our internal test set of 28 molecules. In detail, we selected the top

three combination of feature selection (i.e. Naive with 165 features, Random Forest with 35 features and Boruta) coupled with XGBoost ML model. The final score is given by the majority of the results, i.e collecting at least two votes out of three.

RESULTS

Data Collection

The latest version of G4LDB includes 3695 G4 ligands and 32142 activity entries in total. From the database, we downloaded Interaction and Stabilisation activities with a focus on data related to FRET melting assay. The FRET melting assay is a widely used technique to measure the melting temperature (T_m) of G4 structures, thus retrieving thermal stability and the stabilization induced by small molecules. In this assay the G4 sequence is labelled at the 5′ and 3′ ends with two fluorescent dyes: a donor and an acceptor. Upon excitation of the donor unit, if the acceptor is in close proximity energy transfer may occur, leading to the transfer of the excitation to the acceptor, and hence quenching the donor fluorescence. FRET being highly dependent from the distance between donor and acceptor, it may be considered as a most suitable tool to assess folding and unfolding. At lower temperatures, G4 is in its folded state, keeping the donor and acceptor dyes close together, which allows for FRET to occur. As the temperature increases, the G4 unfolds, causing the dyes to move farther apart, reducing FRET and, hence, favouring the donor fluorescence quantum yield. When testing the effect of a stabilizing compound, the same FRET assay is repeated in the presence of the compound. If the ligand stabilizes the G4, it will increase the T_m, meaning that the structure will remain intact at a higher temperature. This shift in melting temperature (ΔT_m) is, thus, a direct measure of the stabilization of the G4 induced by a specific ligand. The choice of this technique was also driven by consistency with our in-house dataset, which includes stabilization data obtained using FRET assay on various G4-DNA sequences. After filtering duplicates from G4LDB, we obtained 5320 unique activity entries and 1835 unique ligands. On top of these entries, another filtering step was applied to evaluate if T_m data were consistent for molecules presenting multiple activity data. Considering that our research group focuses on metal complexes, especially of salphen based ligands, known for their important G4 stabilisation activity (9, 10) we enriched the G4LDB collection with other annotated salphen-like complexes from literature (13, 56–58). Finally, according with our threshold, we collected a unique

ensemble of 1217 compounds divided into 490 ACTIVE and 727 INACTIVE entries. Of these molecules, 1073 were purely organic compounds while 144 were metals complexes. After QC geometry optimization and properties calculations, we obtained a final dataset composed of 1188 molecules.

Descriptors calculation

For each molecule 5666 different type of molecular descriptors, divided into 33 different classes (Topological indices, Connectivity indices, Geometrical descriptors, 3D autocorrelations, Functional group counts, Charge descriptors, Molecular properties, Drug-like indices, etc), (50) have been calculated using alvaDesc descriptors (see https://www.alvascience.com/alvadesc-descriptors/). A preliminary dimensionality reduction step was performing removing constant values, values with standard deviation inferior to 0.0001, and columns with all missing values. After this filtering 4285 features from alvaDesc have been conserved. Furthermore, trough DFT relevant QC properties and descriptors such as Surfaces (electrostatic potential, average local ionisation energy, electron densities), Atomic Electrostatic Potential Charges (Charge + dipole moment), Mulliken population, Multiple moments, Polarizability and Hyperpolarizability have been calculated.

Data cleaning and preparation (Training-test and validation subset creation)

After Cleaning, Transformation and dimensionality reduction, described into Material and Methods section, the final dataset comprised 1217 compounds and 1723 descriptors used as features for Machine learning models building. After the threshold application, we obtained a dataset of 490 ACTIVE ($\Delta T \geq 15°C$) and 727 INACTIVE molecules ($\Delta T \leq 8°C$). After ligand preparation, this numbers were slightly reduced to 476 ACTIVE and 712 INACTIVE binders.

Feature selection algorithms

To filter the maximum significant features for classification, Decision Tree, Naive and Random Forest feature selection methods were submitted to four different configurations as described in the Methods section (*Feature selection algorithms*) retrieving different number of crucial

features (Table 1). Notably, the Boruta algorithm retrieved 304 features as best performing combination.

	Decision Tree	Näive	Random Forest
1:5 total dataset	238	165	78
1:10 total dataset	119	98	62
sqrt total dataset	35	35	34
1:5 number of ACTIVE	95	95	62

Table 1. *Number of features retrieved for each method and filtering parameters.*

ML models adopted

Analysing all the performance for each combination of FS and ML models, we selected XGBoost as the best-performing algorithm coupled with Naive, Random Forest and Boruta. Indeed, decision Tree FS methods showed poor results when compared with the other mentioned approaches, and have been, thus, discarded. As shown in Figure 2, when comparing the results of the main tested models combined with various feature selection methods, XGBoost clearly outperformed the others, obtaining a nearly 90% accuracy when coupled to Random Forest.

Features Selection Methods

Machine Learning Methods

	Naïve	Random Forest	Boruta
XGBoost	85.7%	89.9%	87.8%
Gradient Boost	83.2%	87.8%	88.2%
Tree Ensemble	85.7%	87.8%	88.2%

Figure2. Overall accuracy of the intersection between ML methods and FS Methods. XGBoost methods outperforms as mean overall accuracy we compared with other methods.

