JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

Harmonization of QSAR Best Practices and Molecular Docking Provides an Efficient Virtual Screening Tool for Discovering New G-Quadruplex Ligands

Daimel Castillo-González,[○,▽] Jean-Louis Mergny,[○,▽] Aurore De Rache,[○,▽] Gisselle Pérez-Machado,[◆,¶,∥] Miguel Angel Cabrera-Pérez,[◆,∥,#] Orazio Nicolotti,[□] Antonellina Introcaso,[□] Giuseppe Felice Mangiatordi,[□] Aurore Guédin,[○,▽] Anne Bourdoncle,[○,▽] Teresa Garrigues,[∥] Federico Pallardó,[¶] M. Natália D. S. Cordeiro,[★] Cesar Paz-y-Miño,[†] Eduardo Tejera,[†] Fernanda Borges,*[,‡] and Maykel Cruz-Monteagudo*[,†,‡]

[○]ARNA Laboratory, IECB, University of Bordeaux, F-33600 Pessac, France

[▽]ARNA Laboratory, INSERM, U869, F-33000 Bordeaux, France

[◆]Molecular Simulation and Drug Design Group, Centro de Bioactivos Químicos (CBQ), Central University of Las Villas, Santa Clara, Villa Clara 54830, Cuba

[¶]Department of Physiology, Faculty of Medicine, University of Valencia, Valencia 46010, Valencia, Spain

[∥]Department of Pharmacy and Pharmaceutical Technology, University of Valencia, Burjassot 46100, Valencia, Spain

[#]Department of Engineering, Area of Pharmacy and Pharmaceutical Technology, Miguel Hernández University, 03550 Sant Joan d'Alacant, Alicante, Alicante, Spain

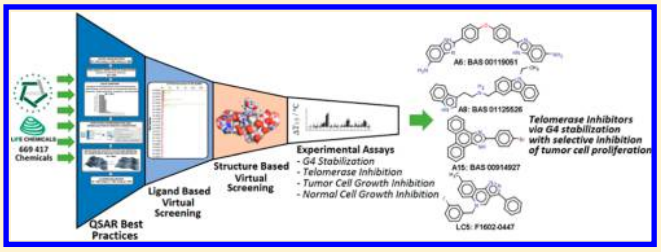
[□]Dipartimento di Farmacia-Scienze, Università degli Studi di Bari "Aldo Moro", Via Orabona 4, 70125 Bari, Bari, Italy

[★]REQUIMTE, Departamento de Química e Bioquímica, Faculdade de Ciências and [‡]CIQUP/Departamento de Química e Bioquímica, Faculdade de Ciências, Universidade do Porto, Porto 4169-007, Portugal

[†]Instituto de Investigaciones Biomédicas (IIB), Universidad de Las Américas, 170513 Quito, Pichincha, Ecuador

Supporting Information

ABSTRACT: Telomeres and telomerase are key players in tumorogenesis. Among the various strategies proposed for telomerase inhibition or telomere uncapping, the stabilization of telomeric G-quadruplex (G4) structures is a very promising one. Additionally, G4 stabilizing ligands also act over tumors mediated by the alternative elongation of telomeres. Accordingly, the discovery of novel compounds able to act on telomeres and/or inhibit the telomerase enzyme by stabilizing DNA telomeric G4 structures as well as the development of approaches efficiently prioritizing such compounds constitute active areas of research in computational medicinal chemistry and anticancer drug discovery. In this direction, we applied a virtual screening strategy based on the rigorous application of QSAR best practices and its harmonized integration with structure-based methods. More than 600,000 compounds from commercial databases were screened, the first 99 compounds were prioritized, and 21 commercially available and structurally diverse candidates were purchased and submitted to experimental assays. Such strategy proved to be highly efficient in the prioritization of G4 stabilizer hits, with a hit rate of 23.5%. The best G4 stabilizer hit found exhibited a shift in melting temperature from FRET assay of +7.3 °C at 5 μM, while three other candidates also exhibited a promising stabilizing profile. The two most promising candidates also exhibited a good telomerase inhibitory ability and a mild inhibition of HeLa cells growth. None of these candidates showed antiproliferative effects in normal fibroblasts. Finally, the proposed virtual screening strategy proved to be a practical and reliable tool for the discovery of novel G4 ligands which can be used as starting points of further optimization campaigns.



INTRODUCTION

The telomerase enzyme, a reverse transcriptase responsible for the addition of a repetitive DNA sequence to the ends of linear eukaryotic chromosomes, is involved in the phenomenon of cellular immortalization.[1,2] This enzyme is active in more than 85% of tumor cells but is not active in the majority of normal cells, with some notable exceptions such as stem cells and

Received: June 29, 2015
Published: September 10, 2015

DOI: 10.1021/acs.jcim.5b00415
J. Chem. Inf. Model. 2015, 55, 2094−2110

germline cells. Telomerase is therefore a promising target for the treatment of malignancies.[3]

Various strategies have been proposed to interfere with telomeric functions of cancer cells and/or inhibit the telomerase enzyme.[3−6] One of the strategies used to indirectly inhibit the telomerase is the stabilization of telomeric G-quadruplex (G4) structures.[7] Although therapeutic interventions targeting telomeric G4 structures have been the most explored,[8−10] G4 forming sequences are found in other contexts. More than 40% of human gene promoters contain one or more G4 motifs, and these motifs have more probabilities of being located in proto-oncogenes promoters regions than in genes that suppress tumor growth. Consequently, the presence of guanine tetrads in multiple regulatory regions and promoters of oncogenes suggest that these structures can provide targets for anticancer strategies,[11−15] opening a door to a multitarget approach.

On the other hand, although many tumor cells avoid replicative senescence via the activation of telomerase, the alternative elongation of telomere mechanism represents an alternative telomere maintenance mechanism which hampers the antitumor efficacy of telomerase inhibitors. It is widely acknowledged that in any of these mechanisms, the participation of G4 structures is potentially important. The differences in telomere length or accessibility between normal and malignant cells may provide specificity and a high therapeutic index for G4 ligands. Such ligands offer the advantage of also acting over tumors mediated by alternative mechanisms, in contrast to telomerase catalytic inhibitors.[16−18]

The search for molecules that bind G4 structures can be performed by resorting to several experimental techniques.[19,20] However, it is important to optimize the ligand design based on an understanding of the correlation between the structure and mechanism of action and to develop screening methods relevant to the mechanism that are reproducible, inexpensive, and highly efficient. In this direction, chemoinformatics can efficiently complement virtual screening (VS) strategies improving the identification and prioritization of compounds able to stabilize G4 structures for further experimental tests, thereby reducing the experimental effort and boosting the overall success rate. Several studies have reported virtual screening aided methodologies for the discovery of novel G4 stabilizers and/or telomerase inhibitors. Most of them mainly relied on structure-based approaches.[7,21−26]

In this work, we used a virtual screening strategy based on the rigorous application of quantitative structure−activity relationships (QSAR) best practices and its harmonized integration with structure-based methods. Such a strategy demonstrated to be highly efficient and led to the identification of several novel G4 stabilizers ligands.

RESULTS AND DISCUSSION

Training Set Optimization. It is well-known that QSAR models based on inaccurate structural or biological data will produce statistically insignificant and unreliable predictions. In this context, Young and co-workers[27] clearly pointed out the importance of chemical data curation. The study concluded that even small structural errors within a data set could significantly deteriorate the generalization ability of QSAR models, while manual curation of structural data leads to a substantial increase in the model predictivity. More recently Tropsha and co-workers[28] strongly encouraged to view chemical data curation not only as a separate and critical component of any chemoinformatics research but also to amend it as the sixth rule of the OECD QSAR modeling and validation principles.[29]

First, a preliminary inspection of duplicate structures in the full data set (see details in Table S1 of the Supporting Information) allowed for detection of two pairs of duplicated compounds (**acp-009** and **acp-038**; **tri-067** and **tri-147**). From the first pair just **acp-038** was removed since the values of the concentration required to cause 50% inhibition of the telomerase enzyme activity (IC_{50}) were very similar (a difference of just 0.03 μM), while both compounds in the second pair were removed due to a higher discrepancy on their IC_{50} values (0.2 μM).

Next, the distribution of the elements present in the remaining 780 compounds was analyzed. The bar graph in Figure 1 shows

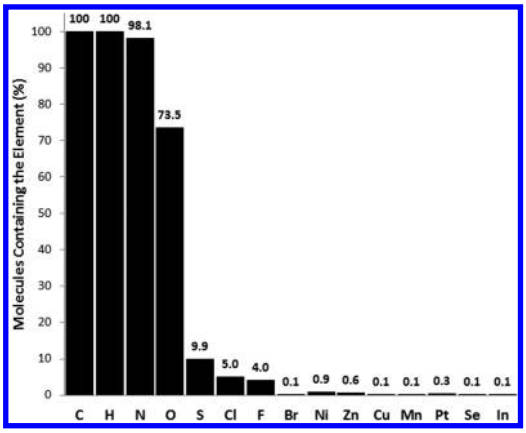


Figure 1. Distribution of the elements present in the compounds remaining after duplicates removal.

the percentage of molecules in the remaining data set containing the respective element. As a result, four molecules containing rare elements (Se, In, and Pt) were detected and removed, in addition to 14 organometallics. It was also possible to confirm that the data set was free of mixtures and inorganics.

The removal of organometallics and molecules containing rare elements are justified by the fact that not all molecular descriptors calculation software correctly handles such compounds. In fact, the molecular structure in this work was codified by DRAGON 6.0 descriptors,[30] which actually could not compute the whole set of molecular descriptors for these specific chemotypes. Nevertheless, we decided to remove these compounds mainly due to a significant underrepresentation of its specific elements in the data set (structural outliers), which represents an important source of noise with a negative influence on a further model learning process. Compound **por-005**, the only compound in the data set containing a bromine atom in its structure, was removed too for identical reasons.

The remaining subset of 761 compounds, constituted just by C, H, N, O, S, Cl, and F atoms, were subject to a standardization process of the chemical structure provided in the respective SMILES codes, as depicted in the Experimental Section. As a result, an SDF file was obtained comprising a standardized 3D representation of the 761 compounds in the protonation state corresponding to pH = 7.2, which was used for further data analysis and modeling.

In this context it is worthy to recall that the presence of structure activity relationship (SAR) continuity provides a fundamental basis for QSAR analysis and resultant compound activity predictions,[31−33] while the presence of SAR discontinuity hinders the proper application of the QSAR paradigm.[33−35] To accept and keep in mind this fundamental limitation is the usual behavior among those practicing QSAR best practices.[36]

DOI: 10.1021/acs.jcim.5b00415
J. Chem. Inf. Model. 2015, 55, 2094−2110

However, little work has been devoted to alleviate it, and to the best of our knowledge no one has been directed to reduce SAR discontinuity on a data set and consequently to restore as much as possible the fundamental principle of QSAR and similarity-based methods.[31,37−40]

In the present work, we instead approached this problem by removing from the training process those problematic compounds inducing SAR discontinuity, with the intended purpose of restoring the SAR continuity necessary for deriving reliable and predictive QSAR models. Such problematic compounds are defined as activity cliffs generators (ACGs).[41] The main assumption of this solution is that a machine learning algorithm that learns from a training set free of the noise induced by ACGs should return models able to identify the structural and/or physicochemical patterns determining the desired activity in a sharper way than an algorithm that learns from a training set including those problematic examples.[40]

However, if the SAR continuity is restored by eliminating compounds, this process might reduce the applicability domain of the model. A possible solution to this problem could be to reach a balance between the predictive ability of the QSAR models developed and their applicability domain. According to the work of Guha,[42] even if a machine learning model captures the most significant activity cliffs they would have to be memorized by the model. So, we may be forced to choose between a predictive model with a certain loss of applicability domain (by removing ACGs) and a model efficiently capturing the SAR but at the cost of a certain degree of overfitting, which hampers its predictive ability (by keeping ACGs).

A remedial measure to soften the loss of applicability domain can be to derive several diverse machine learning models to implement a consensus classifier,[43,44] which was the strategy used in this work. It is well-known that multiclassifiers, ensemble, or consensus classifiers are effective, among other reasons, because they span the decision space. Each base classifier covers a different region of the decision space (chemical space or SAR) and the union of all the base classifiers produces a common region that results in a wider chemical coverage or applicability domain.[43,45] So, in our opinion, it is worth testing the hypothesis of ACGs removal since reduction of the applicability domain seems to have a remedial solution whereas the overfitting does not. However, some questions still remain. To what extent is the learning process affected? What are the effects on the generalization ability of the selected pattern after giving up to the information encoded in the activity cliff pairs?

In this scenario, we can just validate our hypothesis by controlling the ability of the models, trained as above assumed, in predicting activity cliffs members or ACGs. To this end, we first identified 220 pairs of compounds (comprising 221 unique compounds) forming potential activity cliffs by applying the procedure for activity cliffs detection described in the Experimental Section. Next, by applying the discrete definition of activity cliffs described in the Experimental Section, the 220 pairs of potential activity cliffs identified by the Structure Activity Landscape Index (SALI)[46] approach were reduced to only 142 (164 unique compounds) with a maximum value of ${}^{Norm}ED_{ij}$ = 0.06 and a minimum value of ΔPot_{ij} = 0.21. ${}^{Norm}ED_{ij}$ and ΔPot_{ij} are defined in the Experimental Section.

Once we identified the compounds pairs forming activity cliffs we proceeded to remove those compounds with higher influence over the SAR discontinuity of the curated data set of 761 telomerase inhibitors. Such a procedure produced a list of 77 unique compounds to be removed due to their significant influence over the SAR discontinuity. As explained above, this subset was used as an external evaluation set specifically reserved for the evaluation of the ability of the derived models to correctly predict this challenging type of compounds.

For example, the most influencing compound identified in the data set was **ac-029**, which forms cliff pairs with eight other compounds. As detailed in the Experimental Section, in this work those compounds with IC_{50} values $\leq 1\ \mu M$ were assigned to the "active" class (Class_1), otherwise the compound was labeled as "inactive" (Class_0). As can be noted in Table S2 (see the Supporting Information), all cliff partners have a potency value in the class's borderline, and **ac-029** is the only one belonging to Class_0. At the same time, the difference of potency approaches or exceeds 1 order of magnitude for all pairs. This situation induces a significant discontinuity to a preassumed continue SAR. So, the removal of **ac-029** clearly contributes to alleviate this effect by eliminating the only representative with a potency value determining a different class within a subset of compounds with a common chemotype and class. On the other hand, the less significant activity cliff pairs covered by the discrete definition applied are those with the minimum/maximum difference in potency/structure. In this respect, the interested readers can have a look at Table S3 (see the Supporting Information), where can be noted that even in the worst situation the structural similarity is clear and, in the case where the difference of potency is minimal, the cliff partners belong to different classes.

