ACS OMEGA

ACS AUTHORCHOICE

http://pubs.acs.org/journal/acsodf

Article

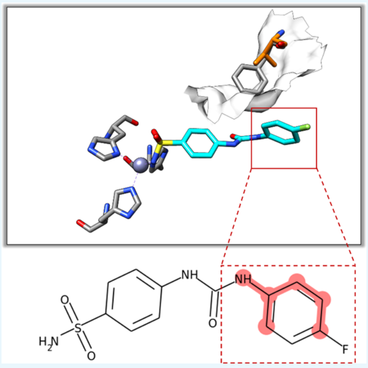
Predicting Isoform-Selective Carbonic Anhydrase Inhibitors via Machine Learning and Rationalizing Structural Features Important for Selectivity

Salvatore Galati, Dimitar Yonchev, Raquel Rodríguez-Pérez, Martin Vogt, Tiziano Tuccinardi,* and Jürgen Bajorath*

Cite This: *ACS Omega* 2021, 6, 4080−4089

Read Online

ACCESS | Metrics & More | Article Recommendations



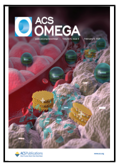
ABSTRACT: Carbonic anhydrases (CAs) catalyze the physiological hydration of carbon dioxide and are among the most intensely studied pharmaceutical target enzymes. A hallmark of CA inhibition is the complexation of the catalytic zinc cation in the active site. Human (*h*) CA isoforms belonging to different families are implicated in a wide range of diseases and of very high interest for therapeutic intervention. Given the conserved catalytic mechanisms and high similarity of many *h*CA isoforms, a major challenge for CA-based therapy is achieving inhibitor selectivity for *h*CA isoforms that are associated with specific pathologies over other widely distributed isoforms such as *h*CA I or *h*CA II that are of critical relevance for the integrity of many physiological processes. To address this challenge, we have attempted to predict compounds that are selective for isoform *h*CA IX, which is a tumor-associated protein and implicated in metastasis, over *h*CA II on the basis of a carefully curated data set of selective and nonselective inhibitors. Machine learning achieved surprisingly high accuracy in predicting *h*CA IX-selective inhibitors. The results were further investigated, and compound features determining successful predictions were identified. These features were then studied on the basis of X-ray structures of *h*CA isoform-inhibitor complexes and found to include substructures that explain compound selectivity. Our findings lend credence to selectivity predictions and indicate that the machine learning models derived herein have considerable potential to aid in the identification of new *h*CA IX-selective compounds.

1. INTRODUCTION

Human carbonic anhydrases (*h*CAs) are metalloenzymes that catalyze a reversible hydration of carbon dioxide producing bicarbonate with the release of a proton.[1] Among the eight genetically distinct CA families (α, β, γ, δ, ζ, η, θ, and ι), 15 α-CA isoforms are known in humans, i.e., *h*CA I−*h*CA XIV, which include two V-type isoforms (*h*CA VA and *h*CA VB) that differ in cellular distribution and functions. These metalloenzymes are involved in numerous physiological processes such as pH regulation, CO_2 homeostasis, bone resorption, and gluconeogenesis.[2] Due to the wide spectrum of physiological roles played by CAs, they have been shown to be involved in different diseases such as glaucoma, obesity, osteoporosis, various types of tumors, epilepsy, and neuropathic pain. Therefore, *h*CAs are regarded as important therapeutic targets, and *h*CA modulators are recognized as promising agents for clinical applications.[3] Among the different *h*CA isoforms, *h*CA IX and XII are predominantly found in tumor cells and show a rather limited diffusion in normal cells. Both isoforms are multidomain trans-membrane proteins with an extracellular CA domain and were demonstrated to participate in the rather complex machinery of pH regulation.[4] In particular, the membrane-associated *h*CA IX is considered a tumor-associated protein due to its low level of expression in normal tissues and high overexpression in almost all hypoxic tumors, where it contributes to survival, proliferation, invasion, and metastasis of cancer cells.[5] For these reasons, the *h*CA IX isoform has attracted the attention of many researchers focusing their efforts on the development of potent *h*CA IX inhibitors. As a result, a plethora of inhibitors has been reported in literature with compounds mainly belonging to the sulfonamide, dithiocarbamate, coumarin, sulfocoumarin, sulfamate, and carboxylate classes. Furthermore, an ongoing clinical trial (NCT03450018) is evaluating the sulfonamide inhibitor SLC-0111 in *h*CA IX-positive patients diagnosed with metastatic pancreatic ductal adenocarcinoma.[6]

At present, many compounds that act as low nanomolar or subnanomolar *h*CA IX inhibitors are known; however, beyond

Received: December 17, 2020
Accepted: January 14, 2021
Published: January 26, 2021



inhibitory potency, an important key feature that must be considered for a potential therapeutic application of these compounds is their selectivity against the other *h*CA isoforms and especially against *h*CA I/II, which are ubiquitously distributed and involved in key physiological processes.[7] This is particularly true for *h*CA II since it has the widest tissue distribution and is highly expressed in red blood cells.[8] Because most of the drugs are administered systemically and are membrane-permeable, *h*CA II is likely to sequester nonselective *h*CA inhibitors reducing their circulating concentrations, decreasing their bioavailability for *h*CA IX, and thus limiting their exposure within tumors.[9]

Overcoming the lack of selectivity for a specific *h*CA isoform represents the major challenge in the development of *h*CA inhibitors for therapy. The difficulty in finding a compound selective for a specific *h*CA isoform is due to the high sequence and structural homology shared by all *h*CA isoforms.[10] The large number of compounds reported in literature tested for their inhibition activity against *h*CA IX and *h*CA II has prompted us to generate a database of compounds with selectivity for *h*CA IX over *h*CA II or nonselective. Machine learning (ML) was then applied to predict isoform-selective inhibitors, and compound features determining successful predictions were identified. The resulting feature patterns were further analyzed on the basis of X-ray structures of *h*CA-inhibitor complexes, revealing individual features that were directly implicated in isoform selectivity.