However, based on these results, and the data reported in Table S1 (SI), it is evident that the best approaches to classify molecules as ACTIVE/INACTIVE towards G4 conformations are the three combinations of XGBoost Model (with Naive, Random Forest and Boruta). Therefore,

these three different combinations were further used in a consensus protocol through a voting system.

	Overall Accuracy	Overall Error	Cohen's Kappa(k)	Correctly classified molecules	Incorrectly classified molecules
BORUTA	88.66%	12.19%	0.742	105	14
Random Forest	89.92%	10.08%	0.787	107	12
Näive	85.71%	14.29%	0.695	102	17

Table2. Overall statistics of XGBoost models according to different feature selection methods. Values are a synthesis of validation and test metrics.

Given the imbalance in the dataset, we evaluated all model performance parameters, as reported in SI Table S4, while also considering the consistency of predictions between the test and validation sets.

Voting System

In order to evaluate the generalization capability of our workflow and to strengthen the classification process, we decided to focus on a voting system based on a consensus vote between the three best performing methods. To support this strategy, we evaluated the voting system on an in-house dataset of 28 molecules that had not been included in the model dataset (Figures 3-4). The 28 molecules included 25 transition metal compounds and 3 organic moieties, which corresponded to the free ligands used in the synthesis of the metal compounds themselves.

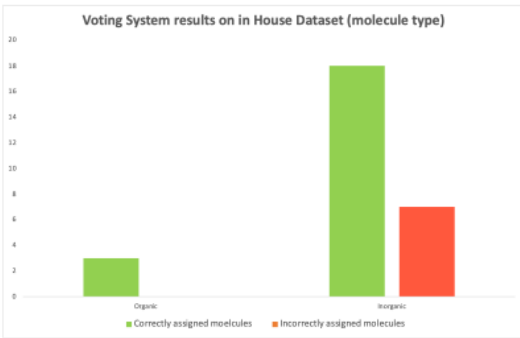
In Figure 3 we report the results after processing the 28 molecules with our voting system. We may see that our model succeeds in correctly assigning 75% of the molecules, with only 7 errors. It is also evident that the voting system is slightly improving the accuracy of the single model, although the combination of XBoost with Boruta model is already significantly high. Considering that the model was completely blind with respect to the 28 molecules used, this represents, in our opinion, an acceptable result. Interestingly, after voting, all the molecules classified as INACTIVE have been correctly identified. This implies that the model is robust enough against false negative and thus should not lead to disregarding potentially ACTIVE compounds.

The importance of the voting system may also be recognized from the example reported in Figure 4. Compound Cpd6 was correctly classified as INACTIVE upon consideration of the majority, but not the unanimity, of the votes. Interestingly, this compound is indeed experimentally correlated to a rather modest increase in the T_m of G4 by about 6 °C, i.e. not far from the threshold. Conversely, Cpd26 ligand, which is unanimously predicted as ACTIVE, shows an experimental stabilization of the G4 by almost 30 °C.

a)

Compounds	XGBOOST_Naive_238	XGBOOST_BORUTA	XGBOOST_RF35	Class Assignment (Voting System)
cpd1	INACTIVE	ACTIVE	ACTIVE	ACTIVE
cpd2	INACTIVE	ACTIVE	ACTIVE	ACTIVE
cpd3	INACTIVE	ACTIVE	INACTIVE	INACTIVE
cpd4	INACTIVE	INACTIVE	ACTIVE	INACTIVE
cpd5	INACTIVE	INACTIVE	ACTIVE	INACTIVE
cpd6	INACTIVE	INACTIVE	ACTIVE	INACTIVE
cpd7	INACTIVE	INACTIVE	ACTIVE	INACTIVE
cpd8	INACTIVE	INACTIVE	INACTIVE	INACTIVE
cpd9	INACTIVE	INACTIVE	INACTIVE	INACTIVE
cpd10	INACTIVE	INACTIVE	ACTIVE	INACTIVE
cpd11	INACTIVE	ACTIVE	ACTIVE	ACTIVE
cpd12	ACTIVE	ACTIVE	ACTIVE	ACTIVE
cpd13	INACTIVE	ACTIVE	ACTIVE	ACTIVE
cpd14	INACTIVE	INACTIVE	ACTIVE	INACTIVE
cpd15	ACTIVE	INACTIVE	ACTIVE	ACTIVE
cpd16	INACTIVE	INACTIVE	ACTIVE	INACTIVE
cpd17	INACTIVE	INACTIVE	INACTIVE	INACTIVE
cpd18	INACTIVE	ACTIVE	ACTIVE	ACTIVE
cpd19	INACTIVE	ACTIVE	ACTIVE	ACTIVE
cpd20	ACTIVE	ACTIVE	ACTIVE	ACTIVE
cpd21	ACTIVE	ACTIVE	ACTIVE	ACTIVE
cpd22	ACTIVE	ACTIVE	ACTIVE	ACTIVE
cpd23	ACTIVE	ACTIVE	ACTIVE	ACTIVE
cpd24	ACTIVE	ACTIVE	ACTIVE	ACTIVE
cpd25	ACTIVE	INACTIVE	ACTIVE	ACTIVE
cpd26	ACTIVE	ACTIVE	ACTIVE	ACTIVE
cpd27	ACTIVE	INACTIVE	ACTIVE	ACTIVE
cpd28	ACTIVE	ACTIVE	ACTIVE	ACTIVE
Overall Accuracy	0,68	0,75	0,50	0,75

b)



c)