It is worthy to say that chemoinformatics data sets, especially those important for drug discovery programs, are characterized by a very small number of active molecules compared to inactive ones. When machine learning classifiers trained under these conditions are applied to similarly distributed validation sets, the number of molecules correctly assigned to the respective class tends to be artificially high, leading to an overestimation of the generalization ability of the classifier, being the classifier probably biased toward the inactive molecules.[36,47] This situation comes from the well-known inability of most standard machine learning algorithms to handle data with an imbalanced class distribution, since they assume a relatively balanced class distribution and equal misclassification costs.[48] On the other hand, studies conducted on the machine learning area evidence that the sensitivity of standard machine learning algorithms to class imbalance increases with the complexity of the classification problem (as the classes become less linearly separable).[49,50] In order to favor the learning process and to ensure a proper generalization ability of machine learning classifiers, it is required to provide data sets with the presumed classes balance. Once applied the classes balancing procedure described in the Experimental Section 267 Class_0 compounds remained, which provides the required classes balance with respect to the remaining set of 262 Class_1 compounds.

Finally, after preprocessing the initial raw data, it is possible to assert that the final result is a curated, standardized, and balanced data set, optimized to train predictive and reliable machine learning classifiers. As shown in Table S4 (see the Supporting Information), we can assume that the structural and activity information encoded on the initial (raw) full set was kept as much as possible on the final (curated) data set. Additionally, we could confirm that the curated data set is free of duplicates using the "Find duplicate structures" option of the EdiSDF program[51] included on the ISIDA project.[52,53] A schematic representation of the training set optimization process applied in this work is shown in Figure 2.

DOI: 10.1021/acs.jcim.5b00415
J. Chem. Inf. Model. 2015, 55, 2094−2110

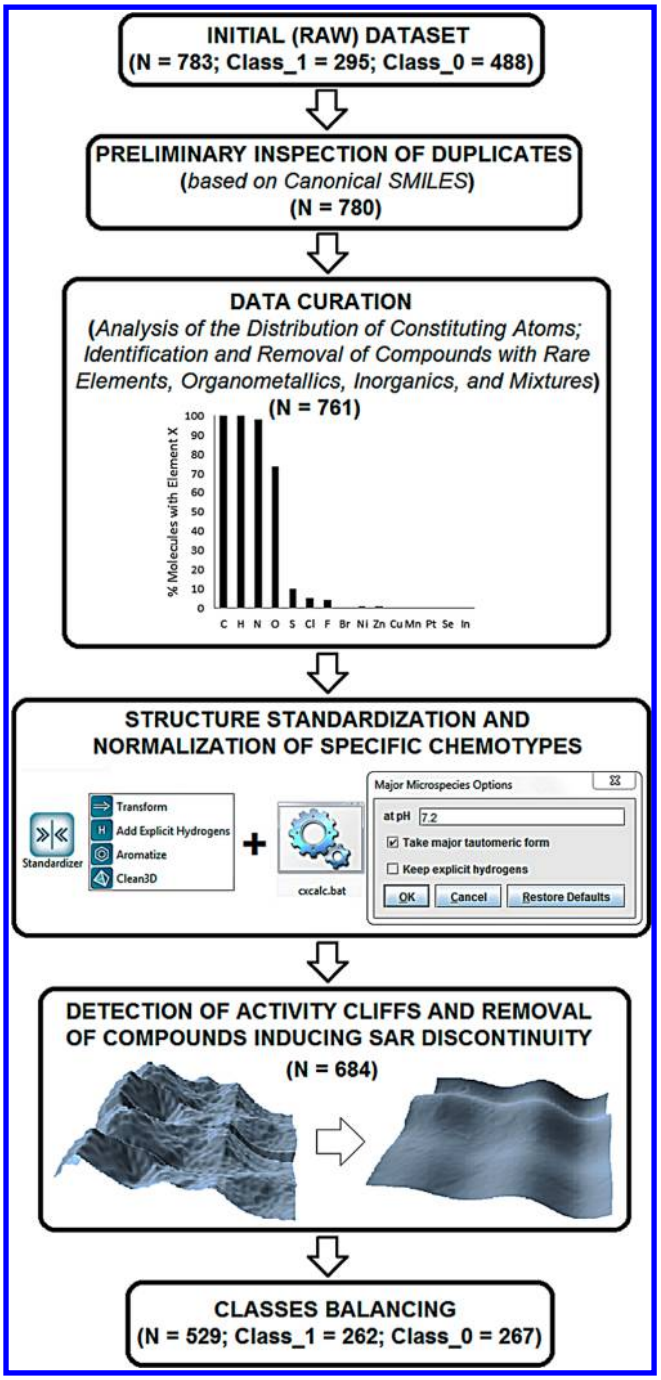


Figure 2. Schematic representation of the training set optimization process.

Models Building and Consensus Classifier. More than 300 classifiers based on DRAGON descriptors and WEKA machine learning classifications algorithms[48] were fitted and evaluated on training (including 10-fold cross-validation), test, and all external evaluation sets, as described in the Experimental Section. Only 81 classifiers with values of accuracy, sensitivity, and specificity ≥0.6 on the training (including 10-fold cross-validation), test, and all external evaluation sets were considered as predictive[36] and reliable, which further advanced to the stage of deriving a consensus classifier.

As depicted in the Experimental Section, only those 27 classifiers with average accuracy-weighted distance values (${}^{Acc.}ED_{ij}$) higher than the overall average accuracy-weighted distance between the 81 valid classifiers were selected to be included in the final consensus classifier. Details on the performance of these 27 classifiers are provided in Table S5 (see the Supporting Information).

The consensus predicted classes were obtained by fusing (summing) the predicted class from each of the 27 diverse and high-performance base classifiers via a majority vote approach. Several weighting schemes were explored to select the scheme(s) rendering the best classification performance. In this sense, the predicted class of each compound according to each base classifier was weighted (multiplied) by the accuracy, kappa coefficient, and correct classification rate (CCR)[54] statistics of the respective classifier.

For the case of unweighted majority vote those compounds predicted as Class_1/Class_0 are labeled as 1/0. The corresponding unweighted consensus score (SUM^U) is thus computed as

$$SUM^U = \sum_{i=1}^{27} PredClass_i \quad (1)$$

In eq 1 $PredClass_i$ accounts for the class assigned by the classifier i to each compound in the data set. Then, if $SUM^U \geq 14$ the compound is predicted as Class_1, otherwise as Class_0.

For the case of weighted majority vote those compounds predicted as Class_1/Class_0 are labeled as 1/−1. The corresponding weighted consensus score (SUM^W) is thus computed as

$$SUM^W = \sum_{i=1}^{27} PredClass_i \times W \quad (2)$$

In eq 2 W represents the weighting feature selected, which can be the accuracy, kappa coefficient, or the CCR statistics associated with the corresponding classifier. Then, if $SUM^W \geq 0$ the compound is predicted as Class_1, otherwise as Class_0.

In order to test the suitability for consensus classification of the 27 selected base classifiers, an initial comparison was conducted on the External Evaluation Set between the base classifier with best performance, the consensus predictions derived from the 81 high-performance classifiers, and the ensemble formed by the 27 diverse and high-performance classifiers.

As can be noted in Table 1, the classification performance of the consensus classifier based on the 27 diverse and high-performance classifiers indeed overcomes the classification performance of the best base classifier and the consensus classifier based on the 81 high-performance classifiers, respectively. Specifically, the results of the majority vote based on the 27 base classifiers weighted by kappa coefficient favorably compares with the rest of the alternative classifiers on all the classification performance metrics except the specificity, which equals the best base classifier. A more detailed comparison on all the subsets (training, test, and external sets) is presented in Table S5 (see the Supporting Information). From this comparison it can be deduced that no significant differences on the classification performance of weighted and unweighted majority vote schemes are observed. So, both schemes can be applied in a further virtual screening campaign.

Retrospective Evaluation of the Ligand-Based Virtual Screening (LBVS) Performance. So far, the 27 base classifiers derived as well as the corresponding consensus classifiers have proven to be predictive and reliable enough to support their use as a virtual screening tool, providing a practical solution to the

DOI: 10.1021/acs.jcim.5b00415
J. Chem. Inf. Model. 2015, 55, 2094−2110

Table 1. Classification Performance Shown by the Best Base Classifier, the Consensus Classifier Based on All 81 High-Performance Classifiers, and the Consensus Classifier Based on the Subset of 27 Diverse and High-Performance Classifiers

classification schemes			*Acc*	*Se*	*Sp*	*FP*	*FN*	*Kappa*
best model			0.876	0.885	0.868	0.132	0.115	0.752
consensus 81 classifiers	majority vote	Unweight	0.867	0.904	0.830	0.170	0.096	0.734
		W(Acc)	0.867	0.904	0.830	0.170	0.096	0.734
		W(CCR)	0.867	0.904	0.830	0.170	0.096	0.734
		W(Kappa)	0.867	0.904	0.830	0.170	0.096	0.734
consensus 27 classifiers		Unweigh	0.886	0.904	0.868	0.132	0.096	0.772
		W(Acc)	0.886	0.904	0.868	0.132	0.096	0.772
		W(CCR)	0.886	0.904	0.868	0.132	0.096	0.772
		W(Kappa)	0.895	0.923	0.868	0.132	0.077	0.791

automatic identification of telomerase inhibitors via G4 stabilization. Anyhow, a good classification performance does not ensure the usefulness of a classifier as a virtual screening tool.[55] So, the potential enrichment ability of the proposed VS strategy needs to be estimated as realistically as possible by means of established enrichment metrics (see the Supporting Information).

In this sense, we decided to apply a virtual screening strategy based on the sequential use of the scores derived from both unweighted (SUM^U) and weighted (SUM^W) majority vote consensus classifiers. Based on the results obtained from the comparison of the different weighting schemes we decided to use only the scores derived from the kappa-weighted (SUM^K) consensus classifier.

A high-quality VS tool should render an ordered list of candidates where promising ones are placed at the top of the list, while irrelevant or detrimental candidates are relegated to the bottom of the same list. For this, the prediction algorithm on which the VS tool is based must be characterized by a good predictive performance. In this respect, VS is expected to return a particularly high/low true positives (*TP*)/false positives (*FP*) rate, which means to maximize/minimize the number of actual active/inactive cases correctly classified by the prediction algorithm. Here the class of interest is regarded as the active class.

In fact, both unweighted and kappa-weighted consensus classifiers show good enough *TP* and *FP* rates (around 88% and 13% respectively, as deduced from the external evaluation set). So, from these results we can expect that a subset of candidate molecules classified as potential telomerase inhibitors (Class_1) using any of the two consensus classifiers (or both) should contain around 88% of TPs (i.e., telomerase inhibitors) and 13% of FPs (i.e., inactive compounds).

The other key feature, a high-quality VS tool, requires us to provide a measure to quantitatively score the target property in such a way that using it as ranking criterion the resultant ordered list resembles as much as possible the actual levels of the target property. In our case, both SUM^U and SUM^K exhibit the variability required for library ranking.

So, we decided to test the suitability of the VS strategy consisting in the sequential use of SUM^U and SUM^K as ranking criteria (decreasingly sort the chemical library by SUM^U, then by SUM^K) for the automatic identification of potent ($IC_{50} \leq 1\ \mu M$) telomerase inhibitors via G4 stabilization dispersed in a data set of moderate, weak, or noninhibitors compounds ($IC_{50} > 1\ \mu M$). In doing so, we initially decided to estimate the enrichment performance using the full data set of 783 inhibitors (295 Class_1 and 488 Class_0). However, the reliability of enrichment metrics estimated from such a sample is hampered by its reduced size as well as by the high ratio of active compounds (Class_1 inhibitors). The problem with using a reduced data set is that the enrichment metrics derived exhibit a higher variance compared to significantly large data sets. Experiments conducted by Truchon and Bayly[56] show that the standard deviation associated with enrichment metrics such as the area under the receiver operating characteristic curve (*ROC*) and the area under the accumulation curve (*AUAC*) are high for small data sets and converge to a constant value when the size of the data set increases. In any case, for the problem at hand it is not possible to set up a large enough decoys set, as is the standard in the performance evaluation of virtual screening tools.[56,57] All we can do is to consider the relative error associated with the use of our data set. For this, the relative error associated with the enrichment metrics derived from this data set will be estimated as recommended in ref 56.

The other problem is related to the high ratio of actives (Class_1 inhibitors) which mainly hinders the early recognition ability, which is known as the "saturation effect". That is, for data sets with a high ratio of actives, once active compounds "saturate" the early part of the ordered list, the enrichment metric cannot get any higher; being this effect more acute as the top fraction considered is smaller.[56]

In order to alleviate as much as possible the saturation effect and thus to estimate in a more realistic way the utility of the virtual screening strategy proposed, we decided to simulate an experiment to evaluate the ability of the proposed approach to retrieve just 14 structurally diverse Class_1 inhibitors from a set of 325 Class_0 inhibitors never used on the classifiers training process.

The 14 Class_1 inhibitors were selected in such a way that every chemical family comprised in the test, and all the external evaluation subsets were represented. For this, one Class_1 representative of each of the 11 chemical families (**ac-022**, **acp-003**, **aq-152**, **fen-005**, **mis-009**, **pip-013**, **qui-015**, **quo-002**, **tet-026**, **tri-002**, and **pta-008**) were randomly selected to be included in the VS experiment together with the three Class_1 inhibitors included in the Real External Evaluation Set (**Ant1,5**, **M1**, and **M2**). The structural diversity of this subset of Class_1 inhibitors can be assessed by inspecting Figure S1 (see the Supporting Information). At the same time, all the Class_0 compounds included on the Test, External Evaluation, Real External Evaluation, External Cliff, and External Negative sets (never used for training) were included as decoys (in this case confirmed negative cases) for the VS experiment. Details on the derivation of the above-mentioned evaluation subsets can be found in the Experimental Section. Finally, the resultant subset of 339 compounds is decreasingly sorted according to the computed values of SUM^U, then by SUM^K, and the enrichment ability of the ligand-based VS strategy proposed is finally assessed according to the enrichment metrics previously detailed.

DOI: 10.1021/acs.jcim.5b00415
J. Chem. Inf. Model. 2015, 55, 2094−2110

We are aware that a ratio of actives R_a = 0.0413 (24 Class_0 "decoys" compounds per each Class_1 "active" inhibitor) of this data set is still insufficient to fulfill the minimum of 36 decoys proposed in ref 58. However, the enrichment metrics derived from such a data set can be used as a proper estimate if we consider the relative error associated with each metric. The relative error associated with the enrichment metrics derived from this data set as well as details on their size and composition are provided in Table 2.