2. MATERIALS AND METHODS

2.1. Compound Data Sets. Our compound collection was assembled from publicly available data extracted from the PubChem BioAssay database (accessed September 2019).[11] Compounds with measured potencies against *h*CA II and *h*CA IX (corresponding to UniProt IDs "P00918" and "Q16790", respectively) were collected. In order to ensure homogeneous experimental conditions and inter-assay data consistency,[12] only assays originating from the laboratory of C. T. Supuran were considered, which amounted to a total of 1138 assays (PubChem AIDs) and 7121 compounds (CIDs). We intended to generate a comprehensive and intrinsically heterogeneous set of inhibitors covering different variants of inhibitory mechanisms, all of which were directed against the active site of *h*CA. Since we aimed at predicting isoform selectivity of *h*CA inhibitors and rationalizing these predictions, we considered it important to comprehensively analyze different types of inhibitors, which further challenged machine learning. Training and test instances were available for all types of inhibitors and considered in combination. Accordingly, the results were generalizable (and not confined to subsets of inhibitors). Only enzyme-inhibitor interactions for which numerically defined inhibition constants (K_i values) were available were considered. No K_i threshold was applied for inhibitors. If two K_i values were available for a compound, then preference was given to the one reported in the source publication. For compounds with three or more measurements, K_i values deviating by more than 25% from the calculated mean K_i were discarded, and the mean K_i value was recalculated and assigned as the final potency annotation. Applying these criteria resulted in a total of 2506 inhibitors tested against both *h*CA isoforms for which a subsequent selectivity analysis was carried out. For each ligand, a selectivity index (SI) was calculated as the difference between the measured negative logarithmic (pK_i) values for *h*CA IX and *h*CA II. Hence, compounds with SI > 0.7, corresponding to at least a five-fold higher potency for *h*CA IX over *h*CA II, were categorized as selective *h*CA IX inhibitors. Conversely, compounds with SI ≤ 0.7 were classified as nonselective *h*CA inhibitors. This classification scheme yielded a data set of 870 *h*CA IX-selective and 1636 nonselective inhibitors.

2.2. Molecular Representation. Building ML models for distinguishing between selective and nonselective *h*CA inhibitors requires the use of molecular representations such as numerical descriptors or fingerprints. Therefore, for each compound, a modified version of the molecular graph-based (i.e., the stereochemically insensitive) extended connectivity fingerprint with bond diameter 4 (ECFP4)[13] was calculated using the Morgan fingerprint implementation of RDKit.[14] ECFPs account for specific atom environments (for ECFP4, those within a radius of two bonds around an atom), which are represented as hash values. In cheminformatic ML applications, ECFP4 has become a widely accepted standard representation for compounds with comparable or superior performance relative to other (fingerprint) descriptors.[15] The ECFP4 hash values for all unique atom environments in each data set compound were computed, resulting in a total of 6061 unique structural features. The hash value positions in the molecular feature vectors were then organized according to their frequency of occurrence in a descending manner. Hence, for each compound, the presence or absence of a specific structural feature determined whether its corresponding bit position in the 6061-dimensional molecular feature vector was set to 1 or 0. This procedure did not include any additional dimensionality reduction such as standard fingerprint "folding" into a predefined fixed-length vector and hence avoided potential bit collisions that may be caused by ambiguous feature-bit mappings. As a result, unambiguous reverse mapping of fingerprints to their corresponding structural features allowed for the visualization and assessment of the importance of individual features during ML classification.

2.3. Structural Organization. To identify analog series (ASs) formed by *h*CA inhibitors, data set compounds were subjected to bond fragmentation according to a set of retrosynthetic rules and organized into analog series (ASs) using the compound-core relationship algorithm.[16] Accordingly, compounds containing the same structural core and different substituents were combined into an AS.

2.4. Machine Learning Methods. *2.4.1. Random Forest.* A random forest (RF) is a supervised ML algorithm that consists of a large number of individual decision trees forming an ensemble classifier. Each individual tree produces a class label prediction for a given data instance, and the final prediction outcome is determined by the majority class vote.[17] Class weight balancing was automatically inferred by the model as inversely proportional to the class label frequencies in the input data. All remaining hyperparameters were set to their default values in scikit-learn version 0.23.1.[18]

2.4.2. Support Vector Machine. A support vector machine (SVM) is a supervised ML algorithm that constructs a hyperplane or set of hyperplanes in a multidimensional feature space, which are used for classification or regression. In the case of classification, acceptable separation is achieved by a hyperplane having the largest distance to the nearest training data points of any class. Thus, maximizing the margin lowers the generalization error of the classifier.[19]

Furthermore, kernel functions enable the algorithm to operate in a high-dimensional implicit feature space. Instead of explicitly computing the data coordinates in that space, the inner products of their pairwise projections are calculated. This approach is commonly referred to as the "kernel trick" and presents a computationally efficient alternative to explicit dimensionality expansion.[20] Accordingly, if linear separation via a hyperplane is not feasible in a given feature space, then the kernel trick facilitates implicit mapping of training compounds into a higher-dimensional feature space where linear separation might become feasible. Herein, the linear and Tanimoto kernels[21] were used, and the better performing kernel was selected for each model during internal cross validation (see below). The regularization parameter *C* determines the magnitude of error penalization and balances model performance in the training set and overfitting. During parameter optimization, *C* values 0.01, 0.1, 1, 10, and 100 were evaluated. SVM training was performed using scikit-learn version 0.23.1.

2.5. Cross Validation and Performance Measures. All ML calculations were carried out by applying a standard double cross validation procedure. First, the ECFP4 representations of selective and nonselective inhibitors were assigned classification labels of "1" and "0", respectively. Then, the data set was recurrently divided 10 times by random sampling into 80% training and 20% test compounds. Calculation parameters specified above were optimized via internal five-fold cross validation on the training set, and the best performing parameter settings were used for test set predictions. Based on the predictions from the 10 independent external cross validation trials, the following measures were computed in order to evaluate model performance: balanced accuracy (BA),[22] F1 score,[23] and Matthew's correlation coefficient (MCC),[24] defined as follows

$$\mathrm{BA} = \frac{0.5\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} + \frac{0.5\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}}$$

$$\mathrm{MCC} = \frac{\mathrm{TP} \times \mathrm{TN} - \mathrm{FP} \times \mathrm{FN}}{\sqrt{(\mathrm{TP} + \mathrm{FP})(\mathrm{TP} + \mathrm{FN})(\mathrm{TN} + \mathrm{FP})(\mathrm{TN} + \mathrm{FN})}}$$

$$\mathrm{F1} = \frac{2\mathrm{TP}}{2\mathrm{TP} + \mathrm{FP} + \mathrm{FN}}$$

TP, TN, FP, and FN abbreviate true positives, true negatives, false positives, and false negatives, respectively. Here, TP + FN corresponds to the total number of selective compounds; conversely, FP + TN corresponds to the total number of nonselective compounds.

As indicated by its formula, MCC takes into account all values of the confusion matrix derived from binary classification. It has the range [−1,1] where MCC = 1 represents a perfect classification (with no FP and FN), MCC = 0 is equivalent to random classification, and MCC = −1 indicates complete disagreement between predicted and actual class labels. BA accounts for the fraction of correct predictions while taking data imbalance into account through equivalent weighting. This was another appropriate measure for our analysis because the compound data set contained approximately twice as many nonselective (negative class) as selective (positive class) compounds. BA has the range [0,1] with BA = 1 describing perfect, BA = 0.5 random, and BA = 0 completely inaccurate classifications, respectively. The F1 score is a composite measure representing the harmonic mean of precision and recall. It strongly emphasizes TP values without taking TN values into consideration. High F1 values indicate good model performance.