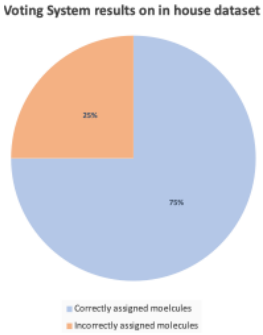


Figure 3. a) Results for each component used in the voting system model and final classification (1 = correct; 0 = incorrect). b) Results of voting system differentiated by molecules type (organic or inorganic). c) Percentage of correctly and incorrectly assigned molecules by voting system.

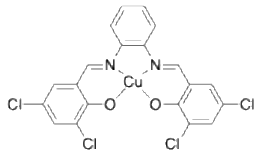
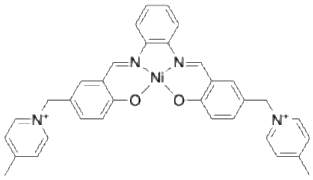
Structure	Compound	Exprimental ΔT (°C)	Class assignment
	Cpd6	5.9	INACTIVE
	Cpd26	28.98	ACTIVE

Figure 4. The experimental ΔT (°C) values represent the average results from multiple FRET assays conducted on three selected G4 structures. The comparison between the experimental data and the classifications generated by the ML model (using a voting system) reveals a high degree of correspondence, confirming the model's predictive accuracy.

Explicability

As for any Drug Design tool, a ML model predicting activity or property classification without explicability lacks useful information, which can be used to rationalise the observed results and for the design and optimisation of new compounds. For this reason, we decided to explore which molecular descriptors were determinants for the classification, identifying those that played a key role in the 9 classification models used in the initial workflow (see SI Table S1*)*. In Figure 5 we report the highly influent descriptors involved in ACTIVE/INACTIVE classification of our dataset and the frequency of appearance out of the nine FS and ML combination. In SI (Table S4) the explanation for each of the most influent descriptors found by our model is fully described. Finally, to compare the difference in the distribution of the descriptors for ACTIVE/ INACTIVE compounds, we evaluated the density distribution (see Table S2). These data confirm that the behaviour of few descriptors is crucial for the correct classification of ligands. Indeed, if on the one hand some descriptors such as CMBL or MAT6e do not present important differences in their values distribution for the two classes, others, and especially the QC descriptors, LDI, SHED_AP, nN, and max_conj_path have shown pronounced differences between ACTIVE and INACTIVE compounds. Such a result indicates that, even if a single descriptor alone may not discriminate between the two classes, a limited number of descriptors may define a relevant subset, which is capable to currently infer the G4 stabilization capacity. In this respect it appears that the combination of constitutional, pharmacophore, and electronic properties allows the algorithm to capture important differences in the ligands, and thus to predict their capacity as G4 binders and stabilizers.