Table 2. Details on the Size and Composition of the Data Subset Used To Evaluate the Virtual Screening Performance of the LBVS Approach Proposed, As Well As the Relative Error Associated to the Enrichment Metrics Derived from This Data Set

data set size and composition			
N = 339	n = 14		R_a = 0.0431
relative error (%) associated with the enrichment metrics			
α	*EF*	*RIE*	*BEDROC*
160.9 (*Top* 1%)	2.77	2.31	5.36
32.2 (*Top* 5%)	1.14	1.02	1.04
16.1 (*Top* 10%)	0.78	0.70	0.48
8 (*Top* 20%)	0.52	0.45	0.21
ROC	0.16		
AUAC	0.15		

As can be noted, the relative error is less than or about 1% for most of the enrichment metrics and never exceeds 5.5%. Actually, the enrichment metrics with associated relative errors near 3% are just those corresponding to the top 1% of the data set, where the saturation effect becomes more acute. So, this data provides sufficient evidence to assert that the effect of using this data set does not significantly affect the inferences on the VS performance deduced from the enrichment metrics computed from it.

The respective values of *AUAC* and *ROC* metrics obtained suggest that by using the proposed VS strategy it is possible to rank a potent telomerase inhibitor earlier than a data set compound of moderate, weak, or noninhibitor capacity with a probability >0.80. The excellent enrichment factor (*EF*) value obtained suggests that the ranking obtained provides a top 1% fraction about 8 times more enriched of potent inhibitors compared to just selecting a random 1% fraction from the data set. As can be noted in Table 3,

Table 3. Virtual Screening Performance of the LBVS Strategy Estimated from Classic Enrichment and Early Recognition Metrics

classic enrichment metrics		early recognition metrics	
$EF_{1\%}$	8.0714 (±0.2239)	$RIE_{1\%}$	9.3749 (±0.2169)
$EF_{5\%}$	5.6975 (±0.0650)	$RIE_{5\%}$	4.5812 (±0.0466)
$EF_{10\%}$	3.5609 (±0.0280)	$RIE_{10\%}$	3.7286 (±0.0259)
$EF_{20\%}$	2.4926 (±0.0130)	$RIE_{20\%}$	2.9824 (±0.0135)
AUAC	0.8051 (±0.0013)	$BEDROC_{1\%}$	0.3877 (±0.0208)
ROC	0.8188 (±0.0012)	$BEDROC_{5\%}$	0.2572 (±0.0027)
		$BEDROC_{10\%}$	0.3171 (±0.0015)
		$BEDROC_{20\%}$	0.4374 (±0.0009)

the *EF* values deteriorates at higher top fractions (top 5%, 10%, and 20%) but still outperforms a random selection. These results point to a good overall (deduced from *AUAC* and *ROC*) and local (deduced from *EF*) enrichment ability of the VS strategy, especially at the top 1% fraction.

However, classic enrichment metrics, such as *ROC*, *AUAC*, and *EF*, cannot discriminate between a VS tool that ranks half of the actives at the beginning of the ordered list and the other half at the end from a VS protocol that ranks all actives at the beginning of the list. This feature is the most important property of a VS tool and is known as the "early recognition" ability. Therefore, the analysis of metrics such as the robust initial enhancement (*RIE*) and the Boltzmann-enhanced discrimination of *ROC* (*BEDROC*) are critical to effectively estimate this essential feature on a VS protocol, especially when very large data sets are intended to be screened.

From the analysis of *RIE* at the respective top 1%, 5%, 10%, and 20% fractions we can deduce that the early recognition ability of the approach exhibits a similar behavior to that observed during the analysis of the overall enrichment ability using the *EF* metric. That is, the early recognition ability of the approach is better at early fractions but starts deteriorating as the size of the considered top fraction increases. This pattern is also observed when the metric analyzed is *BEDROC*, except for a recovering and even improvement at top 20% fractions. The probabilistic interpretation of this metric allows confirming that the ability of the VS protocol to rank most of the Class_1 inhibitors at the top of the filtered fraction is consistently superior at top 1% fractions.

Summarizing, the VS strategy showed not only a good enrichment ability (retrieval of active cases in the filtered fraction) but also, more importantly, a very good early recognition ability (placing active cases at the very beginning of the ordered list). So, from the data presented we can conclude that the proposed approach ensures an efficient VS campaign.

Consensus Classifier Screening of Commercial Databases. A total of 669 417 compounds comprised in two commercial databases (313 473 and 355 944 compounds in ASINEX and LIFE CHEMICALS, respectively) were screened following the ligand-based virtual screening (LBVS) protocol above-described. The 669 417 compounds were scored to produce a sorted list in a decreasing order of probabilities to exhibit a favorable telomerase inhibition profile. We decided to keep those compounds in ASINEX and LIFE CHEMICALS libraries with $SUM^U \geq 20$ (>75% concordance) and $SUM^K \geq 7$; except when the compound belongs to the anticancer subset of LIFE CHEMICALS library, in which case the acceptance criteria was $SUM^U \geq 18$ (>66% concordance) and $SUM^K \geq 4$. Following the strategy above-described the top 99 G4 candidates were finally selected. Figure 3 shows the structural diversity of the candidates selected through the corresponding heatmap based on normalized Euclidean distances and ChemAxon's Chemical Fingerprints (ChFP).

Additionally, to assess to the anticancer potential of the chemotypes represented in the 99 preselected candidates, they were compared in structural coverage terms to the National Cancer Institute (NCI) diversity set. The NCI diversity set is a small library of about 2000 synthetic small molecules selected from the full NCI screening collection, which represents a systemic sampling of the structural diversity of the available repository. In this work the NCI diversity set was used as "anticancer structural reference space" to access the structural coverage of our 99 candidates.

The structural coverage of the anticancer structural reference space determined by the NCI diversity set, for the preselected 99 G4 candidates, was accessed by a multidimensional scaling (MDS) plot. The MDS plot was based on the corresponding chemical structures codified by the Extended Connectivity Fingerprints (ECFP) implemented in PaDEL-Descriptors[59]

DOI: 10.1021/acs.jcim.5b00415
J. Chem. Inf. Model. 2015, 55, 2094−2110

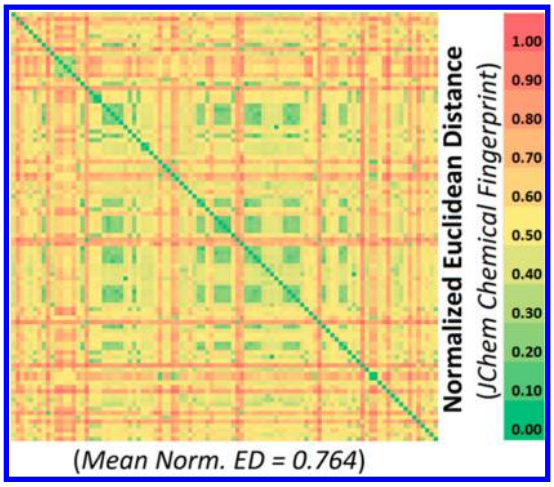


Figure 3. Heatmap of the normalized Euclidean distance between the 99 G4 candidates using ChemAxon's Chemical Fingerprints.

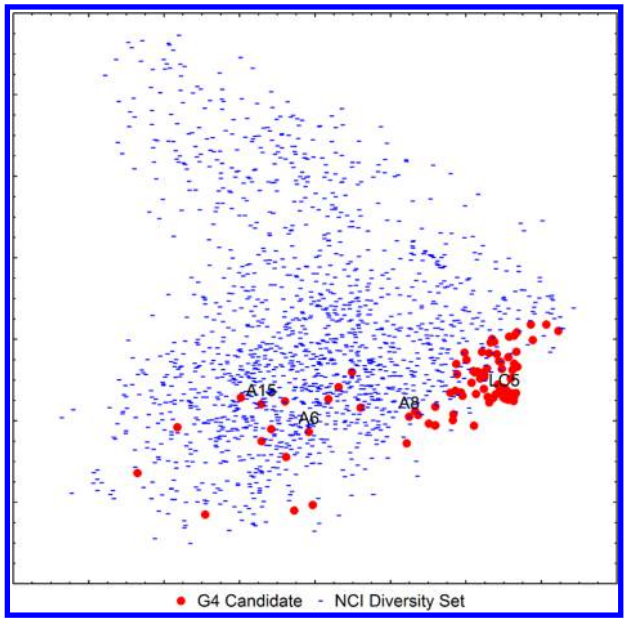


Figure 4. Multidimensional scaling plot of the 99 G4 candidates selected by the LBVS strategy and the 1635 compounds in the NCI diversity set used as "anticancer structural reference space". The four G4 ligand hits experimentally confirmed further are labeled in the graph.

(see Figure 4). The NCI diversity set used for comparison is comprised by 1635 unique compounds included in the Diversity sets III, IV, and V provided in https://wiki.nci.nih.gov/display/NCIDTPdata/Compound+Sets. Duplicate compounds in the three diversity sets were identified and removed using the EdiSDF program[51] included on the ISIDA project.[52,53] The dissimilarity matrix for the MDS was obtained using the *pdist* function implemented in MatLab.[60] The *Jaccard* coefficient was employed as the proximity metric. Then, the MDS was conducted using the *cmdscale* function.

The first observation from the MDS plot provided in Figure 4 is a practically full overlap between the 99 G4 candidates and the 1635 compounds in the NCI diversity set. Such overlap indicates that the 99 G4 candidates exhibit chemotypes similar to a subset of anticancer compounds in the NCI diversity set. This observation supports a potential anticancer utility of such chemotypes. Additionally, the G4 candidates cover a specific but wide region of the anticancer structural space determined by the 1635 compounds in the NCI diversity set. This suggests that these candidates might exhibit a common anticancer mode of action determined by a common structural pattern, which is consistent with the intended purpose of our LBVS strategy of prioritizing potential G4 ligands based on relevant structural information. On the other hand, this wide coverage also supports the conclusions derived from Figure 3 regarding the structural diversity of the 99 G4 candidates.

Refining Screening with Molecular Docking. A molecular docking study was performed on the 99 previously selected candidates in order to prioritize those with more chances to inhibit the telomerase enzyme by a G4 stacking stabilization mechanism. A detailed visual inspection was conducted to shed light on the predicted binding mode of the candidate ligands as well as on the interactions driving the molecular recognition.

The docking calibration carried out with GOLD 5.1[61,62] returned a reliable pose reproducing the binding conformation of the cocrystallized molecule BRACO-19. A root-mean-square deviation (RMSD) value of 1.70 Å was measured by comparing the crystal structure (in green) with the top-scored docking poses (in yellow) of BRACO-19 in the bimolecular G-quadruplex DNA (PDB code: 3CE5) (see Figure 5), providing evidence of the reliability of the docking protocol applied. The top-scored poses of the other four reference G4 stabilizers (Telomestatin, 360A, Piper, TMPyP4) when docked to 3CE5, derived from the docking calibration protocol carried out with GOLD 5.1, are provided in Figure S2 (see the Supporting Information).

Once conducted the docking campaign docking scores were obtained for the 99 preselected candidates. All the compounds were sorted on the basis of the docking average fitness over 10 docking poses (in the range from 93.43 to 40.49 kJ/mol). We doped the list of 99 candidates with the five reference G4 stabilizer compounds to control the reliability of the docking ranking obtained.

In this respect, 79 out of 99 candidates returned an average fitness score higher than Telomestatin (the worst performing reference G4 stabilizer with an average score equal to 67.06 kJ/mol) and were thus preselected. Finally, nine compounds were selected on the basis of a diversity-based selection using as descriptors space the hashed linear binary fingerprints implemented in Canvas,[63] the Soergel distance as structural similarity metric and the maximum sum of pairwise distances for selection. Such an approach should ensure a better sampling of the chemotypes, thus avoiding picking up too similar molecules. The nine proposed molecules were visually inspected to better assess their binding modes. Interestingly, these nine selected compounds returned pretty well superimposed poses. Two representative (top/bottom) compounds are shown in Figure S3 of the Supporting Information.

Integration of Ligand- and Structure-Based Virtual Screening. The 99 candidates prioritized by the LBVS strategy were resorted considering the results of the docking study. The nine most probable and structurally diverse telomerase inhibitors candidates via G-4 stabilization prioritized by the docking study were assigned to the top nine places in the rank list. These nine compounds were specifically sorted in the range [1, 9] using the docking scores, then by SUM^U, then by SUM^K. Six other promising (although less diverse) telomerase inhibitor candidates via G-4 stabilization observed in docking simulations were assigned to the next six places in the ranked list and were specifically sorted in the range [10, 15], using the same sorting

DOI: 10.1021/acs.jcim.5b00415
J. Chem. Inf. Model. 2015, 55, 2094−2110

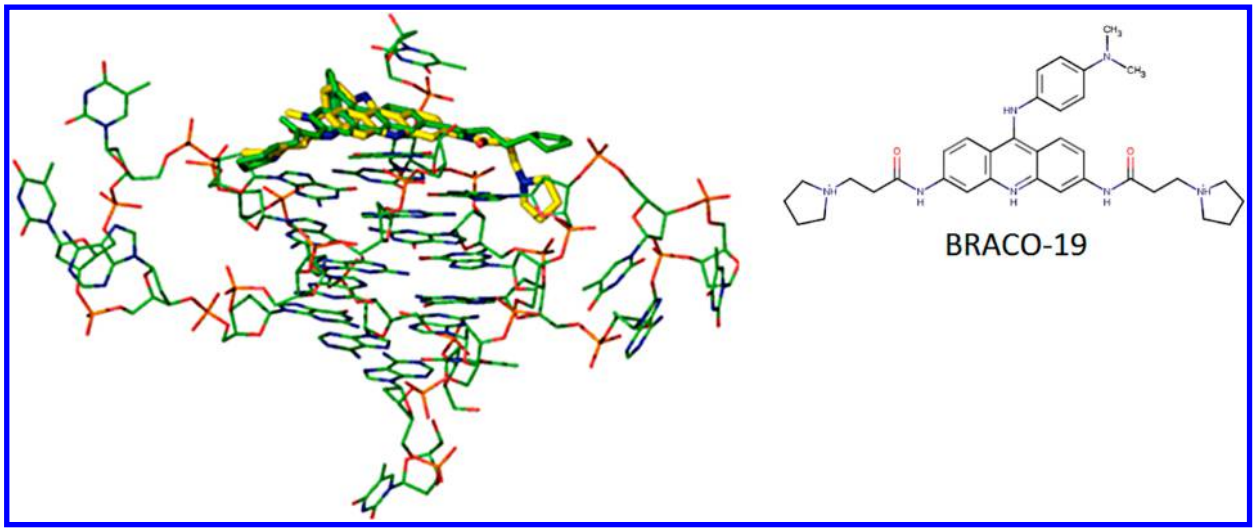


Figure 5. Crystal structure pose (in green) and the top-scored docking pose (in yellow) of BRACO-19 in the bimolecular G4 DNA (PDB code: 3CE5) derived from the docking calibration carried out with GOLD 5.1.

scheme used for the range [1, 9]. The remaining 84 candidates were assigned to the last positions using the same sorting scheme.