In addition, receiver operating characteristic (ROC) curves[25] were computed to compare the TP rate ([0, 1], *y*-axis) to the FP rate ([0, 1], *x*-axis) at different classification thresholds, and the area under the ROC curve (AUROC) was determined.[25] In a ROC curve, the diagonal line is equivalent to a random class prediction and yields an AUROC value of 0.5. Increasing AUROC values between 0.5 and 1 are indicative of increasing model performance, with the value of 1.0 representing a perfect prediction.

Furthermore, to ensure statistically sound comparisons of individual inhibitors, we also required that each inhibitor was predicted as a test set compound in at least five different external cross validation trials. This criterion was met after 26 trials, and for each selective and nonselective inhibitor, "model prediction consistency" (MPC) was calculated as follows

$$\mathrm{MPC}_{\mathrm{selective}}[\%] = \frac{\mathrm{TP}}{\text{Number of predictions}} \times 100$$

$$\mathrm{MPC}_{\mathrm{nonselective}}[\%] = \mathrm{T}\frac{\mathrm{N}}{\text{Number of predictions}} \times 100$$

Accordingly, an MPC value of 100% indicated that a compound was consistently correctly classified. Conversely, an MPC value of 0% resulted from consistently incorrect classification in each trial.

2.6. Feature Weighting and Frequency Analysis. To identify individual structural features determining the classification, corresponding feature weights (FWs) were extracted from SVM models. For FW extraction, two previously introduced methods using the Tanimoto kernel were applicable.[26,27] In addition to their numerical values, FWs were assigned positive or negative signs depending on their relative importance for predicting a specific class label. According to this definition, the SVM model assigned a positive sign to a feature if its presence predominantly determined the prediction of selective inhibitors. In contrast, a feature with a negative sign predominantly determined the identification of nonselective inhibitors.

The corresponding frequency distributions for selective and nonselective compounds, respectively, were defined as

$$f_{\mathrm{selective},i} = \frac{\#\text{Selective compounds with feature } i}{\#\text{Selective compounds}}$$

$$f_{\mathrm{nonselective},i} = \frac{\#\text{Nonselective compounds with feature } i}{\#\text{Nonselective compounds}}$$

In order to determine whether the presence of a given structural feature was more important for the identification of selective or nonselective inhibitors, a frequency difference value ΔF was calculated for each feature *i*

$$\Delta F_i = f_{\mathrm{selective},i} - f_{\mathrm{nonselective},i}$$

Thus, features with positive ΔF values were preferentially found in selective *h*CA IX inhibitors, whereas negative ΔF values indicated features that preferentially occurred in nonselective inhibitors. Furthermore, features that were exclusively found in selective or nonselective compounds were identified and prioritized.

2.7. Analysis of X-ray Structures. Key features determining predictions were further analyzed on the basis of publicly available X-ray structures of *h*CA II and *h*CA IX in complex with inhibitors. A total of 488 *h*CA II, 11 *h*CA IX, and 59 *h*CA IX-mimicking (mutated) proteins in complex with unique inhibitors, many of which were contained in our data set (Table 1), were obtained from the RCSB Protein Data Bank (accessed September 2020).[28] An *h*CA IX-mimicking protein contains the original *h*CA II isoform active site engineered by site-directed mutagenesis to represent the wild-type *h*CA IX isoform by introducing relevant residue replacements. These replacements included A65S, N67Q, E69T, I91L, F131V, K170E, and L204A (*h*CA II sequence numbering).[29] Further analysis revealed that compounds in four of the *h*CA IX and 30 of the *h*CA IX-mimicking structures were also cocrystallized with the *h*CA II isoform. Thus, these compounds provided a meaningful basis for studying different binding modes and specific interactions associated with *h*CA IX/*h*CA II selectivity taking into account structural features that determine ML predictions. Superpositions of X-ray structures were obtained using UCSF Chimera.[30]

Table 1. X-ray Structures[a]

target	PDB entries	unique inhibitors	contained in the ML data set	shared by isoforms	shared in the ML data set
*h*CA II	811	488	93	34	12
*h*CA IX	20	11	4	4	2
*h*CA IX-mimic	94	59	15	30	10

[a] Reported are X-ray structures of *h*CA-inhibitor complexes evaluated in our analysis. For example, from the PDB, 811 structures of *h*CA II-inhibitor complexes were retrieved, which contained 488 unique inhibitors, 93 of which were contained in our data set for ML. Thirty-four of these inhibitors were found in complex structures of all three *h*CA isoforms, and 12 of these shared inhibitors were contained in our data set.

3. RESULTS AND DISCUSSION

3.1. Inhibitors and Analog Series. Initially, the data set of selective and nonselective inhibitors for ML was structurally organized. It was found to contain 328 ASs with two or more compounds, representing ~70% of the 1748 inhibitors. ASs comprised only *h*CA IX-selective inhibitors (48 series), only nonselective (163), or both selective and nonselective inhibitors (117 "mixed" series). Figure 1 shows that these different AS categories displayed similar size distributions, with a clear dominance of small series with less than five compounds. Only small numbers of larger ASs comprising up to 25 compounds were detected. Hence, there was no