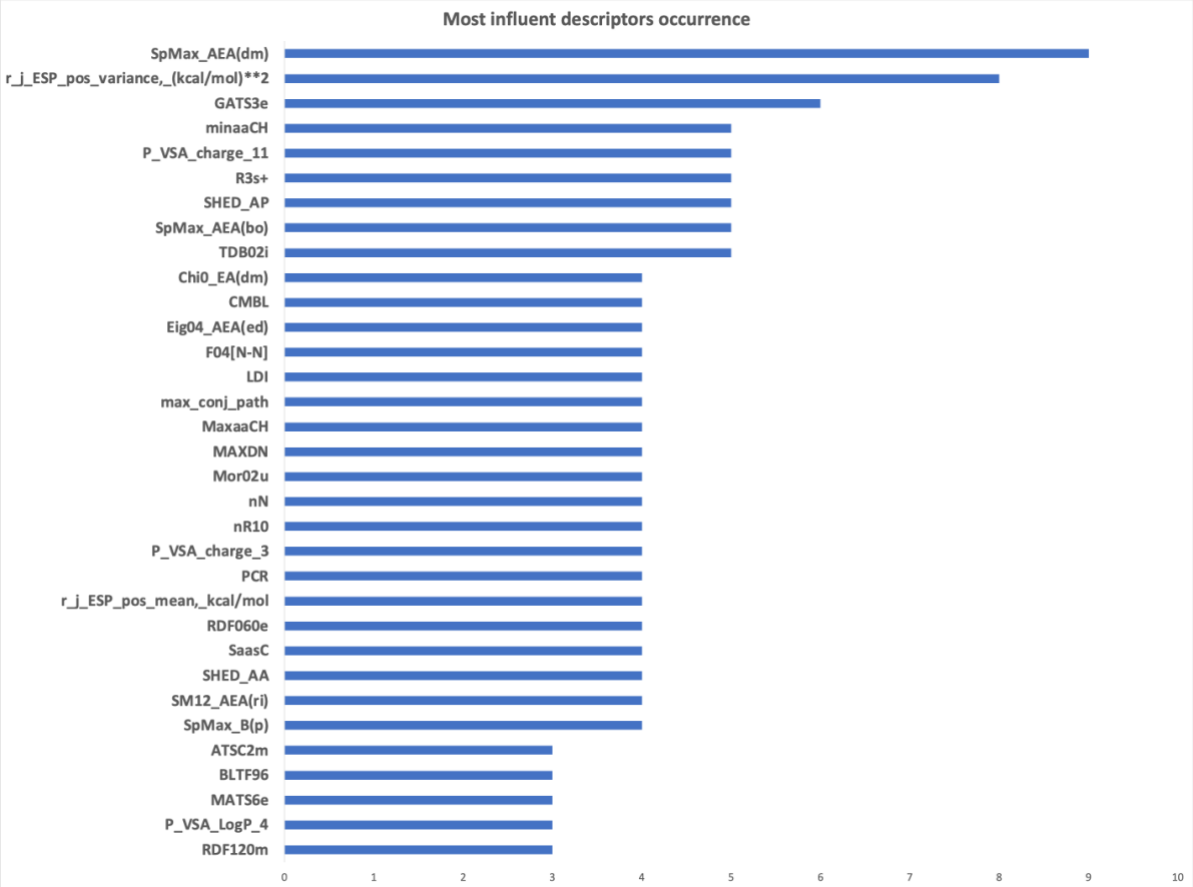


Figure 5. Most important (most frequent) descriptors identified by the best performing models for the classification in ACTIVE/INACTIVE ordered based on the frequency (left).

In SI, we report the Molecular descriptors retrieved for the three-voting system algorithm (Table S3) while the description and descriptor families for the main features retrieved by the best classification models are listed in Table S4.

DISCUSSION

In this work we present **AGAPE** (computAtional G-quadruplex Affinitiy PrEdiction). **AGAPE** is, to our knowledge, the first *in silico* workflow created to infer the affinity prediction of G4 ligands and their stabilization capacity. The workflow is based on a ML approach based on supervised classification exploiting molecular descriptors and QC-derived properties of small molecules to predict their stabilization of G4 structures. To build our models, a dataset of literature compounds composed by 476 ACTIVE and 712 INACTIVE entries was created to train FS/ML models to predict the G4 stabilization capacity and identify the chemical features behind their action. Prior to the final model creation, different feature selection and

classification models were assessed, including different forward and backward propagation feature selection. To this respect, the forward approach outperformed the others in terms of accuracy and overall performance. When analysing FS methods coupled with ML models for classification, we found that XGBoost was the best performing classification method when coupled with Boruta, Näive, and Random Forest feature selection approaches. In order to evaluate the most crucial descriptors for the correct classification of ACTIVE and INACTIVE ligands, we analysed the descriptors that were retrieved by the best combination of FS and ML for the classification. From our analysis, 33 descriptors (31 classical descriptors and 2 QC features) were critical for the correct classification into the two classes. Interestingly, the best performing FS/ML couple (XGBoost/Boruta) is based on all the 33 descriptors retrieved as fundamental for a correct classification. When analysing the molecular descriptor families, we noticed that they are mainly based on charge distribution, electrostatic potential and polarizability of the ligands. Such a characteristic is totally in line with the physical processes behind the G4 stabilisation process by small molecules, which generally occurs through π-π stacking interactions with the G-tetrads or extruded bases, which are mainly driven by dispersion and van der Waals interactions, and, thus, are highly dependent on polarizability. This hypothesis is also confirmed by other descriptors related to the pharmacophore features of the small molecules and the number and type of N atoms within the molecular scaffold. Furthermore, G4 binding is also favoured by electrostatic interaction with the negatively charged nucleic acid backbone, which is strongly influenced by the electrostatic potential of the ligands and their estimated point charges. Altogether, these observations concur in confirming that charge distribution and polarizability are crucial in influencing the ability of small molecules to favourably interact with the quadruplex and induce its stabilization. As a matter of fact, it has already been pointed out that favourable electronic polarizability (59) together with a well-balanced molecular charge could be important for small molecules to adapt to the complex electronic environment of G4, especially in the case of transition metal complexes (60) resulting in a better stabilisation of the host-guest complex.

The possible use of a cheaper model involving only the most important descriptors can also be envisaged in the future, especially for preliminary assessment and classification of large libraries of compounds.

As a final model to classify ligands as ACTIVE or INACTIVE, we adopted a voting system based on the three best performing models (XGBoost coupled wih Boruta, Näive and Random Forest). In order to evaluate the robustness and generalizability of our approach and furtherly test the model prediction capability, we have tested the voting system on an in-house library of 28 compounds (9, 10) that were not used for training, test, and validation. When used to predict G4 stabilisation property for the in-house dataset, our model performed well with an overall accuracy of 0.75. This is a significant result, especially considering that the dataset was mainly composed of inorganic compounds, which represented the minority class in the training, test, and validation sets, which were instead mainly composed by organic compounds. Indeed, when the model system examines an organic compound (see cpd7, 8 and 9 of our in-house dataset) the prediction is always unanimously correct except for compound cpd7, for which the less accurate model fails to provide a correct prediction.