Selection and Purchase of the Most Promising and Diverse Candidates. The selection of promising and structurally diverse compounds in ASINEX and LIFE CHEMICALS databases was done according to the harmonized ligand and structure based final ranking above-described, medicinal chemistry expert's criteria and commercial availability from vendors. Some of the top ranked compounds were not available from vendors at the moment of the purchase. So, the 21 top ranked and commercially available candidates were purchased (15 from ASINEX and 6 from LIFE CHEMICALS) and submitted to experimental assays. The structural diversity of this set of compounds can be appreciated in Figure 6, where the chemical structures and the corresponding heatmap (based on normalized Euclidean distances and ChemAxon's Chemical Fingerprints) are shown. As can be noted in Figure 6, most of the compounds are essentially constituted by an extended aromatic core and electronegative atoms like nitrogen, oxygen, fluorine, or sulfur. Due to the low solubility of some compounds, the stock solutions (1 mM for Life Chemicals and 2 mM for Asinex) were prepared in dimethyl sulfoxide (DMSO) and kept at −20 °C. The compounds **A2**, **A5**, **A7**, and **A11** were not soluble even at 1 mM in DMSO and were discarded from the analyzed set of molecules.

G-Quadruplex Stabilization on Human Telomeric Sequences. The fluorescence resonance energy transfer (FRET) melting assay was used to measure the thermal stabilization ($\Delta T_{1/2}$)[64,65] induced by the 16 remaining compounds on the **F21T** intramolecular human telomeric G-quadruplex which consists of 3.5 copies of the human telomeric guanine-rich strand. Figure 7 shows the $\Delta T_{1/2}$ values measured in K^+ and Na^+ conditions with 5 μM of candidate ligands.

Most compounds showed an induced stabilization at 5 μM lower than 5 °C in both K^+ and Na^+ conditions but **A8** did stabilize **F21T** by 7.3 ± 0.6 °C and by 8 ± 1 °C in K^+ and Na^+ conditions, respectively. This significant stabilization reflects the interaction between **A8** and the human telomeric G-quadruplex. Compounds **A6**, **A15**, and **LC5** showed medium stabilization effects with $\Delta T_{1/2}$ values of minimum 2 °C in both K^+ and Na^+ conditions. From Figure 7 it is possible to note that 4 out of the 17 tested compounds exhibited a moderate G4 stabilizing ability, representing so far, a very good G4 stabilizing hit rate of 23.5% for the VS strategy applied.

Inhibition of Telomerase Activity and Tumor Cell Proliferation. It is well-known that G4 stabilizing compounds can act as indirect inhibitors of the human telomerase enzyme.[66,67] Therefore, based on the FRET melting assay results, the telomerase inhibition ability of the most promising candidate (**A8**) was investigated in a cell-free system by the real time quantitative telomeric repeat amplification protocol (RTQ-TRAP). Other three candidates exhibiting a medium (**A15**) or poor (**LC3** and **LC6**) G4 stabilizing profile were also submitted to the telomerase inhibition and cell growth inhibition assays.

As shown in Figure S5A of the Supporting Information, linearity in fluorescence intensity (*Ct*) was observed with the equivalent of 1 to 1000 human cervical adenocarcinoma (HeLa) cells. However, results for either highly concentrated or extremely diluted samples were less reproducible. The fluorescence signals were occasionally detected in samples with the equivalent lesser than 0.1 cells due to primer-dimer artifacts. The determination coefficient (R^2) of 0.9913 obtained for the HeLa-based standard curve supports its use as a convenient and reliable tool to quantify telomerase activity by cell-to cell comparison (see Figure S5B of the Supporting Information).

Fluorescence signals at late amplification stages (35−40 cycles) were detected for negative controls (CHAPS buffer, water and free cell extracts). These signals were out of the quantitative and linear range in real time quantitative polymerase chain reaction (RTQ-PCR), indicating that primer-dimer formation was negligible. In addition, the control real time polymerase chain reaction (PCR) amplification of TSNT in the presence of DMSO produced amplicons with identical *Ct* values as did the water control (data not shown). This observation provides evidence that DMSO did not inhibit real time PCR amplification of TRAP products.

Figure 8 shows the results from the RTQ-TRAP experiments conducted on standardized protein concentrations at least three times. The remaining telomerase activity was determined relative to control (DMSO-treated cells). For the four compounds, *Ct* values were significantly lower than those found for the DMSO-treated control, showing that all assayed chemicals inhibited the telomerase activity at 1 and 10 μM, including those exhibiting a poor G4 stabilizing profile.

DOI: 10.1021/acs.jcim.5b00415
J. Chem. Inf. Model. 2015, 55, 2094−2110

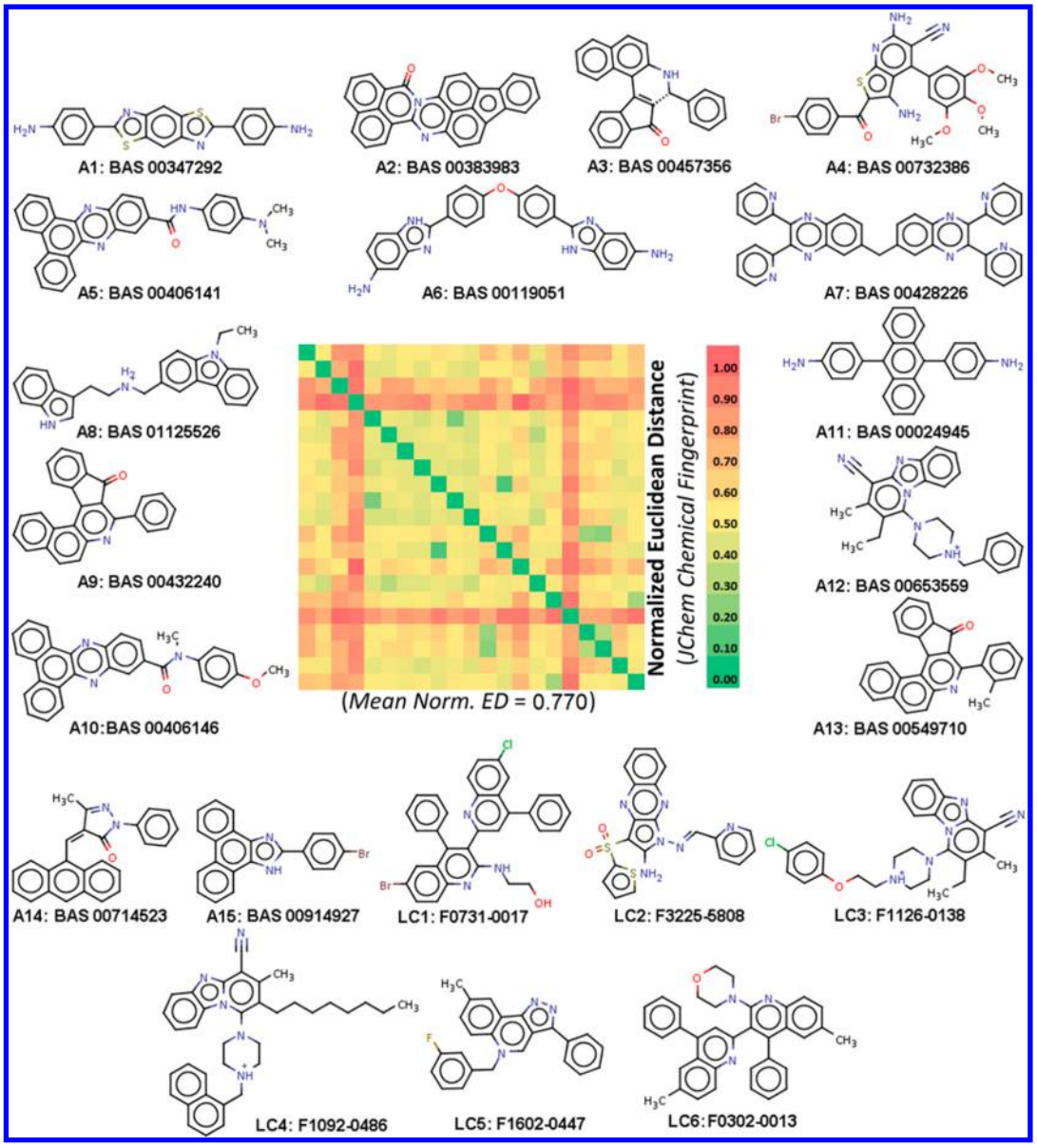


Figure 6. Chemical structures and database identification codes of the 21 telomerase inhibitors via G4 stabilization candidates purchased from ASINEX (labeled with "A") and LIFE CHEMICALS (labeled with "LC") and the corresponding heatmap based on normalized Euclidean distance and ChemAxon's Chemical Fingerprints.

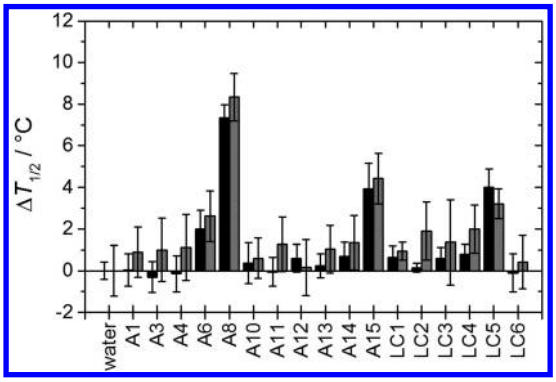


Figure 7. $\Delta T_{1/2}$ values measured in K^+ (black bars) and Na^+ (gray bars) conditions in the presence of 5 μM of the candidate ligands.

The selectivity of these ligands toward cancer cells was evaluated through the effects on proliferation of HeLa cells (telomerase-positive) using a standard sulphorodhamine B (SRB) assay. Cells were exposed to the four compounds at 10 and 1 μM for 24 h (see Figure 9). The four compounds exhibited a mild inhibition of HeLa cells viability, and **A15** presented the most significant inhibiting effect with a remaining cellular viability in this cancer cell line of 80.49% at 10 μM. Remarkably, none of the four compounds evaluated showed antiproliferative effects in normal fibroblasts, suggesting that such chemicals would preferentially limit the growth of cancer cells.

Scaffold Novelty and Drug-likeness. The four G4 hits found were compared in terms of scaffold novelty with 444 already known G4 binders extracted from ChEMBL[68,69] (see structure data in Table S9 of the Supporting Information). None of the G4 hits exhibited a Tanimoto coefficient (Tc) based on ChemAxon's Chemical Fingerprints (ChFP) and Pharmacophore Fingerprints (PhFP) higher than 0.30. See the details in Table S10 of the Supporting Information.

DOI: 10.1021/acs.jcim.5b00415
J. Chem. Inf. Model. 2015, 55, 2094−2110

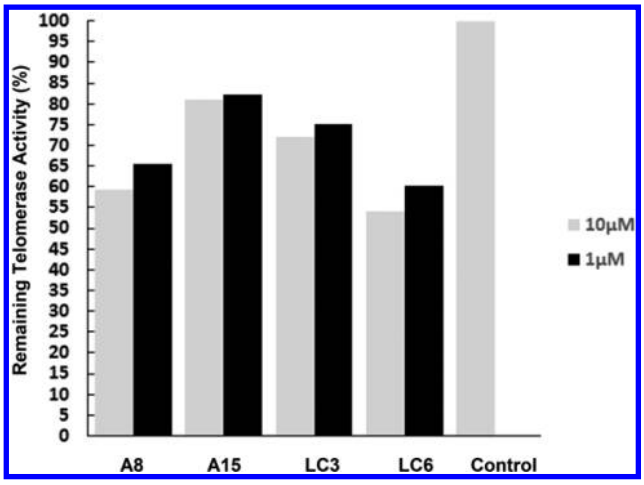


Figure 8. Percentage of remaining telomerase activity in cells treated with the four evaluated ligands and DMSO.

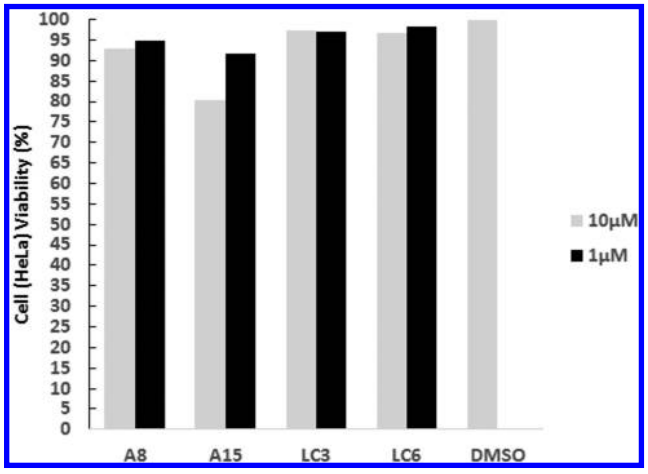


Figure 9. Short-term cytotoxicity for the four evaluated ligands in terms of viability percentage in HeLa line.

A pair of molecules is considered structurally similar if they exhibit a value of Tc based on ECFPs equal or higher than 0.55.[70] On the other hand, ECFPs have a higher structural resolution than both ChFPs and PhFPs. Considering the above-mentioned, we can assert that the four G4 hits found represent novel scaffolds in both chemical and pharmacophoric terms.

Additionally, we decided to estimate the drug-likeness of the four G4 hits found and compare them with seven potent and structurally diverse known G4 binders.[24] See the corresponding structures and references in Table S11 of the Supporting Information. The drug-likeness of the 11 compounds was estimated using a measure of drug-likeness based on the concept of desirability called the quantitative estimate of drug-likeness (QED).[71] Such desirability-based measure of drug-likeness ranges between 0 (for nondrug-like compounds) and 1 (for drug-like compounds). This approach is preferred since evaluation of drug-likeness in absolute terms does not reflect adequately the whole spectrum of compound quality.[71] QED scores were derived from eight widely used molecular properties selected on the basis of published precedence for their relevance in determining drug-likeness:[72−75] molecular weight (MW), octanol/water partition coefficient (LOGP), hydrogen bonds acceptor count (HBA), hydrogen bonds donor count (HBD), polar surface area (PSA), rotatable bonds count (ROTB), aromatic rings count (AROM), and structural alerts count (ALERTS). MW, LOGP, HBA, HBD, PSA, ROTB, and AROM were computed using JChem for Excel functions, while ALERTS was computed by the OCHEM ToxAlerts Web service[76] using 78 reactive, unstable, toxic, genotoxic, carcinogenic, or mutagenic substructures. The corresponding desirability indexes (d_{MW}, d_{LOGP}, d_{HBA}, d_{HBD}, d_{PSA}, d_{ROTB}, d_{AROM}, and d_{ALERTS}) and QED scores were computed using a QED computation example for Excel provided in ref 71 as Supporting Information. From the information provided in Table 4 can be deduced that our four hits exhibit a drug-likeness profile comparable or even better than the seven known potent and structurally diverse G4 binders used as queries.