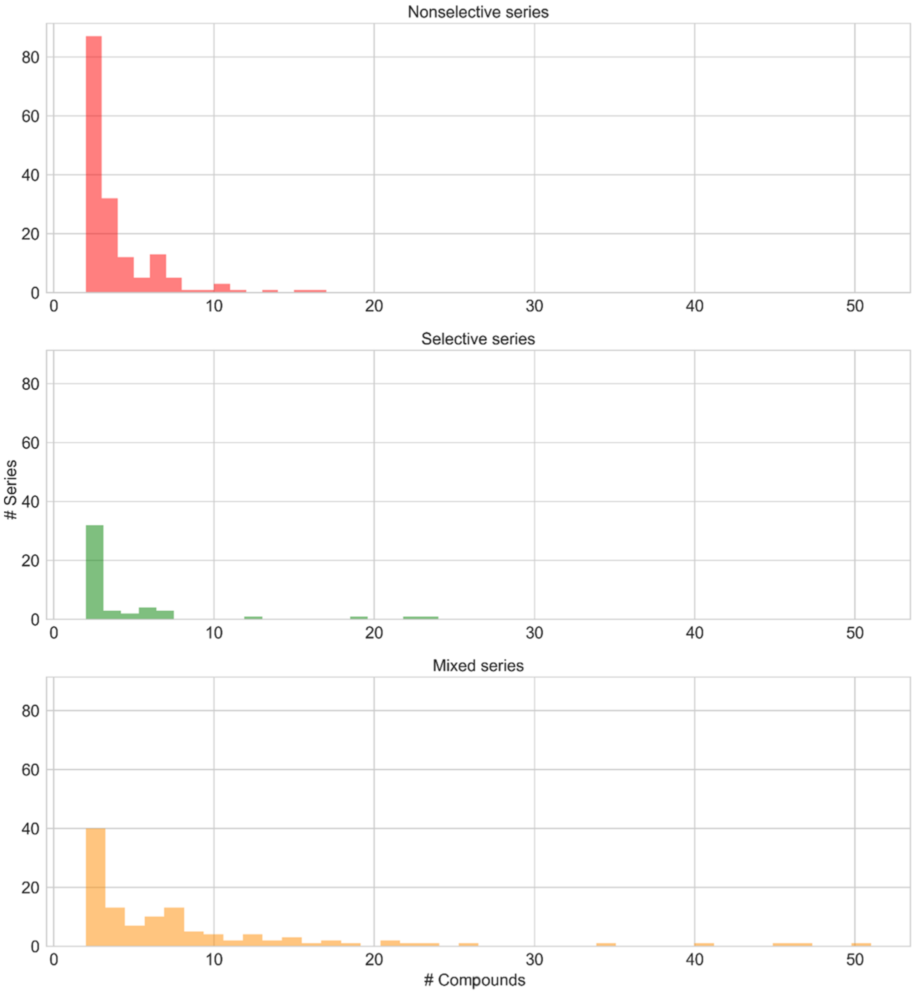


Figure 1. Distribution of analog series. Histograms report the size distributions of ASs exclusively consisting of nonselective (red) or selective (green) inhibitors or combining both nonselective and selective compounds (mixed series, orange). The three histograms are shown on the same scale.

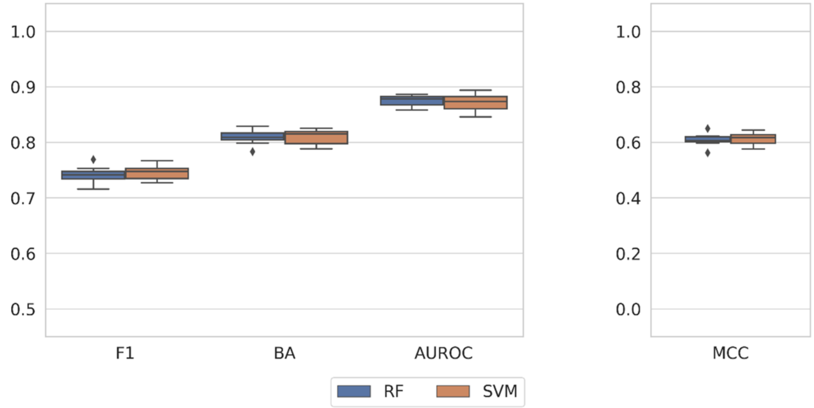


Figure 2. Prediction accuracies. Boxplots report prediction accuracy over 10 independent RF (blue) and SVM (orange) trials using different training and test sets. From the left to the right, results are shown for the F1, BA, AUROC, and MCC measures. Boxplots show the smallest value (lower whisker), lower quartile (lower boundary of the box), median (vertical line in the box), upper quartile (upper boundary of the box), and the maximum value (upper whisker). Values classified as statistical outliers are represented as diamonds.

global or category-centered bias in AS composition toward small numbers of large series, which might limit predictive modeling or conclusions drawn from such investigations. However, as revealed by the presence of 117 mixed ASs, many selective and nonselective inhibitors displayed close structural relationships, which principally challenged the prediction of selective inhibitors. Furthermore, there were essentially twice as many nonselective than selective inhibitors available (applying a moderate SI > 0.7 criterion), which reflected the inherent difficulties in obtaining isoform-selective *h*CA inhibitors, as described above. Rather than balancing the number of compounds with different class labels (positive/selective or negative/nonselective) for training, which generally favors ML predictions, we preferred retaining this intrinsic imbalance, thus attempting predictions under realistic data conditions. Taken together, in light of the statistical and structural characteristics of the inhibitor data set, the selectivity prediction task was considered challenging.

3.2. Prediction of Selective Inhibitors.

We then attempted to systematically predict *h*CA IX-selective inhibitors in cross validation trials. Contrary to our expectations, generally high prediction accuracy was achieved, for both RF and SVM models and on the basis of all performance measures, as summarized in Figure 2. The performance of RF and SVM classification was very similar with only little variation over different trials. With median F1 values of ~0.75, median BA of >0.8, and AUROC values close to 0.9, the predictions consistently yielded reasonable to high accuracy, as further indicated by median MCC values of ~0.6. We also assessed the predictions at the level of ASs, which mirrored the structural organization of test data. As reported in Table 2, 31 of 48 ASs exclusively comprising selective inhibitors were consistently correctly predicted (MPC = 100%), corresponding to 173 of 219 compounds contained in selective ASs. Moreover, 145 of 163 ASs exclusively consisting of nonselective inhibitors were consistently correctly predicted, including 508 of 549 nonselective inhibitors. Overall, 83% of ASs consisting of either only selective or nonselective inhibitors were always correctly predicted. Hence, assessing the predictions at the level of ASs further confirmed their global accuracy.

Table 2. Prediction of Analog Series[a]

		compounds	analog series
selective	total data set	219	48
	$\mathrm{MPC}_{\mathrm{selective}}$ = 100%	173	31
nonselective	total data set	559	163
	$\mathrm{MPC}_{\mathrm{nonselective}}$ = 100%	508	145

[a] Reported are ASs exclusively consisting of selective and nonselective inhibitors and their subsets that were consistently correctly predicted (MPC = 100%).

3.3. Feature Relevance Analysis.

In light of the observed accuracy, we further assessed the predictions by exploring structural features that were responsible for the predictions. In ML, diagnostic approaches are still rare but essential for rationalizing successful predictions or failures. Given the equivalence of the results obtained for RF and SVM classification and the consistently better predictive performance of the Tanimoto over the linear kernel, we focused the analysis on SVM calculations, for which feature weighting approaches were applicable (see Materials and Methods). Accordingly, we determined ECFP4 features with positive and negative SVM weights contributing to the correct prediction of selective and nonselective inhibitors, respectively, and searched for contributing features that exclusively occurred in selective or nonselective compounds. Figure 3 shows that large numbers of features were identified that contributed with varying weights to positive or negative predictions and exclusively occurred in selective and nonselective inhibitors, respectively. As indicated by generally low ΔF values, exclusive features typically only occurred in small subsets of compounds. Hence, there were no distinguishing features that could be generalized, consistent with the structural heterogeneity of selective and nonselective compounds, as revealed by their partitioning into many different ASs of mostly small size. Furthermore, most of the exclusive features had absolute weights <0.10, and comparably few features with absolute weights >0.15 were detected. While many features contributed to meaningful SVM predictions, the latter features largely determined correct predictions of selective or nonselective inhibitors. Figure 4 shows the top 10 features with the largest weights that exclusively occurred in selective inhibitors and thus made the most important contribution to the prediction of selectivity. These ECFP4 features defined different structural fragments that occurred in test compounds as substructures. Notably, these features included two distinct sulfonamide-containing