Overall, we have described the properties and the performance of **AGAPE,** the first *in silico* workflow to assess stabilisation property of G4 stabilisers. In addition to evaluating the model's performance and versatility, this work has provided valuable insights into the factors influencing the G4 stabilization capacity. This led to the identification of key molecular descriptors that facilitate the classification of a ligand stabilization ability. This work paves the way for important advancements in the ML-driven design of G4 stabilizers.

In the future, we plan to refine our model by incorporating more tailored descriptors and a more balanced training set to enhance the predictive capability. Extending the model to predict selectivity toward G4 over canonical nucleic acid structures, as well as selectivity across different G4 topologies or sequences, is particularly appealing and will be also considered in the future. Finally, we plan to integrate AGAPE with generative AI frameworks to suggest structure modifications in existing G4 stabilisers or entirely new scaffold to be synthesized. In this context **AGAPE** will be used to virtually label large datasets as ACTIVE or INACTIVE for G4 stabilisations. Such virtually labelled datasets will then be used to train generative AI or large language model (LLM) to suggest structural modifications, thus mitigating the relative scarcity of available experimental data on these systems.

DATA AVAILABILITY

Here we will insert the link to the KNIME repository with the whole ML workflow and the simplified workflow to be directly adopted by users "https://github.com/Ugomolinf/AGAPE" (Figure 6).

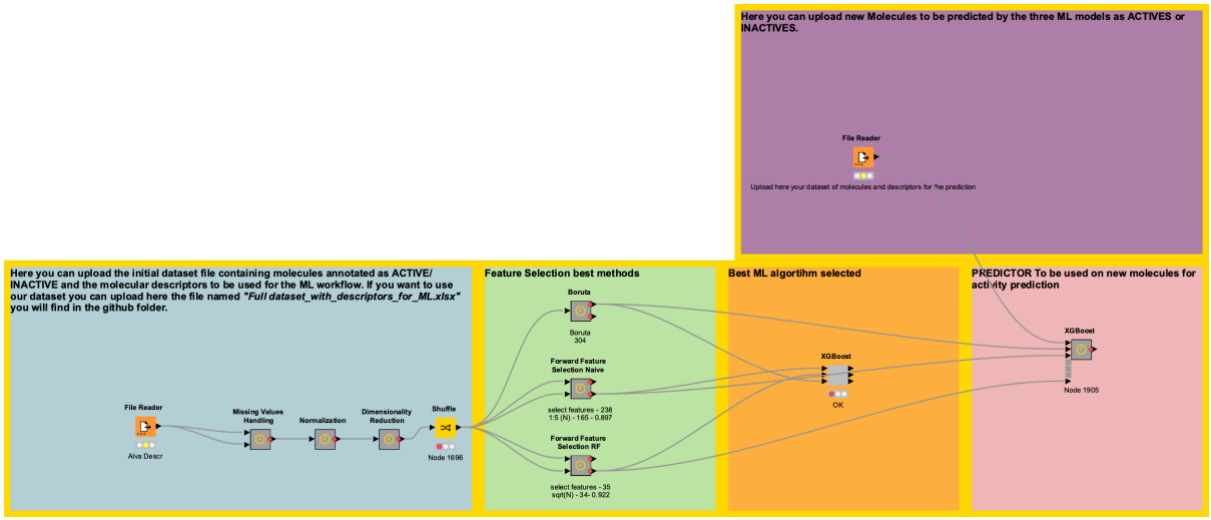


Figure 6. ML simplified workflow available on GitHub at the following link: "https://github.com/Ugomolinf/AGAPE"

SUPPLEMENTARY DATA

Supplementary Data are available in S.D. file.

AUTHOR CONTRIBUTIONS

Luisa D'Anna and Ugo Perricone : conceptualization, dataset creation, ML methodology, Explicabilityformal analysis, visualization, writing – original draft; Giada De Simone: dataset creation, ML methodology; Antonio Monari: conceptualization, writing – original draft, formal analysis, visualization; Giampaolo Barone: conceptualization, writing – original draft, formal analysis, visualization; Alessio Terenzi: conceptualization, writing – original draft, formal analysis, visualization.

ACKNOWLEDGEMENTS

ANR and CGI (Commissariat à l'Investissement d'Avenir) for their financial support of this work through Labex SEAM (Science and Engineering for Advanced Materials and devices), ANR 11 LABX 086 and ANR 11 IDEX 05 02. The support of the IdEx "Université Paris 2019" ANR-18-IDEX-0001 and of the Platform P3MB is gratefully acknowledged.