It is important to note that in general, G4 ligands are not Lipinski compliant, which agrees with the low QED values exhibited by the 11 compounds compared. In order to confirm this particular behavior a network representation of the similarity relationships (according to Tanimoto and MACCS keys) between these 11 compounds plus the 444 known G4 binders extracted from ChEMBL is provided, jointly with their drug-likeness

Table 4. Drug-likeness of the Four G4 Hits Found in the Present Study and the Seven Queries Selected in Ref 24 Estimated by QED

	A6	A8	A15	LC5	BRACO-19	SYUIQ5	RHPS4	CX-3543	triazine-115405	Phen-DC3	BMVC
MW	432.5	368.5	373.3	367.4	593.8	318.4	347.4	604.7	525.7	550.6	403.5
LOGP	4.4	5.3	6	5.9	4.6	3.2	0.9	3.5	6.2	−3	−2.6
HBA	4	0	1	2	7	3	1	7	9	4	0
HBD	4	2	1	0	3	2	0	1	2	2	1
PSA	118.6	37.3	28.7	30.7	96.5	46.4	7.1	92.1	97.5	91.7	23.6
ROTB	4	6	1	3	11	5	0	6	7	4	4
AROM	6	5	5	5	4	4	5	6	5	7	5
ALERTS	9	4	6	3	10	5	8	8	8	7	6
d_{MW}	0.40	0.76	0.73	0.77	0.06	0.99	0.88	0.06	0.13	0.10	0.55
d_{LOGP}	0.72	0.33	0.14	0.17	0.65	0.99	0.74	0.96	0.11	0.04	0.06
d_{HBA}	0.89	0.03	0.23	0.93	0.33	0.98	0.23	0.33	0.11	0.89	0.03
d_{HBD}	0.15	0.79	0.99	0.59	0.38	0.79	0.59	0.99	0.79	0.79	0.99
d_{PSA}	0.49	0.92	0.8	0.83	0.72	0.99	0.33	0.77	0.71	0.77	0.69
d_{ROTB}	0.97	0.69	0.64	0.99	0.15	0.85	0.40	0.69	0.53	0.97	0.97
d_{AROM}	0.01	0.01	0.01	0.01	0.03	0.03	0.01	0.01	0.01	0.01	0.01
d_{ALERT}	0.00	0.03	0.00	0.08	0.00	0.01	0.00	0.00	0.00	0.00	0.00
QED	0.11	0.18	0.16	0.30	0.08	0.34	0.12	0.11	0.08	0.11	0.11

DOI: 10.1021/acs.jcim.5b00415
J. Chem. Inf. Model. 2015, 55, 2094−2110

(estimated by the respective QED values). Such a similarity/drug-likeness network was constructed with the software SARANEA[77] for a set of 455 G4 binders (four G4 hits, seven potent and structurally diverse known G4 queries, and 444 G4 binders extracted from ChEMBL) using MACCS keys and a Tanimoto similarity threshold of 0.85. Nodes represent G4 binders connected by edges if they share a 2D similarity above the predefined threshold. The color of the nodes reflects their drug-likeness quantified by QED, where the drug-likeness increases from red (QED = 0) to green (QED = 1). As can be observed in Figure 10, most G4 binders exhibit a moderate to

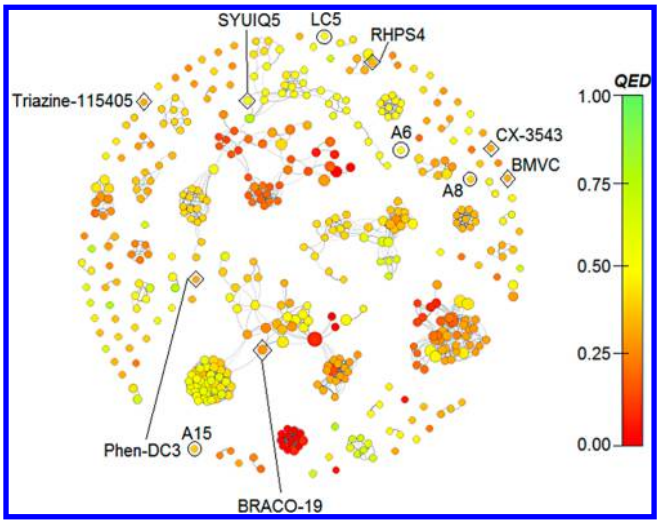


Figure 10. Similarity/drug-likeness network for G4 hits and known G4 binders.

poor drug-likeness (the color of most nodes goes from yellow to orange to red). It can also be noted in this figure that our four G4 hits exhibit a comparable drug-likeness profile. Additionally, the scaffold novelty of our four G4 hits is deduced from this network since none of the four hits are connected (at a similarity threshold of 0.85) to any known G4 binder.

Summarizing, the four G4 hits found can be considered as moderate G4 stabilizers, structurally novel with respect to the existing G4 ligands, and with comparable or superior drug-likeness (as deduced from QED[71]) with respect to about 400 known G4 binders extracted from ChEMBL, including seven well-known potent and structurally diverse G4 binders.

CONCLUSIONS

The proposed virtual screening strategy proved to be highly efficient in the prioritization of G-quadruplex stabilizers, with a hit rate of 23.5%. Notably, four compounds (**A6**, **A8**, **A15**, and **LC5**) were confirmed to be moderate G-quadruplex stabilizers among 17 assayed candidates. Among all, **A8** exhibited the most promising G-quadruplex stabilization profile with the largest shift in melting temperature (+7.3 ± 0.6 °C at 5 μM). The **A8** and **A15** G-quadruplex stabilizers also exhibited a good telomerase inhibitory ability and a mild inhibition of HeLa cells growth. However, no compound showed antiproliferative effects in normal fibroblasts suggesting that these compounds would preferentially limit the growth of cancer cells.

Finally, the results demonstrate that the rigorous application of QSAR best practices and its harmonized integration with structure-based approaches can provide practical and reliable virtual screening tools for the discovery of novel telomerase inhibitors via G-quadruplex stabilization which can be used as a starting point of further optimization campaigns.

EXPERIMENTAL SECTION

Computational Details. *Data Set.* All the modeling efforts conducted in this work were based on an initial large and structurally diverse data set of 783 compounds reported as telomerase inhibitors via G4 stabilization initially compiled by Castillo and co-workers.[7] The concentration reported in the respective primary sources to cause 50% inhibition of the enzyme (IC_{50}) as measured by the TRAP assay was the end point selected to encode the G4-induced telomerase inhibition activity of the compounds in the data set. This data set covers a relatively large and evenly distributed structural and activity space (see details in Table S4 of the Supporting Information) and so can be considered an appropriate learning space to obtain predictive classifiers. In this work the compounds with IC_{50} values ≤ 1 μM are assigned to the "active" class (Class_1), otherwise the compounds are labeled as "inactive" (Class_0). The corresponding identification codes, SMILES, and IC_{50} values are provided in Table S1 (see the Supporting Information). Although this data set has been subject to previous modeling efforts, we decided to conduct a careful data curation on it, previous to any modeling effort, as recommended by Tropsha.[36] The organized protocols for data curation proposed by Fourches et al.[28] constituted (essentially) the guide for the data curation process applied in this work.

Data Set Curation. Previous to the application of any established data curation procedure, a preliminary inspection of duplicate structures was conducted based on the original SMILES codes provided in the raw data set. The original SMILES were converted to the respective canonical form with OpenBabel.[78] Next, the molecular formulas were obtained with the *JCMolFormula* function implemented in JChem for Excel[79] and used to analyze the distribution of the elements present in the data set. This step was conducted to detect rare or under-represented elements, organometallics, mixtures, and inorganics.

The chemical structures provided in the respective SMILES codes were subject to a standardization process. First, the corresponding SDF file was generated using the JChem for the Excel program.[79] The molecular structure representation was then standardized using the ChemAxon's Standardizer.[80] The parameters of the standardization process were set to obtain clean 3D molecular structural representations in SDF format with benzenes in the aromatic form, explicit hydrogens, and a common representation (normalization) of specific chemotypes such as nitro or sulfoxide groups. The SDF file containing the standardized molecular representations generated by the ChemAxon's Standardizer was used as input for the ChemAxon's Calculator (*cxcalc*). By using the *majormicrospecies* option, *cxcalc* produced a SDF file with protonated structures at pH = 7.2, the pH value corresponding to the conditions of further experimental assays.

Detection of Activity Cliffs and Removal of Compounds Inducing SAR Discontinuity. In addition to the standard data curation procedures applied until now, we propose to include the detection of pairs of compounds that form activity cliffs and the removal of those compounds with a significant influence over SAR discontinuity.[40,41] The specific procedure applied to identify activity cliff pairs and the corresponding elimination of the most influencing compounds (ACGs) is described below.

A vector of 306 DRAGON[81] molecular descriptors (including 96 Burden Eigenvalues, 22 Constitutional Indices, 32 Ring

DOI: 10.1021/acs.jcim.5b00415
J. Chem. Inf. Model. 2015, 55, 2094−2110

Descriptors, 71 Functional Group Counts, 73 Atom-centered Fragments, and 12 Molecular Properties) was used as reference space for the structural proximity assessment of the curated set of 761 compounds. The Euclidean distance (ED) was used as the structural proximity measure. The Euclidean distance matrix for the curated set of 761 compounds was obtained using the *Generalized k-means Cluster Analysis* implemented in STATISTICA 8.0.[82] In order to quantify the degree of structural proximity in a more intuitive way, the ED value between each pair of compounds *ij* (ED_{ij}) was normalized or range scaled to [0, 1] by dividing it by the maximum ED_{ij} value ($^{max}ED_{ij}$), obtaining an structural proximity matrix (**S**) based on normalized values ($^{Norm}ED_{ij}$).

$$^{Norm}ED_{ij} = ED_{ij}/^{\max}ED_{ij} \quad (3)$$

At the same time, the absolute difference of potency between each pair of compounds *ij* (ΔPot_{ij}) was computed to obtain the corresponding activity difference matrix (**A**).

$$\Delta Pot_{ij} = |Pot_i - Pot_j| \quad (4)$$

In eq 4 Pot_i and Pot_j represent the potencies of the *i*th and the *j*th compounds according to the decadic logarithm of the IC_{50} values expressed in nM units.

Finally, each element in **A** is divided by the corresponding element in **S** to obtain the corresponding **SALI** matrix whose elements are the Structure Activity Landscape Index[46] for each pair of compounds *ij* ($SALI_{ij}$).

$$SALI_{ij} = \Delta Pot_{ij}/^{Norm}ED_{ij} \quad (5)$$

SALI assigns to each pair of compounds a score that combines their pairwise structural proximity and the difference between their potencies. Therefore, the corresponding elements in the **SALI** matrix can be used to identify and to quantify pairs of compounds forming potential activity cliffs. For this, it is required to apply an arbitrary cutoff $SALI_{ij}$ value intended to detect increasingly significant activity cliffs. The details on the selection of this cutoff are provided in the Supporting Information.

The pairs of potential activity cliffs identified by the SALI approach were screened and refined by the application of a discrete activity cliff definition,[70] adapted to deal with a classification problem. We considered as activity cliffs only those pairs of compounds having a value of $^{Norm}ED_{ij} \leq 0.10$ (structural similarity ≥90%) and a value of $\Delta Pot_{ij} \geq 1$ (potency difference ≥1 order of magnitude). Those pairs involving compounds of different classes were considered as activity cliffs, independently of their values of ΔPot_{ij}. This last was done to reduce the noise induced by structurally similar compounds assigned to different classes (classes overlapping).

Finally, once identified the pairs of compounds forming activity cliffs we proceeded to remove those with more influence on the SAR discontinuity in the curated data set. We tried to minimize the number of compounds removed from the minority class (Class_1). For this, the compounds involved in the cliff pairs identified (according to the discrete definition) were sorted decreasingly according to the number of cliff pairs in which the respective compound was involved. Then, the compounds were sequentially removed from the cliff pairs list according to the above-mentioned ranking criterion until only those involved in only one cliff pair remained, in which case the Class_0 or the less potent partner was removed.

Classes Balancing. Although the imbalance toward inactive cases in our data set is not extreme, the size and composition of the data set allows conducting an undersampling procedure of inactive cases. Rather than simply applying a random procedure, the undersampling approach followed in this work was directed to keep as much as possible the structural and activity information encoded in the initial full set of 422 Class_0 compounds. An undersampled set of 267 Class_0 compounds was finally obtained, which provides the required classes balance with respect to the remaining set of 262 Class_1 compounds. A total of 155 Class_0 compounds were removed to balance the classes. Instead of totally excluding this subset, it was reserved as an additional external evaluation set. The details of the classes balancing procedure applied are provided in the Supporting Information.

Training/Evaluation Data Splitting. The curated, standardized, and balanced data set of 529 telomerase inhibitors was split into three subsets: training, test, and external evaluation sets, as part of the model validation scheme.[36] First, 105 compounds (~20%) were randomly selected as an "external evaluation set" using the *Create Subset/Random Sampling* option implemented in the software package STATISTICA 8.0.[82] This procedure was applied to each class separately. So, our external evaluation set includes 52/53 Class_1/Class_0 compounds. The goal of this external evaluation set is to reproduce in the best possible way a real life situation, where any subset of compounds can be provided for evaluation using the predictive model derived. Thus, the performance of the prediction model on this subset will be the most important indicator of its predictive or generalization ability.

The remaining set of 424 telomerase inhibitors was divided into training and test sets by the application of a *Generalized k-means Cluster Analysis*,[32] as implemented in the *Data Mining* module of STATISTICA 8.0.[82] The vector of 306 DRAGON molecular descriptors used on the activity cliffs identification and classes balancing stages was used as structural reference space. The generalized *k*-means cluster analysis was independently applied to the members of each class (210 Class_1 and 214 Class_0). The Euclidean distance was used as structural proximity measure, and the optimal number of clusters was determined through the 5-fold cross-validation procedure implemented in the module. Approximately 20% of the respective Class_1/Class_0 compounds were reserved for the test set in such a way that each cluster is represented in both training and test subsets. Therefore, 339 (168 Class_1 and 171 Class_0) out of 424 compounds were used for training, while the remaining 85 (42 Class_1 and 43 Class_0) were reserved for the test set and never used for training. This procedure ensures that both training and test subsets are uniformly populated from the molecular structure point of view and that each structure pattern on the test subset is represented on the training subset. The goal here is to guarantee that predictions of new compounds based on models derived from such a training subset will be based on interpolations, avoiding the lack of reliability associated with extrapolations.[36,83]

A small set of 19 G4-induced telomerase inhibitors (3 Class_1 and 16 Class_0 compounds) collected from studies[84−88] published after the classifiers were obtained was also used as a "real external evaluation set". The corresponding identification codes, SMILES, and IC_{50} values are provided in Table S6 (see the Supporting Information). It is important to recall that the set of 77 compounds (19 Class_1 and 58 Class_0 compounds) removed due to their cliff nature ("external cliff set") and the 155 Class_0 compounds removed in the classes balancing process ("external negative set") were also used to evaluate

DOI: 10.1021/acs.jcim.5b00415
J. Chem. Inf. Model. 2015, 55, 2094−2110

the generalization ability of the machine learning classifiers generated.