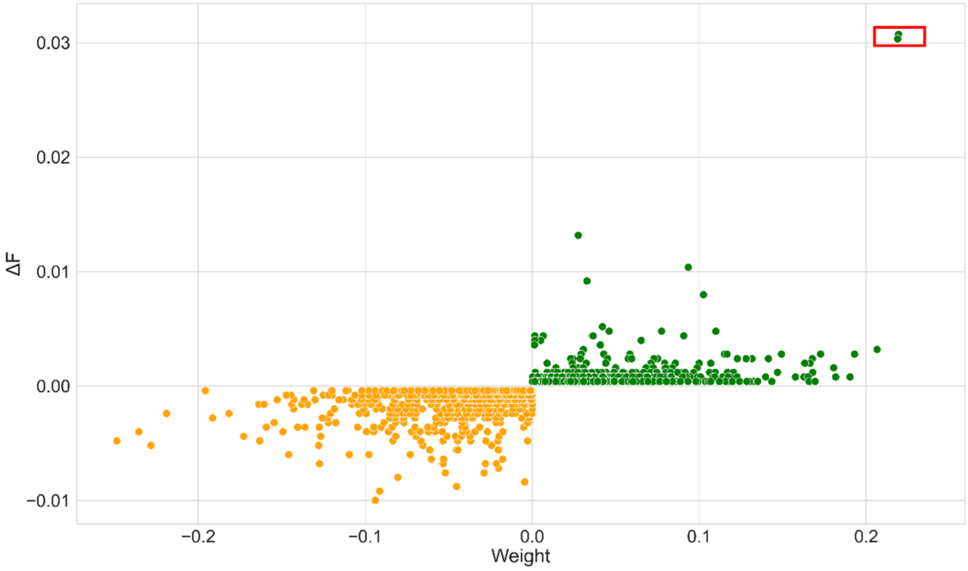


Figure 3. Distribution of exclusive features. The scatterplot shows the distribution of ECFP4 features that are exclusively found in nonselective (orange) or selective (green) inhibitors. Each dot represents a unique feature. The relative frequency of occurrence of a feature in nonselective or selective compounds (ΔFrequency; nonselective < 0, selective > 0) is plotted against the mean feature weight from SVM classification. Negative and positive weights represent contributions to the prediction of nonselective and selective compounds, respectively. Two features with the largest Δfrequency values and the highest weights are highlighted (red box, upper right corner).

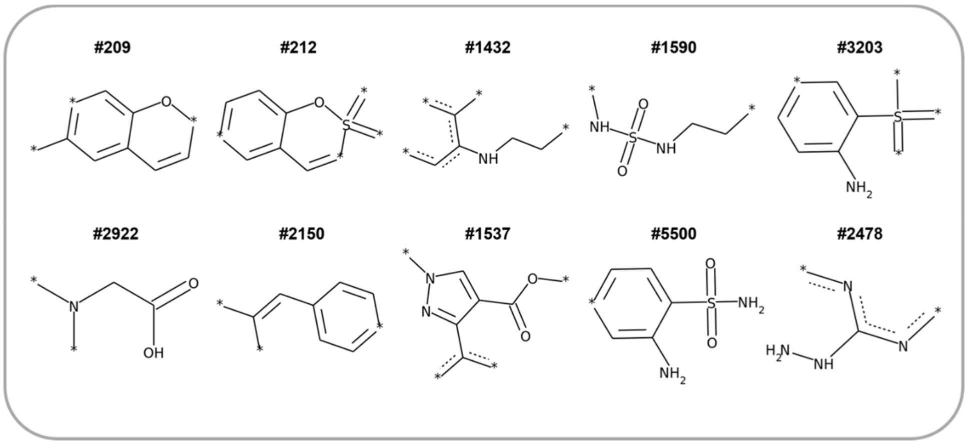


Figure 4. Exclusive features. Shown are the top 10 features with the largest SVM weights that exclusively occurred in selective inhibitors (ordered from upper left, top 1, to lower right, top 10). Features 209 and 212 are highlighted in Figure 3.

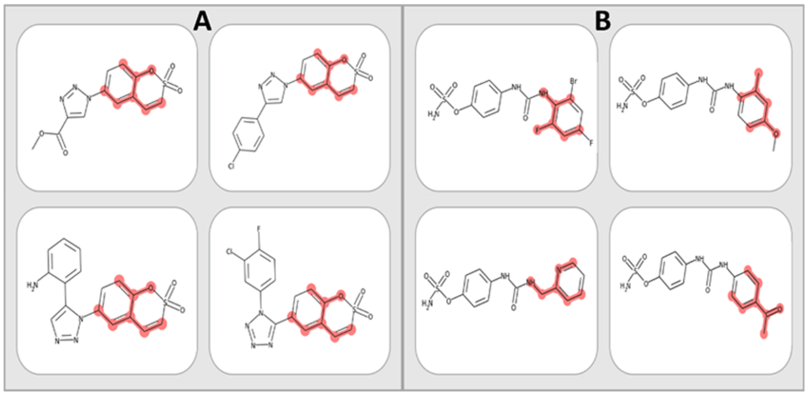


Figure 5. (A,B) Feature mapping. In (A), feature 209 from Figure 4 is mapped (red) on exemplary analogs from a selective AS with $\mathrm{MPC}_{\mathrm{selective}}$ = 100%. In (B), members of another selective AS with $\mathrm{MPC}_{\mathrm{selective}}$ = 100% are shown. Features with the highest SVM weights are mapped on the analogs. For a compound from this AS, X-ray structures of complexes with *h*CA II and *h*CA IX were available.

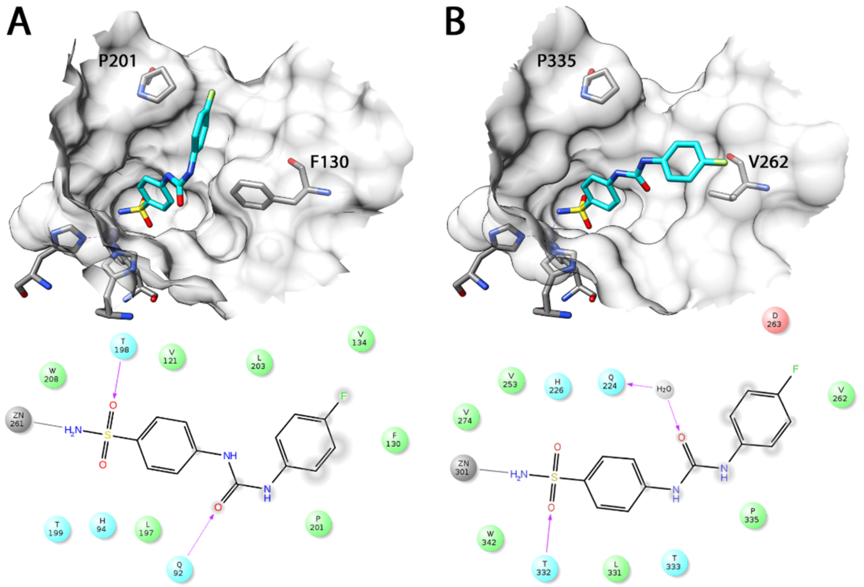


Figure 6. Structure-based analysis. X-ray structures of SLC-0111 in complex with (A) *h*CA II and (B) *h*CA IX-mimic forms. (upper section) The catalytic zinc cation interacting with the sulfonamide moiety of the inhibitor is depicted as a sphere; (lower section) an interaction map is shown for the inhibitor and amino acid residues lining the active site (green, lipophilic residues; sky blue, polar residues; red, charged residues).