FUNDING

CONFLICT OF INTEREST

There are no conflicts to declare

REFERENCES

1. Varshney,D., Spiegel,J., Zyner,K., Tannahill,D. and Balasubramanian,S. (2020) The regulation and functions of DNA and RNA G-quadruplexes. Nat Rev Mol Cell Biol, 21, 459–474.
2. Spiegel,J., Adhikari,S. and Balasubramanian,S. (2020) The Structure and Function of DNA G-Quadruplexes. Trends in Chemistry, 2, 123–136.
3. Georgiades,S.N., Abd Karim,N.H., Suntharalingam,K. and Vilar,R. (2010) Interaction of Metal Complexes with G‐Quadruplex DNA. Angew Chem Int Ed, 49, 4020–4034.
4. Tian,T., Chen,Y.-Q., Wang,S.-R. and Zhou,X. (2018) G-Quadruplex: A Regulator of Gene Expression and Its Chemical Targeting. Chem, 4, 1314–1344.
5. Kosiol,N., Juranek,S., Brossart,P., Heine,A. and Paeschke,K. (2021) G-quadruplexes: a promising target for cancer therapy. Mol Cancer, 20, 40.
6. Wang,E., Thombre,R., Shah,Y., Latanich,R. and Wang,J. (2021) G-Quadruplexes as pathogenic drivers in neurodegenerative disorders. Nucleic Acids Res, 49, 4816–4830.
7. Liu,L.-Y., Ma,T.-Z., Zeng,Y.-L., Liu,W. and Mao,Z.-W. (2022) Structural Basis of Pyridostatin and Its Derivatives Specifically Binding to G-Quadruplexes. J Am Chem Soc, 144, 11878–11887.
8. Neidle,S. (2016) Quadruplex Nucleic Acids as Novel Therapeutic Targets. J. Med. Chem., 59, 5987–6011.
9. D'Anna,L., Rubino,S., Pipitone,C., Serio,G., Gentile,C., Palumbo Piccionello,A., Giannici,F., Barone,G. and Terenzi,A. (2023) Salphen metal complexes as potential anticancer agents: interaction profile and selectivity studies toward the three G-quadruplex units in the KIT promoter. Dalton Trans., 10.1039/D2DT03229E.
10. Froux,A., D'Anna,L., Rainot,A., Neybecker,C., Spinello,A., Bonsignore,R., Rouget,R., Harlé,G., Terenzi,A., Monari,A., et al. (2024) Metal centers and aromatic moieties in Schiff base complexes: impact on G-quadruplex stabilization and oncogene downregulation. Inorg. Chem. Front., 11, 5725–5740.
11. Farine,G., Migliore,C., Terenzi,A., Lo Celso,F., Santoro,A., Bruno,G., Bonsignore,R. and Barone,G. (2021) On the G‐Quadruplex Binding of a New Class of Nickel(II), Copper(II), and Zinc(II) Salphen‐Like Complexes. Eur. J. Inorg. Chem., 2021, 1332–1336.
12. Terenzi,A., Bonsignore,R., Spinello,A., Gentile,C., Martorana,A., Ducani,C., Högberg,B., Almerico,A.M., Lauria,A. and Barone,G. (2014) Selective G-quadruplex stabilizers: Schiff-base metal complexes with anticancer activity. RSC Adv., 4, 33245–33256.
13. Bonsignore,R., Russo,F., Terenzi,A., Spinello,A., Lauria,A., Gennaro,G., Almerico,A.M.,