Structure Codification. Four blocks of DRAGON molecular descriptors were computed for each molecule in the data set. The first block comprises constitutional, physicochemical, and topological information and is denoted as 0-2D. The second block encodes conformational or tridimensional information and is denoted as 3D. The third block includes 45 P_VSA-like descriptors and is denoted as P_VSA. The fourth block comprises all the structural information available in DRAGON 6.0 except the charge descriptors, which is denoted as FULL. These four blocks were saved without considering the pair correlations between descriptors. Four similar blocks of molecular descriptors, excluding descriptors with a pair correlation higher than 0.7, were saved and identified as 0-2D_70, 3D_70, P_VSA_70, and FULL_70, respectively. In this way we favor the development of diverse classifiers based on different structural information and/or influenced or not by the effect of the degree of multicolinearity present on the molecular descriptors matrix. Details are provided in Table S7 (see the Supporting Information).

Feature Selection. Each block of molecular descriptors was subject to an independent process of feature selection. Such a feature selection process relies on a consensus strategy based on seven different ranking feature selection algorithms implemented on WEKA.[89] This consensus ranking feature selection process is described in the Supporting Information.

Modeling and Validation. For each block, the optimal subset of molecular descriptors was evaluated on all the classification algorithms implemented on WEKA 3.6. Contrary to the stage of identification of the optimal subsets of molecular descriptors, the parameters for the classifiers used at this stage were tuned to reach the best performance in a 10-fold cross-validation experiment. The classifiers with best performance were retained and evaluated on the respective test, external evaluation, external cliff, external negative, and real external evaluation sets.

Both the learning and the predictive ability of all the classifiers obtained were assessed by checking the overall and the class-specific performance measures on training (including 10-fold cross-validation), test, and external evaluation sets, respectively.[90] The *accuracy*[91] was used to quantify the overall predictive ability of the classifier. The *Kappa statistic* was used to measure the agreement between predicted and observed categorizations of a data set, while correcting for the agreement that occurs by chance.[90] *Sensitivity* and *specificity* were computed to quantify the class-specific predictive performance of the classifier since they encode its ability to identify positive and negative cases, respectively.

Only those classifiers with values of *accuracy*, *sensitivity*, and *specificity* ≥ 0.6 on training (including 10-fold cross-validation), test, and all external evaluation sets were considered as predictive and reliable and, thus, further advanced to the stage of deriving a consensus classifier.

Consensus Classifier. The selection of the classifiers finally employed for consensus prediction was based on the combined analysis of the diversity between the 81 high-performance classifiers and their respective classification performance. First, we evaluated the pair wise normalized Euclidean distance between the 81 high-performance classifiers, using as reference space the outputs (predicted classes) vector of compounds in the external evaluation set. The result is a classifier's prediction dissimilarity matrix (**ED**) composed of elements ED_{ij}. Accordingly, the pairwise mean classification accuracy was also evaluated as

$$Acc._{ij} = \frac{Acc._i + Acc._j}{2} \quad (6)$$

In eq 6 $Acc._i$ and $Acc._j$ represent the accuracy on the external evaluation set of the classifiers i and j, respectively. Accordingly, $Acc._{ij}$ represents the elements of the mean classification accuracy matrix (**Acc**) of the 81 high-performance classifiers.

Finally, the geometric mean between the respective elements in **ED** and **Acc** is computed to conform a resultant matrix whose elements ${}^{Acc.}ED_{ij}$ act as diversity metric weighted by classifier's accuracy. This matrix is first used to identify and remove those classifiers with identical accuracy and prediction outputs in the external test set (${}^{Acc.}ED_{ij} = 0$). As a consequence, a refined matrix of 59 high-performance and diverse classifiers was obtained. For each classifier we computed its average accuracy-weighted distance with respect to the rest of the 58 classifiers. Only those classifiers with average accuracy-weighted distance values (obtained from each row/column) higher than the overall average accuracy-weighted distance (obtained from the whole 59×59 matrix) were selected to be included in the final consensus classifier.

Virtual Screening Performance. The main goal in a virtual screening effort is to select a subset from a large pool of compounds (typically a compound database or a virtual library) and try to maximize the number of known actives in this subset. That is, to select the most "enriched" subset possible. Several enrichment metrics have been proposed in the literature to measure the enrichment ability of a VS protocol.[56] In this work, we used some of the most extended metrics, which are detailed in the Supporting Information.

Molecular Docking Simulation. A computer model to study the stacking of G-quadruplex DNA was built. The X-ray crystal structure of the bimolecular G4 DNA cocrystallized with BRACO19 (PDB code: 3CE5) was taken from the Protein Data Bank and used as the initial model. The DNA structure was prepared using the Protein Preparation Wizard tool[92] in Maestro[93] (Schrodinger Suite). Missing hydrogen atoms were added to the structure followed by local minimization. The 3D ligand structures were built with the LigPrep[94] program accessible from Maestro GUI, retaining the chirality specified in the SMILES codes. The Epik[95] program was used to generate the most probable ionized and tautomerized structures within a pH range of 7.0 ± 1.0. Molecular docking simulations were carried out using GOLD (Version 5.1),[59,60] choosing GoldScore as fitness function. GoldScore is made up of four components that account for protein ligand binding energy: protein−ligand hydrogen bond (HB) energy (external HB), protein−ligand van der Waals (VDW) energy (external VDW), ligand internal VDW energy (internal VDW), and ligand torsional strain energy (internal torsion). The parameters used in the fitness function (HB energies, atom radii and polarizabilities, torsion potentials, HB directionalities, and so forth) were taken from the GOLD parameter file. GOLD was set to generate 10 docking poses per molecule within a sphere of 15 Å radius centered on the centroid atom of the cocrystallized ligand.

Docking studies were executed on the basis of the recommendations proposed by Haider and Neidle.[96] Table S8 (see the Supporting Information) shows the values of variation in melting temperature (ΔTm) obtained from the same experimental conditions for 1 μM of each of the five reference G4

DOI: 10.1021/acs.jcim.5b00415
J. Chem. Inf. Model. 2015, 55, 2094−2110

stabilizers (BRACO-19, Telomestatin, 360A, Piper, TMPyP4) used for the calibration of the docking protocol.

Experimental Details. *Oligonucleotides.* Oligodeoxynucleotide probes were synthesized by Eurogentec (Belgium). All concentrations are expressed in strand molarity and were determined using a nearest-neighbor approximation for the absorption coefficients of the unfolded species.[97] **F21T** is a 21-nucleotide human telomeric sequence d-$G_3(T_2AG_3)_3$ modified with 6-carboxyfluorescein (FAM) at the 5′ end and tetramethylrhodamine (TAMRA) at the 3′ end.

Fluorescence Resonance Energy Transference (FRET) Melting Studies. The F21T oligonucleotide is used as a probe for the G-quadruplex formation of the human telomeric sequence.[98] A real-time PCR apparatus (MX3005P, Stratagene) was used to record the fluorescence of 96 samples during a thermal denaturation process. The experiments were performed at concentrations of 5 μM of the ligand and 0.2 μM of **F21T**. The "K^+" conditions correspond to 10 mM of lithium cacodylate (pH 7.2), 10 mM of potassium chloride, and 90 mM of lithium chloride, while the "Na^+" conditions were 10 mM of lithium cacodylate (pH 7.2) and 100 mM of sodium chloride. In K^+ conditions, LiCl was added in order to approach the physiological ionic strength without further stabilizing the G-quadruplex. All experiments were performed in duplicate on at least 3 separate plates. The emission of fluorescein was normalized between 0 and 1, and the $T_{1/2}$ was defined as the temperature for which the normalized emission is 0.5.

Cell and Culture Conditions. HeLa cells and normal fibroblasts (WI38) (American Type Culture Collection, Manassas, VA) were maintained at 37 °C in a humidified incubator with 5% of CO2 in Dulbecco's modified Eagle's medium supplemented with 10% of heat-inactivated fetal bovine serum (FBS), 100 units/mL of penicillin, and 100 μg/mL of streptomycin (Invitrogen, Carlsbad, CA). The growth media was changed every 2−3 days per week and subcultured when 80% confluent.

Chemicals and Reagents. Dulbecco's modified Eagle's medium, fetal bovine serum, penicillin, streptomycin, and trypsin/EDTA were purchased from Invitrogen (Carlsbad, CA). Test compounds were purchased from Life Chemicals (http://www.lifechemicals.com) and Asinex (http://www.asinex.com), and dimethyl sulfoxide (DMSO) was purchased from Sigma Chemical (St. Louis, MO). The test compounds were dissolved at 10 mM in 100% DMSO and stored at −20 °C in the dark. Further dilutions were made in water.

Protein Extraction for Telomerase Activity. The HeLa cells in exponential phase of growth (approximately 5×10^6 cells) were harvested using trypsin-EDTA (Gibco/Invitrogen, Carlsbad, CA) and pelleted by centrifugation at 5000 × *g* for 3 min at 4 °C. Cell pellets were washed with phosphate-buffered saline (PBS), centrifuged, and then resuspended for 30 min in 200 μL of ice-cold 1 × CHAPS lysis buffer.

Protein lysates were centrifuged at 12 000 rpm for 30 min at 4 °C, and the supernatant was collected. The protein concentration was measured by the Bradford method.[99] The final extracts were diluted to a concentration of 1 mg/mL protein with lysis buffer and immediately stored in aliquots at −80 °C.

Telomerase Assay (RTQ-TRAP). The ability of G4 ligands to inhibit telomerase in an extract-based assay was assessed by a modified real-time quantitative telomeric repeat amplification protocol (RTQ-TRAP assay).[100−102] The RTQ-TRAP assay was performed in two steps: (i) telomerase-mediated extension of the forward primer (TS: 5′-AATCCGTCGAGCAGAGTT-3′), (Oswel, Southampton, U.K.)[103] and (ii) a PCR amplification step.

In details, 1 μg of the telomerase extract was incubated for 20 min at 25 °C in the reaction mix with the query compounds at final concentrations of 1 and 10 μM. The total volume of the reaction mixture was 20 μL containing 10 μL of LightCycler 480 SYBR Green I Master (Roche Diagnostics) mix, 5 pmol TS primer, 1.25 pmol ACX primer (5′-GCGCGGCTTACCCTTAC-CCTTACCCTAACC-3′), 2 μL of protein extract (500 ng/μL), and 2 μL of test sample.

As a second step, the real-time TRAP was performed in a LightCycler System 480 (Roche Diagnostics, Mannheim, Germany). In order to activate the modified Taq polymerase and inactivate telomerase activity, the PCR was initiated at 95 °C for 5 min, followed by 40 cycles (95 °C for 30 s, 55 °C for 30 s, and 72 °C for 30 s) and one cycle corresponding to the melting curve.[104]

The following controls were included: (i) telomerase positive extract control (2 μL of HeLa protein extract); (ii) lysis buffer control (2 μL of CHAPS Lysis Buffer); (iii) vehicle control (2 μL of 0.1% DMSO plus cell extract); and (iv) no template control (2 μL of Nuclease Free Water). In addition, a standard curve was generated using serial dilutions of telomerase-positive HeLa cell extracts (10000, 1000, 100, 10, 1, and 0.1 cells).

The presence of internal telomerase assay standard (ITAS) allowed us to determine the selectivity of potential inhibitors regarding Taq polymerase inhibition. The TRAP internal control, TSNT (5′-AATCCGTCGAGCAGAGTTAAAAG-GCCGAGAAGCGAT-3′) at 0.01 amol was amplified by 7.5 pmol of primer TS and its own dedicated return primer, NT (5′-ATCGCTTCTCGGCCTTTT-3′), respectively.

Each sample was analyzed in duplicates in three independent experiments. The fluorescence intensity threshold cycle (*Ct*) values were determined from semilog amplification plots (log increase in fluorescence versus cycle number) and compared with standard curves generated from serial dilutions of telomerase-positive HeLa cell extracts.

The telomerase activity was comparatively assessed based on *Ct*, where values greater than 35 were considered false positives due to primer-dimers artifacts. The inhibition of the telomerase activity was determined as the percentage of remaining telomerase activity:

$$Activity\ Remaining\ (\%) = \frac{1}{1.18^{\Delta Ct}} \times 100 \tag{7}$$

$$\Delta Ct = Ct_{Treatment} - Ct_{Control} \tag{8}$$

In eq 7 $Ct_{Treatment}$ and $Ct_{Control}$ are the threshold cycle values of cells treated with the candidate ligand and DMSO treated cells, respectively.

Cellular Proliferation Assay. Short-term growth inhibition in carcinoma and normal cells was measured using the Sulforhodamine B (SRB) assay as described previously.[105] The results were determined by three independent experiments. The absorbance of untreated cells was considered as 100%. The percentage of cell viability was determined from the mean absorbance at 540 nm for each ligand concentration and was expressed as the percentage of the control untreated well absorbance.[106]

Statistical Analysis. All experiments were repeated three times. Data are expressed as a mean standard deviation. Data were analyzed using Model I ANOVAs with SPSS for windows, version 19.0.[107] The relationship of interest was ligand vs control; therefore, a Dunnett's test was used to determine statistical significance of the results at p-values <0.05. Significant changes were assessed by a *t* test for unpaired data at p-values <0.05.

DOI: 10.1021/acs.jcim.5b00415
J. Chem. Inf. Model. 2015, 55, 2094−2110

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jcim.5b00415.

Selection of the arbitrary $SALI_{ij}$ value to be used as cutoff; classes balancing; consensus ranking feature selection; virtual screening performance; Tables S1−S11; and Figures S1−S5 (PDF)

■ AUTHOR INFORMATION

Corresponding Authors

*E-mail: gmailkelcm@yahoo.es.

*Phone: 351-220402502. Fax: 351-220402659. E-mail: fborges@fc.up.pt.

Funding

Postdoctoral grant [SFRH/BPD/90673/2012] was financed by the FCT − Fundação para a Ciência e a Tecnologia, Portugal, cofinanced by the European Social Fund. Project [1-A1/036687/11] was financed by the AECID − Agencia Española de Cooperación Iberoamericana para el Desarrollo.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

M.C.-M. acknowledges the postdoctoral grant [SFRH/BPD/90673/2012] financed by the FCT − Fundação para a Ciência e a Tecnologia, Portugal, cofinanced by the European Social Fund. D.C.-G., G.P.-M., and M.A.C.-P. acknowledge the financial support of Agencia Española de Cooperación Iberoamericana para el Desarrollo (AECID) for the project [1-A1/036687/11] Montaje de un Laboratorio de Quimica Computacional, con Fines Académicos y Científicos, para el Diseño de Potenciales Candidatos a Fármacos, en Enfermedades de Alto Impacto Social.