substructures (numbers 1590 and 5500). As discussed above, the sulfonamide group complexing the catalytic zinc cation in the active site is a hallmark of many potent *h*CA inhibitors, which is contained in both selective and nonselective inhibitors. Thus, the presence or absence of a sulfonamide group alone is insufficient to distinguish between selective and nonselective inhibitors. Rather, the way in which sulfonamide is embedded in substructures/compounds or specific feature combinations in which it occurs might contribute to the prediction of selective inhibitors. Furthermore, the two top ranked features in Figure 4 were features 209 and 212, which delineated overlapping substructures and had the largest weights and by far the highest Δ*F* value among positive features, as shown in Figure 3 (where features 209 and 212 are highlighted). Thus, these two features accounting for similar structural fragments made overall the most important contributions to the predictions of selective inhibitors.

3.4. Feature Mapping.

The keyed design of the feature fingerprint with 1:1 bit-to-feature correspondence made it possible to map key features to structures of test compounds. Therefore, we searched for ASs exclusively comprising selective inhibitors that were consistently correctly predicted and contained features 209 and/or 212. Several ASs were identified. Figure 5A shows an exemplary series of sulfocoumarin derivatives in which both features were present and formed a substructure covering most of the sulfocoumarin core. These compounds are potent and selective inhibitors of the tumor-associated *h*CA IX and *h*CA XII isoforms.[31] Of note, coumarin and sulfocoumarin derivatives can act by complex mechanisms. These compounds are known to undergo hydrolysis upon binding to the catalytic site of *h*CAs. However, prior to hydrolysis, they bind within the *h*CA active site similarly to phenols, i.e., by anchoring to the zinc-bound water molecule/hydroxide ion,[32] as confirmed by an X-ray structure of 2-thioxocoumarine in complex with *h*CA II.[33] This recognition mechanism was intentionally included in our ML analysis, yielding promising results.

In Figure 5B, another selective AS is shown in which features making the largest contributions to consistently correct predictions were mapped on individual analogs containing them. All of these features delineated an extended terminal pyridyl or substituted phenyl ring systems distant from the sulfonamide moiety. Thus, in both cases, key features for correct predictions defined coherent substructures of corresponding regions of analogs, which provided a basis of interpreting predictions.

3.5. Relating Important Features to Selectivity.

Feature weighting and mapping identified a number of features that determined accurate SVM predictions of *h*CA IX-selective inhibitors. However, although these features made major contributions to ML predictions, it could not be concluded that they were implicated in or responsible for selectivity. Structural features determining predictions may or may not be of biological relevance, the assessment of which goes beyond ML analysis. Hence, the question whether substructures defined by the most important features we identified were indeed implicated in inhibitor selectivity required additional analysis.

3.6. Structure-Based Analysis.

To address this question, we searched for selective inhibitors for which X-ray structures of complexes with *h*CA II and *h*CA IX or *h*CA IX-mimics were available. Such structures provided a basis for viewing mapped key features in light of enzyme-inhibitor interactions and exploring potential differences implicated in selectivity. Among the large number of publicly available *h*CA isoform X-ray structures (Table 1), a limited number of suitable *h*CA II/IX structures with selective inhibitors we predicted were identified and compared. A particularly instructive example was obtained by comparing X-ray structures of *h*CA II and *h*CA IX-mimicking protein in complex with the *h*CA IX-selective inhibitor SLC-0111 that belongs to the series in Figure 5B (PDB entries 3N4B and 5JN3, respectively). The binding mode of SLC-0111 in the *h*CA II and *h*CA IX-mimic structures is shown in Figure 6A and B, respectively. As observed for all members of the corresponding AS, the feature with the highest

```
                        134     4         5         6         7            8      9
                          456789012345678901234567890123456789012345678901234567890123
sp|P00918|CAH2_HUMAN      MSHHWGYGKHNGPEHWHKDFPIAKGERQSPVDIDTHTAKYDPSLKPLSVS---YDQATSL
sp|Q16790|CAH9_HUMAN      EGDDQSHWRYGGDPPWPRVSPACAGRFQSPVDIRPQLAAFCPALRPLELLGFQLPPLPEL
                           ... .: ::.*   * :  * . *. ****** .: * : *:*:**.:        ..*

                                0         1             2         3         4
                          4567890123456789012345677----89012345678901234567890123456789
sp|P00918|CAH2_HUMAN      RILNNGHAFNVEFDDSQDKAVLKGGPLDGTYRLIQFHFHWGSLDGQGSEHTVDKKKYAAE
sp|Q16790|CAH9_HUMAN      RLRNNGHSVQLTLPPGLEMALGPG----REYRALQLHLHWGAAGRPGSEHTVEGHRFPAE
                          *: ****:.:: :  . : *:  *      ** :*:*:***: .  ******: :::.**

                          5         6         7         8         9         0
                          012345678901234567890123456789012345678901234567890123456789
sp|P00918|CAH2_HUMAN      LHLVHWNTKYGDFGKAVQQPDGLAVLGIFLKVG-SAKPGLQKVVDVLDSIKTKGKSADFT
sp|Q16790|CAH9_HUMAN      IHVVHLSTAFARVDEALGRPGGLAVLAAFLEEGPEENSAYEQLLSRLEEIAEEGSETQVP
                          :*:** .* :. ..:*: :*.*****. **: * . :.. ::::. *:.*  :*..::..

                          1         2         3         4         5         6
                          01234567890123456789012345678901234567890123456789012---3456
sp|P00918|CAH2_HUMAN      NFDPRGLLPESLD-YWTYPGSLTTPPLLECVTWIVLKEPISVSSEQVLKFRKLNFNGEGE
sp|Q16790|CAH9_HUMAN      GLDISALLPSDFSRYFQYEGSLTTPPCAQGVIWTVFNQTVMLSAKQLHTLSDT---LWGP
                          .:*  .***..:. *: * *******  : * * *:::.: :*::*: .: .      *

                             7         8         9         0         1         2
                          789012345678901234567890123456789012345678901234567890123456
sp|P00918|CAH2_HUMAN      PEELMVDNWRPAQPLKNRQIKASFK-----------------------------------
sp|Q16790|CAH9_HUMAN      GDSRLQLNFRATQPLNGRVIEASFPAGVDSSPRAAEPVQLNSCLAAGDILALVFGLLFAV
                           :. :  *:*.:***:.* *:***
```

Figure 7. Sequence alignment of *h*CA II and *h*CA IX. Shown is the alignment of the *h*CA II (CAH2_HUMAN) and *h*CA IX (CAH9_HUMAN) amino acid sequences taken from UniProt.[35] Binding site residues are highlighted in yellow, and nonconserved residues participating in the formation of the binding site are shown in bold. Identical residues are indicated with "*", while conservative residue replacements are marked with ":" and ".".