Keppler,B.K. and Barone,G. (2018) The interaction of Schiff Base complexes of nickel(II) and zinc(II) with duplex and G-quadruplex DNA. Journal of Inorganic Biochemistry, 178, 106–114.
14. Jiménez-Luna,J., Grisoni,F., Weskamp,N. and Schneider,G. (2021) Artificial intelligence in drug discovery: recent advances and future perspectives. Expert Opin Drug Discov, 16, 949–959.
15. Vamathevan,J., Clark,D., Czodrowski,P., Dunham,I., Ferran,E., Lee,G., Li,B., Madabhushi,A., Shah,P., Spitzer,M., et al. (2019) Applications of machine learning in drug discovery and development. Nat Rev Drug Discov, 18, 463–477.
16. Paul,D., Sanap,G., Shenoy,S., Kalyane,D., Kalia,K. and Tekade,R.K. (2021) Artificial intelligence in drug discovery and development. Drug Discov Today, 26, 80–93.
17. Isert,C., Atz,K. and Schneider,G. (2023) Structure-based drug design with geometric deep learning. Current Opinion in Structural Biology, 79, 102548.
18. Azevedo,P.H.R.D.A., Peçanha,B.R.D.B., Flores-Junior,L.A.P., Alves,T.F., Dias,L.R.S., Muri,E.M.F. and Lima,C.H.D.S. (2024) In silico drug repurposing by combining machine learning classification model and molecular dynamics to identify a potential OGT inhibitor. Journal of Biomolecular Structure and Dynamics, 42, 1417–1428.
19. Noé,F., De Fabritiis,G. and Clementi,C. (2020) Machine learning for protein folding and dynamics. Curr Opin Struct Biol, 60, 77–84.
20. Xia,S., Chen,E. and Zhang,Y. (2023) Integrated Molecular Modeling and Machine Learning for Drug Design. J. Chem. Theory Comput., 19, 7478–7495.
21. Gong,Y., Teng,D., Wang,Y., Gu,Y., Wu,Z., Li,W., Tang,Y. and Liu,G. (2022) In silico prediction of potential drug-induced nephrotoxicity with machine learning methods. J Appl Toxicol, 42, 1639–1650.
22. Rossi,F. and Paiardini,A. (2022) A Machine Learning Perspective on DNA and RNA G-quadruplexes. CBIO, 17, 305–309.
23. Cagirici,H.B., Budak,H. and Sen,T.Z. (2022) G4Boost: a machine learning-based tool for quadruplex identification and stability prediction. BMC Bioinformatics, 23, 240.
24. Teng,F.-Y., Jiang,Z.-Z., Guo,M., Tan,X.-Z., Chen,F., Xi,X.-G. and Xu,Y. (2021) G-quadruplex DNA: a novel target for drug design. Cell Mol Life Sci, 78, 6557–6583.
25. Barshai,M., Engel,B., Haim,I. and Orenstein,Y. (2023) G4mismatch: Deep neural networks to predict G-quadruplex propensity based on G4-seq data. PLoS Comput Biol, 19, e1010948.
26. Yang,B., Guneri,D., Yu,H., Wright,E.P., Chen,W., Waller,Z.A.E. and Ding,Y. (2024) Prediction of DNA i-motifs via machine learning. Nucleic Acids Res, 52, 2188–2197.
27. Gallegos,M., Vassilev-Galindo,V., Poltavsky,I., Martín Pendás,Á. and Tkatchenko,A. (2024) Explainable chemical artificial intelligence from accurate machine learning of real-space chemical descriptors. Nat Commun, 15, 4345.
28. Kanev,G.K., de Graaf,C., Westerman,B.A., de Esch,I.J.P. and Kooistra,A.J. (2021) KLIFS: an overhaul after the first 5 years of supporting kinase research. Nucleic Acids Research, 49, D562–D569.
29. De Simone,G., Sardina,D.S., Gulotta,M.R. and Perricone,U. (2022) KUALA: a machine learning-driven framework for kinase inhibitors repositioning. Sci Rep, 12, 17877.
30. Todeschini,R. and Consonni,V. (2009) Molecular Descriptors for Chemoinformatics 1st ed. Wiley.
31. Grisoni,F., Ballabio,D., Todeschini,R. and Consonni,V. (2018) Molecular Descriptors for Structure-Activity Applications: A Hands-On Approach. Methods Mol Biol, 1800, 3–53.
32. Amar,Y., Schweidtmann,A.M., Deutsch,P., Cao,L. and Lapkin,A. (2019) Machine learning and molecular descriptors enable rational solvent selection in asymmetric catalysis. Chem.

Sci., 10, 6697–6706.
33. Baptista,D., Correia,J., Pereira,B. and Rocha,M. (2022) Evaluating molecular representations in machine learning models for drug response prediction and interpretability. J Integr Bioinform, 19, 20220006.
34. Kabylda,A., Vassilev-Galindo,V., Chmiela,S., Poltavsky,I. and Tkatchenko,A. (2023) Efficient interatomic descriptors for accurate machine learning force fields of extended molecules. Nat Commun, 14, 3562.
35. Orosz,Á., Héberger,K. and Rácz,A. (2022) Comparison of Descriptor- and Fingerprint Sets in Machine Learning Models for ADME-Tox Targets. Front. Chem., 10, 852893.
36. Zagidullin,B., Wang,Z., Guan,Y., Pitkänen,E. and Tang,J. (2021) Comparative analysis of molecular fingerprints in prediction of drug combination effects. Brief Bioinform, 22, bbab291.
37. Boobier,S., Hose,D.R.J., Blacker,A.J. and Nguyen,B.N. (2020) Machine learning with physicochemical relationships: solubility prediction in organic solvents and water. Nat Commun, 11, 5753.
38. Shimakawa,H., Kumada,A. and Sato,M. (2024) Extrapolative prediction of small-data molecular property using quantum mechanics-assisted machine learning. npj Comput Mater, 10, 11.
39. McNaughton,A.D., Joshi,R.P., Knutson,C.R., Fnu,A., Luebke,K.J., Malerich,J.P., Madrid,P.B. and Kumar,N. (2023) Machine Learning Models for Predicting Molecular UV-Vis Spectra with Quantum Mechanical Properties. J Chem Inf Model, 63, 1462–1471.
40. Ko,T.W., Finkler,J.A., Goedecker,S. and Behler,J. (2021) A fourth-generation high-dimensional neural network potential with accurate electrostatics including non-local charge transfer. Nat Commun, 12, 398.
41. Stuyver,T. and Coley,C.W. (2022) Quantum chemistry-augmented neural networks for reactivity prediction: Performance, generalizability, and explainability. The Journal of Chemical Physics, 156, 084104.
42. Guan,Y., Coley,C.W., Wu,H., Ranasinghe,D., Heid,E., Struble,T.J., Pattanaik,L., Green,W.H. and Jensen,K.F. (2021) Regio-selectivity prediction with a machine-learned reaction representation and on-the-fly quantum mechanical descriptors. Chem. Sci., 12, 2198–2208.
43. Huang,B. and Von Lilienfeld,O.A. (2021) Ab Initio Machine Learning in Chemical Compound Space. Chem. Rev., 121, 10001–10036.
44. Tortorella,S., Carosati,E., Sorbi,G., Bocci,G., Cross,S., Cruciani,G. and Storchi,L. (2021) Combining machine learning and quantum mechanics yields more chemically aware molecular descriptors for medicinal chemistry applications. J Comput Chem, 42, 2068–2078.
45. Berthold,M.R., Cebron,N., Dill,F., Gabriel,T.R., Kötter,T., Meinl,T., Ohl,P., Sieb,C., Thiel,K. and Wiswedel,B. (2008) KNIME: The Konstanz Information Miner. In Preisach,C., Burkhardt,H., Schmidt-Thieme,L., Decker,R. (eds), Data Analysis, Machine Learning and Applications, Studies in Classification, Data Analysis, and Knowledge Organization. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 319–326.
46. Wang,Y.-H., Yang,Q.-F., Lin,X., Chen,D., Wang,Z.-Y., Chen,B., Han,H.-Y., Chen,H.-D., Cai,K.-C., Li,Q., et al. (2022) G4LDB 2.2: a database for discovering and studying G-quadruplex and i-Motif ligands. Nucleic Acids Res, 50, D150–D160.
47. Verga,D., Granzhan,A. and Teulade-Fichou,M.-P. (2023) Targeting Quadruplex Nucleic Acids: The Bisquinolinium Saga. In Sugimoto,N. (ed), Handbook of Chemical Biology of Nucleic Acids. Springer Nature Singapore, Singapore, pp. 1–57.
48. Ma,T.-Z., Liu,L.-Y., Zeng,Y.-L., Ding,K., Zhang,H., Liu,W., Cao,Q., Xia,W., Xiong,X., Wu,C., et