■ REFERENCES

(1) Riou, J. F.; Guittat, L.; Mailliet, P.; Laoui, A.; Renou, E.; Petitgenet, O.; Megnin-Chanet, F.; Helene, C.; Mergny, J. L. Cell Senescence and Telomere Shortening Induced by a New Series of Specific G-Quadruplex DNA Ligands. *Proc. Natl. Acad. Sci. U. S. A.* **2002**, *99*, 2672−2677.

(2) Feng, J.; Funk, W. D.; Wang, S. S.; Weinrich, S. L.; Avilion, A. A.; Chiu, C. P.; Adams, R. R.; Chang, E.; Allsopp, R. C.; Yu, J.; Le, S.; West, M. D.; Harley, C. B.; Andrews, W. H.; Greider, C. W.; Villeponteau, B. The Rna Component of Human Telomerase. *Science* **1995**, *269*, 1236−1241.

(3) Sprouse, A. A.; Steding, C. E.; Herbert, B. S. Pharmaceutical Regulation of Telomerase and Its Clinical Potential. *J. Cell. Mol. Med.* **2012**, *16*, 1−7.

(4) Agrawal, A.; Dang, S.; Gabrani, R. Recent Patents on Anti-Telomerase Cancer Therapy. *Recent Pat. Anti-Cancer Drug Discovery* **2012**, *7*, 102−117.

(5) Xu, Y.; He, K.; Goldkorn, A. Telomerase Targeted Therapy in Cancer and Cancer Stem Cells. *Clin. Adv. Hematol. Oncol.* **2011**, *9*, 442−455.

(6) Shay, J. W.; Wright, W. E. Role of Telomeres and Telomerase in Cancer. *Semin. Cancer Biol.* **2011**, *21*, 349−353.

(7) Castillo-Gonzalez, D.; Perez-Machado, G.; Guedin, A.; Mergny, J. L.; Cabrera-Perez, M. A. FDA-Approved Drugs Selected Using Virtual Screening Bind Specifically to G-Quadruplex DNA. *Curr. Pharm. Des.* **2013**, *19*, 2164−2173.

(8) Incles, C. M.; Schultes, C. M.; Kempski, H.; Koehler, H.; Kelland, L. R.; Neidle, S. A G-Quadruplex Telomere Targeting Agent Produces P16-Associated Senescence and Chromosomal Fusions in Human Prostate Cancer Cells. *Mol. Cancer. Ther.* **2004**, *3*, 1201−1206.

(9) Moore, M. J.; Schultes, C. M.; Cuesta, J.; Cuenca, F.; Gunaratnam, M.; Tanious, F. A.; Wilson, W. D.; Neidle, S. Trisubstituted Acridines as G-Quadruplex Telomere Targeting Agents. Effects of Extensions of the 3,6- and 9-Side Chains on Quadruplex Binding, Telomerase Activity, and Cell Proliferation. *J. Med. Chem.* **2006**, *49*, 582−599.

(10) Gunaratnam, M.; Neidle, S. An Evaluation Cascade for G-Quadruplex Telomere Targeting Agents in Human Cancer Cells. *Methods Mol. Biol.* **2010**, *613*, 303−313.

(11) Siddiqui-Jain, A.; Grand, C. L.; Bearss, D. J.; Hurley, L. H. Direct Evidence for a G-Quadruplex in a Promoter Region and Its Targeting with a Small Molecule to Repress C-Myc Transcription. *Proc. Natl. Acad. Sci. U. S. A.* **2002**, *99*, 11593−11598.

(12) Lim, K. W.; Lacroix, L.; Yue, D. J.; Lim, J. K.; Lim, J. M.; Phan, A. T. Coexistence of Two Distinct G-Quadruplex Conformations in the Htert Promoter. *J. Am. Chem. Soc.* **2010**, *132*, 12331−12342.

(13) Lemarteleur, T.; Gomez, D.; Paterski, R.; Mandine, E.; Mailliet, P.; Riou, J. F. Stabilization of the C-Myc Gene Promoter Quadruplex by Specific Ligands' Inhibitors of Telomerase. *Biochem. Biophys. Res. Commun.* **2004**, *323*, 802−808.

(14) Cogoi, S.; Paramasivam, M.; Filichev, V.; Geci, I.; Pedersen, E. B.; Xodo, L. E. Identification of a New G-Quadruplex Motif in the Kras Promoter and Design of Pyrene-Modified G4-Decoys with Antiproliferative Activity in Pancreatic Cancer Cells. *J. Med. Chem.* **2009**, *52*, 564−568.

(15) Verma, A.; Halder, K.; Halder, R.; Yadav, V. K.; Rawal, P.; Thakur, R. K.; Mohd, F.; Sharma, A.; Chowdhury, S. Genome-Wide Computational and Expression Analyses Reveal G-Quadruplex DNA Motifs as Conserved Cis-Regulatory Elements in Human and Related Species. *J. Med. Chem.* **2008**, *51*, 5641−5649.

(16) Laronze-Cochard, M.; Kim, Y. M.; Brassart, B.; Riou, J. F.; Laronze, J. Y.; Sapi, J. Synthesis and Biological Evaluation of Novel 4,5-Bis(Dialkylaminoalkyl)-Substituted Acridines as Potent Telomeric G-Quadruplex Ligands. *Eur. J. Med. Chem.* **2009**, *44*, 3880−3888.

(17) Corey, D. R. Telomeres and Telomerase: From Discovery to Clinical Trials. *Chem. Biol.* **2009**, *16*, 1219−1223.

(18) Huang, F. C.; Chang, C. C.; Wang, J. M.; Chang, T. C.; Lin, J. J. Induction of Senescence in Cancer Cells by the G-Quadruplex Stabilizer, BMVC4, Is Independent of Its Telomerase Inhibitory Activity. *Br. J. Pharmacol.* **2012**, *167*, 393−406.

(19) Zhou, Q.; Li, L.; Xiang, J.; Tang, Y.; Zhang, H.; Yang, S.; Li, Q.; Yang, Q.; Xu, G. Screening Potential Antitumor Agents from Natural Plant Extracts by G-Quadruplex Recognition and NMR Methods. *Angew. Chem., Int. Ed.* **2008**, *47*, 5590−5592.

(20) Zhou, Q.; Li, L.; Xiang, J.; Sun, H.; Tang, Y. Fast Screening and Structural Elucidation of G-Quadruplex Ligands from a Mixture Via G-Quadruplex Recognition and NMR Methods. *Biochimie* **2009**, *91*, 304−308.

(21) Pinto, I. G.; Guilbert, C.; Ulyanov, N. B.; Stearns, J.; James, T. L. Discovery of Ligands for a Novel Target, the Human Telomerase Rna, Based on Flexible-Target Virtual Screening and NMR. *J. Med. Chem.* **2008**, *51*, 7205−7215.

(22) Cosconati, S.; Marinelli, L.; Trotta, R.; Virno, A.; Mayol, L.; Novellino, E.; Olson, A. J.; Randazzo, A. Tandem Application of Virtual Screening and Nmr Experiments in the Discovery of Brand New DNA Quadruplex Groove Binders. *J. Am. Chem. Soc.* **2009**, *131*, 16336−16337.

(23) Chen, S. B.; Tan, J. H.; Ou, T. M.; Huang, S. L.; An, L. K.; Luo, H. B.; Li, D.; Gu, L. Q.; Huang, Z. S. Pharmacophore-Based Discovery of Triaryl-Substituted Imidazole as New Telomeric G-Quadruplex Ligand. *Bioorg. Med. Chem. Lett.* **2011**, *21*, 1004−1009.

(24) Alcaro, S.; Musetti, C.; Distinto, S.; Casatti, M.; Zagotto, G.; Artese, A.; Parrotta, L.; Moraca, F.; Costa, G.; Ortuso, F.; Maccioni, E.; Sissi, C. Identification and Characterization of New DNA G-Quadruplex Binders Selected by a Combination of Ligand and Structure-Based Virtual Screening Approaches. *J. Med. Chem.* **2013**, *56*, 843−855.

(25) Artese, A.; Costa, G.; Ortuso, F.; Parrotta, L.; Alcaro, S. Identification of New Natural DNA G-Quadruplex Binders Selected by

a Structure-Based Virtual Screening Approach. *Molecules* **2013**, *18*, 12051−12070.
(26) Rocca, R.; Moraca, F.; Costa, G.; Alcaro, S.; Distinto, S.; Maccioni, E.; Ortuso, F.; Artese, A.; Parrotta, L. Structure-Based Virtual Screening of Novel Natural Alkaloid Derivatives as Potential Binders of H-Telo and C-Myc DNA G-Quadruplex Conformations. *Molecules* **2015**, *20*, 206−223.
(27) Young, D.; Martin, T.; Venkatapathy, R.; Harten, P. Are the Chemical Structures in Your QSAR Correct? *QSAR Comb. Sci.* **2008**, *27*, 1337−1345.
(28) Fourches, D.; Muratov, E.; Tropsha, A. Trust, but Verify: On the Importance of Chemical Structure Curation in Cheminformatics and QSAR Modeling Research. *J. Chem. Inf. Model.* **2010**, *50*, 1189−1204.
(29) Nicolotti, O.; Benfenati, E.; Carotti, A.; Gadaleta, D.; Gissi, A.; Mangiatordi, G. F.; Novellino, E. Reach and in Silico Methods: An Attractive Opportunity for Medicinal Chemists. *Drug Discovery Today* **2014**, *19*, 1757−1768.
(30) Todeschini, R.; Consonni, V. *Molecular Descriptors for Chemoinformatics*; Wiley-VCH: Weinheim, Germany, 2009; Vol. *1*.
(31) Breiman, L.; Friedman, J. H.; Olshen, R. A.; Stone, C. J. *Classification and Regression Trees*; Chapman and Hall/CRC Press: Boca Ratón, FL, 1984.
(32) Burden, F. R.; Ford, M. G.; Whitley, D. C.; Winkler, D. A. Use of Automatic Relevance Determination in QSAR Studies Using Bayesian Neural Networks. *J. Chem. Inf. Model.* **2000**, *40*, 1423−1430.
(33) Eckert, H.; Bajorath, J. Molecular Similarity Analysis in Virtual Screening: Foundations, Limitations and Novel Approaches. *Drug Discovery Today* **2007**, *12*, 225−233.
(34) Jaworska, J.; Nikolova-Jeliazkova, N.; Aldenberg, T. QSAR Applicability Domain Estimation by Projection of the Training Set in Descriptor Space: A Review. *ATLA, Altern. Lab. Anim.* **2005**, *33*, 445−459.
(35) Sushko, I.; Novotarskyi, S.; Körner, R.; Pandey, A. K.; Cherkasov, A.; Li, J.; Gramatica, P.; Hansen, K.; Schroeter, T.; Müller, K.-R.; Xi, L.; Liu, H.; Yao, X.; Öberg, T.; Hormozdiari, F.; Dao, P.; Sahinalp, C.; Todeschini, R.; Polishchuk, P.; Artemenko, A.; Kuz'min, V.; Martin, T. M.; Young, D. M.; Fourches, D.; Muratov, E.; Tropsha, A.; Baskin, I.; Horvath, D.; Marcou, G.; Muller, C.; Varnek, A.; Prokopenko, V. V.; Tetko, I. V. Applicability Domains for Classification Problems: Benchmarking of Distance to Models for Ames Mutagenicity Set. *J. Chem. Inf. Model.* **2010**, *50*, 2094−2111.
(36) Tropsha, A. Best Practices for QSAR Model Development, Validation, and Exploitation. *Mol. Inf.* **2010**, *29*, 476−488.
(37) Johnson, M. A.; Maggiora, G. M. *Concepts and Applications of Molecular Similarity*; John Wiley & Sons: New York, NY, 1990.
(38) Peng, H.; Long, F.; Ding, C. Feature Selection Based on Mutual Information: Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy. *IEEE Trans. Pattern Anal. Machine Intel.* **2005**, *27*, 1226−1238.
(39) Willett, P. Similarity-Based Virtual Screening Using 2d Fingerprints. *Drug Discovery Today* **2006**, *11*, 1046−1053.
(40) Cruz-Monteagudo, M.; Medina-Franco, J. L.; Perez-Castillo, Y.; Nicolotti, O.; Cordeiro, M. N.; Borges, F. Activity Cliffs in Drug Discovery: Dr Jekyll or Mr Hyde? *Drug Discovery Today* **2014**, *19*, 1069−1080.
(41) Méndez-Lucio, O.; Pérez-Villanueva, J.; Castillo, R.; Medina-Franco, J. L. Identifying Activity Cliff Generators of PPAR Ligands Using SAS Maps. *Mol. Inf.* **2012**, *31*, 837−846.
(42) Guha, R. The Ups and Downs of Structure-Activity Landscapes. *Methods Mol. Biol.* **2010**, *672*, 101−117.
(43) Polikar, R. Ensemble Based Systems in Decision Making. *IEEE Circuit Syst. Mag.* **2006**, *6*, 21−44.
(44) Zhang, L.; Fourches, D.; Sedykh, A.; Zhu, H.; Golbraikh, A.; Ekins, S.; Clark, J.; Connelly, M. C.; Sigal, M.; Hodges, D.; Guiguemde, A.; Guy, R. K.; Tropsha, A. Discovery of Novel Antimalarial Compounds Enabled by QSAR-Based Virtual Screening. *J. Chem. Inf. Model.* **2013**, *53*, 475−492.
(45) Kuncheva, L. I. *Combining Pattern Classifiers, Methods and Algorithms*; Wiley Interscience: New York, NY, 2004.
(46) Guha, R.; Van Drie, J. H. Structure−Activity Landscape Index: Identifying and Quantifying Activity Cliffs. *J. Chem. Inf. Model.* **2008**, *48*, 646−658.
(47) Scior, T.; Medina-Franco, J. L.; Do, Q.-T.; Martínez-Mayorga, K.; Yunes Rojas, J. A.; Bernard, P. How to Recognize and Workaround Pitfalls in QSAR Studies: A Critical Review. *Curr. Med. Chem.* **2009**, *16*, 4297−4313.
(48) Witten, I. H.; Frank, E. *Data Mining: Practical Machine Learning Tools and Techniques*, 2nd ed.; Morgan Kaufmann: San Francisco, CA, 2005.
(49) Japkowicz, N. *The Class Imbalance Problem: Significance and Strategies. In International Conference on Artificial Intelligence*; Arabnia, H. R., de la Fuente, D., Kozerenko, E. B., Olivas, J. A., Chang, R., LaMonica, P. M., Liuzzi, R. A.; G, S. A. M., Eds.; CSREA Press: Las Vegas, NV, 2000.
(50) Japkowicz, N. *Learning from Imbalanced Data Sets: A Comparison of Various Solutions. In AAAÍ2000 Workshop on Learning from Imbalanced Data Sets*; Holte, R., Japkowicz, N., Ling, C., Matwin, S., Eds.; AAAI Press: Austin, TX, 2000.
(51) Solov'ev, V. P.; Varnek, A. *EdiSDF* [For Windows], version 5.03. http://infochim.u-strasbg.fr/recherche/isida/index.php (accessed Sept 1, 2015).
(52) Ruggiu, F.; Marcou, G.; Varnek, A.; Horvath, D. ISIDA Property-Labelled Fragment Descriptors. *Mol. Inf.* **2010**, *29*, 855−868.
(53) Varnek, A.; Fourches, D.; Horvath, D.; Klimchuk, O.; Gaudin, C.; Vayer, P.; Solov'ev, V.; Hoonakker, F.; Tetko, I.; G, M. ISIDA - Platform for Virtual Screening Based on Fragment and Pharmacophoric Descriptors. *Curr. Comput.-Aided Drug Des.* **2008**, *4*, 191−198.
(54) Kovatcheva, A.; Golbraikh, A.; Oloff, S.; Xiao, Y. D.; Zheng, W.; Wolschann, P.; Buchbauer, G.; Tropsha, A. Combinatorial QSAR of Ambergris Fragrance Compounds. *J. Chem. Inf. Model.* **2004**, *44*, 582−595.
(55) Perez-Castillo, Y.; Cruz-Monteagudo, M.; Lazar, C.; Taminau, J.; Froeyen, M.; Cabrera-Perez, M. A.; Nowe, A. Toward the Computer-Aided Discovery of FabH Inhibitors. Do Predictive Qsar Models Ensure High Quality Virtual Screening Performance? *Mol. Diversity* **2014**, *18*, 637−654.
(56) Truchon, J. F.; Bayly, C. I. Evaluating Virtual Screening Methods: Good and Bad Metrics for the "Early Recognition" Problem. *J. Chem. Inf. Model.* **2007**, *47*, 488−508.
(57) Kirchmair, J.; Markt, P.; Distinto, S.; Wolber, G.; Langer, T. Evaluation of the Performance of 3d Virtual Screening Protocols: Rmsd Comparisons, Enrichment Assessments, and Decoy Selection−What Can We Learn from Earlier Mistakes? *J. Comput.-Aided Mol. Des.* **2008**, *22*, 213−228.
(58) Huang, N.; Shoichet, B. K.; Irwin, J. J. Benchmarking Sets for Molecular Docking. *J. Med. Chem.* **2006**, *49*, 6789−6801.
(59) Yap, C. W. Padel-Descriptor: An Open Source Software to Calculate Molecular Descriptors and Fingerprints. *J. Comput. Chem.* **2011**, *32*, 1466−1474.
(60) MathWorks. *MatLab* [For Windows], version 7.2. http://www.mathworks.com/products/matlab/ (accessed Sept 1, 2015).
(61) Jones, G.; Willett, P.; Glen, R. C.; Leach, A. R.; Taylor, R. Development and Validation of a Genetic Algorithm for Flexible Docking. *J. Mol. Biol.* **1997**, *267*, 727−748.
(62) Jones, G.; Willett, P.; Glen, R. C. Molecular Recognition of Receptor Sites Using a Genetic Algorithm with a Description of Desolvation. *J. Mol. Biol.* **1995**, *245*, 43−53.
(63) Schrödinger. *Canvas* [For Windows], version 1.6. http://www.schrodinger.com/ (accessed Sept 1, 2015).
(64) Darby, R. A.; Sollogoub, M.; McKeen, C.; Brown, L.; Risitano, A.; Brown, N.; Barton, C.; Brown, T.; Fox, K. R. High Throughput Measurement of Duplex, Triplex and Quadruplex Melting Curves Using Molecular Beacons and a Lightcycler. *Nucleic Acids Res.* **2002**, *30*, e39.
(65) De Cian, A.; Guittat, L.; Kaiser, M.; Sacca, B.; Amrane, S.; Bourdoncle, A.; Alberti, P.; Teulade-Fichou, M. P.; Lacroix, L.; Mergny, J. L. Fluorescence-Based Melting Assays for Studying Quadruplex Ligands. *Methods* **2007**, *42*, 183−195.