positive SVM weight mapped to the terminal ring (in this case, a 4-fluorophenyl moiety) distant from the sulfonamide group complexing the catalytic zinc ion. In both complexes, the benzenesulfonamide fragment position was superimposable and interacted with the catalytic zinc ion; in the *h*CA II structure, the *N*,*N*′-ureic portion of the ligand adopted a less stable cis/trans conformation with the 4-fluorophenyl moiety that interacted with residue P201. This orientation was determined by the steric hindrance between the 4-fluorophenyl moiety and the phenyl ring of the F130 side chain that determined the observed orientation of the compound. The phenylalanine residue was not conserved in *h*CA IX where it was replaced by a smaller valine residue (V262). Figure 7 shows the corresponding sequence alignment. This substitution led to the absence of steric hindrance between the protein and the 4-fluorophenyl moiety of the ligand. As a consequence, the inhibitor was able to maintain a more stable trans/trans *N*,*N*′-ureic conformation with a strong lipophilic interaction between 4-fluorophenyl and V262. By contrast, suboptimal interactions in this region of *h*CA II resulted in a loss of potency of the inhibitor compared to *h*CA IX and hence in selectivity of the compound for *h*CA IX over *h*CA II. The importance of inhibitor interactions with residue 131 in *h*CA isoforms has also been pointed out in the literature,[34] providing corroborating evidence. These considerations were equally applicable to most of the other analogs comprising the *h*CA IX-selective series in Figure 5B, which was consistently correctly predicted. In all instances, features with the highest positive weights determining the predictions were mapped to the corresponding ring structures, which were implicated in selectivity-determining interactions with *h*CA isoforms. Therefore, in this case, features that determined ML predictions were directly implicated in critical enzyme-inhibitor interactions determining compound selectivity and thus biologically relevant.

3.7. Conclusions.

Predicting target-selective compounds typically represents a challenging task. In this work, we have attempted to predict inhibitors with selectivity for the tumor-associated *h*CA IX isoform over the ubiquitous *h*CA II isoform via ML. Surprisingly accurate and robust predictions were obtained using RF and SVM models, including many selective or nonselective ASs that were consistently correctly predicted, lending credence to the computational approach. These rather encouraging findings prompted us to further analyze the predictions. SVM feature weight analysis revealed numerous features that exclusively occurred in selective or nonselective compounds and contributed to positive and negative predictions. Highly weighted features were found to map to corresponding regions in ASs, hence rationalizing origins of successful predictions. For selectivity analysis and compound design, signature features of compound selectivity are of prime interest. However, there is no guarantee that features that make large contributions to or determine positive ML predictions are indeed biologically relevant. Therefore, we have gone a step further and evaluated important features on the basis of X-ray structures of complexes formed by the *h*CA IX and *h*CA II isoforms and selective inhibitors. For an exemplary selective AS, comparisons of corresponding X-ray structures revealed that features determining correct predictions defined substructures of inhibitors that were involved in selectivity-conferring interactions, thus establishing proof-of-principle. Demonstrating biological relevance of distinguishing features identified by ML is far from being routine, and to our knowledge, this may be one of the first studies doing so. Our findings also indicate that the ML models reported herein should have potential for practical applications in the search for

https://dx.doi.org/10.1021/acsomega.0c06153

new *h*CA IX-selective inhibitors. Therefore, as a part of our study, trained RF and SVM models are made available upon request.

■ AUTHOR INFORMATION

Corresponding Authors

Tiziano Tuccinardi − *Department of Pharmacy, University of Pisa, 56126 Pisa, Italy;* orcid.org/0000-0002-6205-4069; Phone: 39-050-2219595; Email: tiziano.tuccinardi@unipi.it

Jürgen Bajorath − *Department of Life Science Informatics, B-IT, LIMES Program Unit Chemical Biology and Medicinal Chemistry, Rheinische Friedrich-Wilhelms-Universität, D-53115 Bonn, Germany;* orcid.org/0000-0002-0557-5714; Phone: 49-228-7369-100; Email: bajorath@bit.uni-bonn.de

Authors

Salvatore Galati − *Department of Life Science Informatics, B-IT, LIMES Program Unit Chemical Biology and Medicinal Chemistry, Rheinische Friedrich-Wilhelms-Universität, D-53115 Bonn, Germany; Department of Pharmacy, University of Pisa, 56126 Pisa, Italy;* orcid.org/0000-0002-1959-5839

Dimitar Yonchev − *Department of Life Science Informatics, B-IT, LIMES Program Unit Chemical Biology and Medicinal Chemistry, Rheinische Friedrich-Wilhelms-Universität, D-53115 Bonn, Germany*

Raquel Rodríguez-Pérez − *Department of Life Science Informatics, B-IT, LIMES Program Unit Chemical Biology and Medicinal Chemistry, Rheinische Friedrich-Wilhelms-Universität, D-53115 Bonn, Germany*

Martin Vogt − *Department of Life Science Informatics, B-IT, LIMES Program Unit Chemical Biology and Medicinal Chemistry, Rheinische Friedrich-Wilhelms-Universität, D-53115 Bonn, Germany;* orcid.org/0000-0002-3931-9516

Complete contact information is available at: https://pubs.acs.org/10.1021/acsomega.0c06153

Author Contributions

The study was carried out, and the manuscript was written with contributions of all authors. All authors have approved the final version of the manuscript.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

S.G. is supported by Rotary Club Livorno “Mascagni”/The Rotary Foundation (global grant no. GG2095429).

■ REFERENCES

(1) Angeli, A.; Carta, F.; Supuran, C. T. Carbonic Anhydrases: Versatile and Useful Biocatalysts in Chemistry and Biochemistry. *Catalysts* **2020**, *10*, 1008.

(2) Mishra, C. B.; Tiwari, M.; Supuran, C. T. Progress in the Development of Human Carbonic Anhydrase Inhibitors and Their Pharmacological Applications: Where Are We Today? *Med. Res. Rev.* **2020**, *40*, 2485−2565.

(3) Supuran, C. T. Carbonic Anhydrases: Novel Therapeutic Applications for Inhibitors and Activators. *Nat. Rev. Drug Discovery* **2008**, *7*, 168−181.

(4) Angeli, A.; Carta, F.; Nocentini, A.; Winum, J.-Y.; Zalubovskis, R.; Akdemir, A.; Onnis, V.; Eldehna, W. M.; Capasso, C.; De Simone, G.; et al. Carbonic Anhydrase Inhibitors Targeting Metabolism and Tumor Microenvironment. *Metabolites* **2020**, *10*, 412.

(5) Supuran, C. T.; Alterio, V.; Di Fiore, A.; D'Ambrosio, K.; Carta, F.; Monti, S. M.; De Simone, G. Inhibition of Carbonic Anhydrase IX Targets Primary Tumors, Metastases, and Cancer Stem Cells: Three for the Price of One. *Med. Res. Rev.* **2018**, *38*, 1799−1836.