al. (2024) G-quadruplex-guided cisplatin triggers multiple pathways in targeted chemotherapy and immunotherapy. Chem. Sci., 15, 9756–9774.
49. Chung,W.J., Heddi,B., Hamon,F., Teulade‐Fichou,M. and Phan,A.T. (2014) Solution Structure of a G‐quadruplex Bound to the Bisquinolinium Compound Phen‐DC 3. Angew Chem Int Ed, 53, 999–1002.
50. Mauri,A. (2020) alvaDesc: A Tool to Calculate and Analyze Molecular Descriptors and Fingerprints. In Roy,K. (ed), Ecotoxicological QSARs, Methods in Pharmacology and Toxicology. Springer US, New York, NY, pp. 801–820.
51. Mauri,A. and Bertola,M. (2022) Alvascience: A New Software Suite for the QSAR Workflow Applied to the Blood-Brain Barrier Permeability. Int J Mol Sci, 23, 12882.
52. Berman,J.J. (2016) Understanding Your Data. In Data Simplification. Elsevier, pp. 135–187.
53. Schöning,V. and Hammann,F. (2018) How far have decision tree models come for data mining in drug discovery? Expert Opinion on Drug Discovery, 13, 1067–1069.
54. Talevi,A., Morales,J.F., Hather,G., Podichetty,J.T., Kim,S., Bloomingdale,P.C., Kim,S., Burton,J., Brown,J.D., Winterstein,A.G., et al. (2020) Machine Learning in Drug Discovery and Development Part 1: A Primer. CPT Pharmacom & Syst Pharma, 9, 129–142.
55. Obaido,G., Mienye,I.D., Egbelowo,O.F., Emmanuel,I.D., Ogunleye,A., Ogbuokiri,B., Mienye,P. and Aruleba,K. (2024) Supervised machine learning in drug discovery and development: Algorithms, applications, challenges, and prospects. Machine Learning with Applications, 17, 100576.
56. Arola-Arnal,A., Benet-Buchholz,J., Neidle,S. and Vilar,R. (2008) Effects of Metal Coordination Geometry on Stabilization of Human Telomeric Quadruplex DNA by Square-Planar and Square-Pyramidal Metal Complexes. Inorg. Chem., 47, 11910–11919.
57. Ruehl,C.L., Lim,A.H.M., Kench,T., Mann,D.J. and Vilar,R. (2019) An Octahedral Cobalt(III) Complex with Axial NH 3 Ligands that Templates and Selectively Stabilises G‐quadruplex DNA. Chemistry A European J, 25, 9691–9700.
58. Zhou,C., Liao,T., Li,Z., Gonzalez‐Garcia,J., Reynolds,M., Zou,M. and Vilar,R. (2017) Dinickel–Salphen Complexes as Binders of Human Telomeric Dimeric G‐Quadruplexes. Chemistry A European J, 23, 4713–4722.
59. Salsbury,A.M. and Lemkul,J.A. (2019) Molecular Dynamics Simulations of the c-kit1 Promoter G-Quadruplex: Importance of Electronic Polarization on Stability and Cooperative Ion Binding. J Phys Chem B, 123, 148–159.
60. Figueiredo,J., Mergny,J.-L. and Cruz,C. (2024) G-quadruplex ligands in cancer therapy: Progress, challenges, and clinical perspectives. Life Sci, 340, 122481.