DOI: 10.1021/acs.jcim.5b00415
J. Chem. Inf. Model. 2015, 55, 2094−2110

(66) Perry, P. J.; Jenkins, T. C. DNA Tetraplex-Binding Drugs: Structure-Selective Targeting Is Critical for Antitumour Telomerase Inhibition. *Mini-Rev. Med. Chem.* **2001**, *1*, 31−41.
(67) Cuesta, J.; Read, M. A.; Neidle, S. The Design of G-Quadruplex Ligands as Telomerase Inhibitors. *Mini-Rev. Med. Chem.* **2003**, *3*, 11−21.
(68) Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; Overington, J. P. ChEMBL: A Large-Scale Bioactivity Database for Drug Discovery. *Nucleic Acids Res.* **2012**, *40*, D1100−1107.
(69) Bento, A. P.; Gaulton, A.; Hersey, A.; Bellis, L. J.; Chambers, J.; Davies, M.; Kruger, F. A.; Light, Y.; Mak, L.; McGlinchey, S.; Nowotka, M.; Papadatos, G.; Santos, R.; Overington, J. P. The ChEMBL Bioactivity Database: An Update. *Nucleic Acids Res.* **2014**, *42*, D1083−1090.
(70) Stumpfe, D.; Bajorath, J. Exploring Activity Cliffs in Medicinal Chemistry. *J. Med. Chem.* **2012**, *55*, 2932−2942.
(71) Bickerton, G. R.; Paolini, G. V.; Besnard, J.; Muresan, S.; Hopkins, A. L. Quantifying the Chemical Beauty of Drugs. *Nat. Chem.* **2012**, *4*, 90−98.
(72) Ghose, A. K.; Viswanadhan, V. N.; Wendoloski, J. J. A Knowledge-Based Approach in Designing Combinatorial or Medicinal Chemistry Libraries for Drug Discovery. 1. A Qualitative and Quantitative Characterization of Known Drug Databases. *J. Comb. Chem.* **1999**, *1*, 55−68.
(73) Oprea, T. I. Property Distribution of Drug-Related Chemical Databases. *J. Comput.-Aided Mol. Des.* **2000**, *14*, 251−264.
(74) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and Computational Approaches to Estimate Solubility and Permeability in Drug Discovery and Development Settings. *Adv. Drug Delivery Rev.* **2001**, *46*, 3−26.
(75) Veber, D. F.; Johnson, S. R.; Cheng, H. Y.; Smith, B. R.; Ward, K. W.; Kopple, K. D. Molecular Properties That Influence the Oral Bioavailability of Drug Candidates. *J. Med. Chem.* **2002**, *45*, 2615−2623.
(76) Sushko, I.; Salmina, E.; Potemkin, V. A.; Poda, G.; Tetko, I. V. Toxalerts: A Web Server of Structural Alerts for Toxic Chemicals and Compounds with Potential Adverse Reactions. *J. Chem. Inf. Model.* **2012**, *52*, 2310−2316.
(77) Lounkine, E.; Wawer, M.; Wassermann, A. M.; Bajorath, J. Saranea: A Freely Available Program to Mine Structure-Activity and Structure-Selectivity Relationship Information in Compound Data Sets. *J. Chem. Inf. Model.* **2010**, *50*, 68−78.
(78) OpenEye. *OpenBabel: The Opensource Chemistry Toolbox* [For Windows], version 2.2.99. http://openbabel.org (accessed Sept 1, 2015).
(79) ChemAxon. *JChem for Excel* [For Windows], version 5.10.2.725. http://www.chemaxon.com (accessed Sept 1, 2015).
(80) ChemAxon. *Standardizer* [For Windows], version 5.10.2. http://www.chemaxon.com (accessed Sept 1, 2015).
(81) Talete. *DRAGON* [For Windows], version 6.0. http://www.talete.mi.it/ (accessed Sept 1, 2015).
(82) StatSoft. *STATISTICA* [For Windows], version 8.0. www.statsoft.com (accessed Sept 1, 2015).
(83) Kubinyi, H. *Virtual Screening - the Road to Success. In XIX International Symposium on Medicinal Chemistry*; European Federation for Medicinal Chemistry: Istanbul, Turkey, 2006.
(84) Peduto, A.; Pagano, B.; Petronzi, C.; Massa, A.; Esposito, V.; Virgilio, A.; Paduano, F.; Trapasso, F.; Fiorito, F.; Florio, S.; Giancola, C.; Galeone, A.; Filosa, R. Design, Synthesis, Biophysical and Biological Studies of Trisubstituted Naphthalimides as G-Quadruplex Ligands. *Bioorg. Med. Chem.* **2011**, *19*, 6419−6429.
(85) Li, Z.; Tan, J. H.; He, J. H.; Long, Y.; Ou, T. M.; Li, D.; Gu, L. Q.; Huang, Z. S. Disubstituted Quinazoline Derivatives as a New Type of Highly Selective Ligands for Telomeric G-Quadruplex DNA. *Eur. J. Med. Chem.* **2012**, *47*, 299−311.
(86) Peng, D.; Tan, J. H.; Chen, S. B.; Ou, T. M.; Gu, L. Q.; Huang, Z. S. Bisaryldiketene Derivatives: A New Class of Selective Ligands for C-Myc G-Quadruplex DNA. *Bioorg. Med. Chem.* **2010**, *18*, 8235−8242.
(87) Folini, M.; Pivetta, C.; Zagotto, G.; De Marco, C.; Palumbo, M.; Zaffaroni, N.; Sissi, C. Remarkable Interference with Telomeric Function by a G-Quadruplex Selective Bisantrene Regioisomer. *Biochem. Pharmacol.* **2010**, *79*, 1781−1790.
(88) Chen, C. Y.; Wang, Q.; Liu, J. Q.; Hao, Y. H.; Tan, Z. Contribution of Telomere G-Quadruplex Stabilization to the Inhibition of Telomerase-Mediated Telomere Extension by Chemical Ligands. *J. Am. Chem. Soc.* **2011**, *133*, 15036−15044.
(89) Holmes, G.; Donkin, A.; Witten, I. H. *Waikato Environment for Knowledge Analysis (WEKA)*, 3.6.1. http://www.cs.waikato.ac.nz/ml/weka/ (accessed Sept 1, 2015).
(90) Witten, I. H.; Frank, E. Chapter 5: Credibility: Evaluating What's Been Learned. *Data Mining: Practical Machine Learning Tools and Techniques*; Gray, J., Ed.; Morgan Kaufman: San Francisco, CA, 2005; pp 143−186.
(91) Cannon, E. O.; Amini, A.; Bender, A.; Sternberg, M. J.; Muggleton, S. H.; Glen, R. C.; Mitchell, J. B. Support Vector Inductive Logic Programming Outperforms the Naive Bayes Classifier and Inductive Logic Programming for the Classification of Bioactive Chemical Compounds. *J. Comput.-Aided Mol. Des.* **2007**, *21*, 269−280.
(92) Schrödinger. *Protein Preparation Wizard* [For Windows], version 2.4. http://www.schrodinger.com/ (accessed Sept 1, 2015).
(93) Schrödinger. *Maestro* [For Windows], version 9.4. http://www.schrodinger.com/ (accessed Sept 1, 2015).
(94) Schrödinger. *LigPrep* [For Windows], version 2.6. http://www.schrodinger.com/ (accessed Sept 1, 2015).
(95) Schrödinger. *Epik* [For Windows], version 2.4. http://www.schrodinger.com/ (accessed Sept 1, 2015).
(96) Haider, S.; Neidle, S. Molecular Modeling and Simulation of G-Quadruplexes and Quadruplex-Ligand Complexes. *Methods Mol. Biol.* **2010**, *608*, 17−37.
(97) Cantor, C. R.; Warshaw, M. M.; Shapiro, H. Oligonucleotide Interactions. 3. Circular Dichroism Studies of the Conformation of Deoxyoligonucleotides. *Biopolymers* **1970**, *9*, 1059−1077.
(98) Mergny, J. L.; Maurizot, J. C. Fluorescence Resonance Energy Transfer as a Probe for G-Quartet Formation by a Telomeric Repeat. *ChemBioChem* **2001**, *2*, 124−132.
(99) Bradford, M. M. A Rapid and Sensitive Method for the Quantitation of Microgram Quantities of Protein Utilizing the Principle of Protein-Dye Binding. *Anal. Biochem.* **1976**, *72*, 248−254.
(100) Gomez, D.; Mergny, J. L.; Riou, J. F. Detection of Telomerase Inhibitors Based on G-Quadruplex Ligands by a Modified Telomeric Repeat Amplification Protocol Assay. *Cancer Res.* **2002**, *62*, 3365−3368.
(101) Wege, H.; Chui, M. S.; Le, H. T.; Tran, J. M.; Zern, M. A. Sybr Green Real-Time Telomeric Repeat Amplification Protocol for the Rapid Quantification of Telomerase Activity. *Nucleic Acids Res.* **2003**, *31*, E3−3.
(102) Hou, M.; Xu, D.; Bjorkholm, M.; Gruber, A. Real-Time Quantitative Telomeric Repeat Amplification Protocol Assay for the Detection of Telomerase Activity. *Clin. Chem.* **2001**, *47*, 519−524.
(103) Shim, W. Y.; Park, K. H.; Jeung, H. C.; Kim, Y. T.; Kim, T. S.; Hyung, W. J.; An, S. H.; Yang, S. H.; Noh, S. H.; Chung, H. C.; Rha, S. Y. Quantitative Detection of Telomerase Activity by Real-Time Trap Assay in the Body Fluids of Cancer Patients. *Int. J. Mol. Med.* **2005**, *16*, 857−863.
(104) Ohuchida, K.; Mizumoto, K.; Ogura, Y.; Ishikawa, N.; Nagai, E.; Yamaguchi, K.; Tanaka, M. Quantitative Assessment of Telomerase Activity and Human Telomerase Reverse Transcriptase Messenger RNA Levels in Pancreatic Juice Samples for the Diagnosis of Pancreatic Cancer. *Clin. Cancer Res.* **2005**, *11*, 2285−2292.
(105) Gunaratnam, M.; Greciano, O.; Martins, C.; Reszka, A. P.; Schultes, C. M.; Morjani, H.; Riou, J.-F.; Neidle, S. Mechanism of Acridine-Based Telomerase Inhibition and Telomere Shortening. *Biochem. Pharmacol.* **2007**, *74*, 679−689.
(106) Sparapani, S.; Haider, S. M.; Doria, F.; Gunaratnam, M.; Neidle, S. Rational Design of Acridine-Based Ligands with Selectivity for Human Telomeric Quadruplexes. *J. Am. Chem. Soc.* **2010**, *132*, 12263−12272.
(107) IBM. *IBM SPSS Statistics* [For Windows], Version 19.0. http://www-01.ibm.com/ (accessed Sept 1, 2015).

DOI: 10.1021/acs.jcim.5b00415
J. Chem. Inf. Model. 2015, 55, 2094−2110