(6) Strapcova, S.; Takacova, M.; Csaderova, L.; Martinelli, P.; Lukacikova, L.; Gal, V.; Kopacek, J.; Svastova, E. Clinical and Pre-Clinical Evidence of Carbonic Anhydrase IX in Pancreatic Cancer and Its High Expression in Pre-Cancerous Lesions. *Cancers* **2020**, *12*, 2005.

(7) Nocentini, A.; Supuran, C. T. Carbonic Anhydrase Inhibitors as Antitumor/Antimetastatic Agents: A Patent Review (2008-2018). *Expert Opin. Ther. Pat.* **2018**, *28*, 729−740.

(8) Lindskog, S. Purification and Properties of Bovine Erythrocyte Carbonic Anhydrase. *Biochim. Biophys. Acta* **1960**, *39*, 218−226.

(9) Singh, S.; Lomelino, C. L.; Mboge, M. Y.; Frost, S. C.; McKenna, R. Cancer Drug Development of Carbonic Anhydrase Inhibitors beyond the Active Site. *Molecules* **2018**, *23*, 1045.

(10) De Simone, G.; Alterio, V.; Supuran, C. T. Exploiting the Hydrophobic and Hydrophilic Binding Sites for Designing Carbonic Anhydrase Inhibitors. *Expert Opin. Drug Discovery* **2013**, *8*, 793−810.

(11) Kim, S.; Chen, J.; Cheng, T.; Gindulyte, A.; He, J.; He, S.; Li, Q.; Shoemaker, B. A.; Thiessen, P. A.; Yu, B.; Zaslavsky, L.; Zhang, J.; Bolton, E. E. PubChem 2019 Update: Improved Access to Chemical Data. *Nucleic Acids Res.* **2019**, D1102−D1109.

(12) Poli, G.; Galati, S.; Martinelli, A.; Supuran, C. T.; Tuccinardi, T. Development of a Cheminformatics Platform for Selectivity Analyses of Carbonic Anhydrase Inhibitors. *J. Enzyme Inhib. Med. Chem.* **2020**, *35*, 365−371.

(13) Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742−754.

(14) *RDKit: Cheminformatics and Machine Learning Software*; 2013; http://www.rdkit.org (accessed November 2020).

(15) Riniker, S.; Landrum, G. A. Open-Source Platform to Benchmark Fingerprints for Ligand-Based Virtual Screening. *J. Cheminf.* **2013**, *5*, 26.

(16) Naveja, J. J.; Vogt, M.; Stumpfe, D.; Medina-Franco, J. L.; Bajorath, J. Systematic Extraction of Analogue Series from Large Compound Collections Using a New Computational Compound−Core Relationship Method. *ACS Omega* **2019**, *4*, 1027−1032.

(17) Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5−32.

(18) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; Duchesnay, E. Scikit-Learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825−2830.

(19) Vapnik, V. N. *The Nature of Statistical Learning Theory*; Springer Science & Business Media: New York, 2013.

(20) Boser, B. E.; Guyon, I. M.; Vapnik, V. N. A Training Algorithm for Optimal Margin Classifiers. In *COLT '92: Proceedings of the Fifth Annual Workshop on Computational Learning Theory*; Association for Computing Machinery: New York, NY, USA, 1992; pp. 144−152.

(21) Ralaivola, L.; Swamidass, S. J.; Saigo, H.; Baldi, P. Graph Kernels for Chemical Informatics. *Neural Networks* **2005**, *18*, 1093−1110.

(22) Brodersen, K. H.; Ong, C. S.; Stephan, K. E.; Buhmann, J. M. The Balanced Accuracy and Its Posterior Distribution. In *2010 20th International Conference on Pattern Recognition*; IEEE: 2010; pp. 3121−3124.

(23) Van Rijsbergen, C. J. *Information Retrieval*; 2 nd ed.; Butterworth-Heinemann: Oxford, UK, 1979.

(24) Matthews, B. W. Comparison of the Predicted and Observed Secondary Structure of T4 Phage Lysozyme. *Biochim. Biophys. Acta, Protein Struct.* **1975**, *405*, 442−451.

(25) Bradley, A. P. The Use of the Area under the ROC Curve in the Evaluation of Machine Learning Algorithms. *Pattern Recognit.* **1997**, *30*, 1145−1159.

(26) Balfer, J.; Bajorath, J. Visualization and Interpretation of Support Vector Machine Activity Predictions. *J. Chem. Inf. Model.* **2015**, *55*, 1136−1147.

(27) Rodríguez-Pérez, R.; Vogt, M.; Bajorath, J. Support Vector Machine Classification and Regression Prioritize Different Structural Features for Binary Compound Activity and Potency Value Prediction. *ACS Omega* **2017**, *2*, 6371−6379.

(28) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235−242.

(29) Pinard, M. A.; Boone, C. D.; Rife, B. D.; Supuran, C. T.; McKenna, R. Structural Study of Interaction between Brinzolamide and Dorzolamide Inhibition of Human Carbonic Anhydrases. *Bioorg. Med. Chem.* **2013**, *21*, 7210−7215.

(30) Pettersen, E. F.; Goddard, T. D.; Huang, C. C.; Couch, G. S.; Greenblatt, D. M.; Meng, E. C.; Ferrin, T. E. UCSF Chimera-a Visualization System for Exploratory Research and Analysis. *J. Comput. Chem.* **2004**, *25*, 1605−1612.

(31) Grandane, A.; Tanc, M.; Zalubovskis, R.; Supuran, C. T. 6-Triazolyl-Substituted Sulfocoumarins Are Potent, Selective Inhibitors of the Tumor-Associated Carbonic Anhydrases IX and XII. *Bioorg. Med. Chem. Lett.* **2014**, *24*, 1256−1260.

(32) Maresca, A.; Temperini, C.; Pochet, L.; Masereel, B.; Scozzafava, A.; Supuran, C. T. Deciphering the Mechanism of Carbonic Anhydrase Inhibition with Coumarins and Thiocoumarins. *J. Med. Chem.* **2010**, *53*, 335−344.

(33) Ferraroni, M.; Carta, F.; Scozzafava, A.; Supuran, C. T. Thioxocoumarins Show an Alternative Carbonic Anhydrase Inhibition Mechanism Compared to Coumarins. *J. Med. Chem.* **2016**, *59*, 462−473.

(34) Lomelino, C. L.; Mahon, B. P.; McKenna, R.; Carta, F.; Supuran, C. T. Kinetic and X-Ray Crystallographic Investigations on Carbonic Anhydrase Isoforms I, II, IX and XII of a Thioureido Analog of SLC-0111. *Bioorg. Med. Chem.* **2016**, *24*, 976−981.

(35) UniProt Consortium. Reorganizing the Protein Space at the Universal Protein Resource (UniProt). *Nucleic Acids Res.* **2012**, *40*, D71−D75.