



Identifying potential selective fluorescent probes for cancer-associated protein carbonic anhydrase IX using a computational approach



Rhiannon L. Kamstra^{a,b}, Wely B. Floriano^{a,b,*}

^a Lakehead University, Department of Chemistry, Thunder Bay, ON P7B 5E1, Canada

^b Thunder Bay Regional Research Institute, Thunder Bay, ON P7A 7T1, Canada

ARTICLE INFO

Article history:

Accepted 18 October 2014

Available online 24 October 2014

Keywords:

Carbonic anhydrase IX
Fluorescent molecular probes
Virtual ligand screening
HierVLS
Tumor hypoxia
ATTO680

ABSTRACT

Carbonic anhydrase IX (CAIX) is a biomarker for tumor hypoxia. Fluorescent inhibitors of CAIX have been used to study hypoxic tumor cell lines. However, these inhibitor-based fluorescent probes may have a therapeutic effect that is not appropriate for monitoring treatment efficacy. In the search for novel fluorescent probes that are not based on known inhibitors, a database of 20,860 fluorescent compounds was virtually screened against CAIX using hierarchical virtual ligand screening (HierVLS). The screening database contained 14,862 compounds tagged with the ATTO680 fluorophore plus an additional 5998 intrinsically fluorescent compounds. Overall ranking of compounds to identify hit molecular probe candidates utilized a principal component analysis (PCA) approach. Four potential binding sites, including the catalytic site, were identified within the structure of the protein and targeted for virtual screening. Available sequence information for 23 carbonic anhydrase isoforms was used to prioritize the four sites based on the estimated “uniqueness” of each site in CAIX relative to the other isoforms. A database of 32 known inhibitors and 478 decoy compounds was used to validate the methodology. A receiver–operating characteristic (ROC) analysis using the first principal component (PC1) as predictive score for the validation database yielded an area under the curve (AUC) of 0.92. AUC is interpreted as the probability that a binder will have a better score than a non-binder. The use of first component analysis of binding energies for multiple sites is a novel approach for hit selection. The very high prediction power for this approach increases confidence in the outcome from the fluorescent library screening. Ten of the top scoring candidates for isoform-selective putative binding sites are suggested for future testing as fluorescent molecular probe candidates.

© 2014 Elsevier Inc. All rights reserved.

1. Introduction

Carbonic anhydrase IX (CAIX) is a protein that is expressed in several tumor types, including cervical cancer, and is rarely expressed in healthy tissues outside of the gastrointestinal tract [1]. Carbonic anhydrases belong to a family of proteins that catalyze the reversible hydration of carbon dioxide. CAIX activity and expression are associated with tumor-related hypoxia. There is increasing evidence that CAIX activity in tumors correlates significantly with increased invasiveness, increased likelihood of metastasis, and poor overall survival, suggesting that CAIX expression is an important indicator for prognostically relevant tumor pathology [2,3].

A reliable method of measuring of CAIX expression *in vivo* through imaging could become an important tool for assessing the cellular and organism-level response to novel tumor treatments during pre-clinical research. Optically active probes could be used to selectively image changes in CAIX expression throughout a course of treatment, either *in vitro* or *in vivo*. However, CAIX-based assessments of treatment efficacy should not be biased by any effects the probe may have on the activity of CAIX or other CA isoforms. Thus, an ideal probe for tumor monitoring should have high affinity and specificity for the CAIX isoform and should not interfere with its normal function.

Fluorescence-based imaging techniques are improving and may be used to obtain images through shallow tissue depths of up to a few centimeters [4]. Range is improved and interference from autofluorescence is minimized by utilizing fluorescent moieties that emit light in the near-infrared (NIR) range of the electromagnetic spectrum [4]. NIR wavelengths are poorly absorbed by water and

* Corresponding author at: Lakehead University, Department of Chemistry, Thunder Bay, ON P7B 5E1, Canada. Tel.: +1 807 766 7215; fax: +1 807 346 7775.
E-mail address: wely.floriano@lakeheadu.ca (W.B. Floriano).

hemoglobin, and therefore, use of NIR-emitting probes improves signal penetration and reduce changes in signal due to physiological composition [4]. Relatively few available molecules possess the desired optical qualities, making it unlikely to identify a small molecule intrinsically suitable for imaging, while also conferring strong affinity and selectivity for the desired target, CAIX. However, it is possible to conjugate a suitable reactive dye to a small molecule containing specific chemical functional groups to create a target-specific fluorescent probe.

In this paper, we propose 10 fluorescent CAIX probe candidates for experimental validation. Candidates were selected from a virtual database of 20,860 compounds, which were either fluorescently labeled or had published evidence of intrinsic fluorescence. The database was screened against a CAIX model using hierarchical virtual ligand screening (HierVLS), which combines a force field-based approach with a hierarchical framework to maximize efficiency for high-throughput screening applications [5]. Previously published experimental data on CAIX inhibitors was used to validate the methodology used [6]. Available sequence information for other CA isoforms was used to prioritize candidates based on the estimated “uniqueness” of each (putative or catalytic) binding site. The extensively studied structure and mechanism of carbonic anhydrases, in addition to the availability of a crystal structure for CAIX [7], influenced the decision to adopt a computational approach.

2. Methods and procedures

2.1. Preparation of the CAIX model structure

The only available experimental structure of the target protein, CAIX, was downloaded from the RCSB Protein Data Bank (www.rcsb.com). This structure (PDB Code: 3IAI) was obtained using X-ray diffraction at a resolution of 2.20 Å [7]. The biological assemblies associated with PDB: 3IAI are proposed homodimers. Chain A, corresponding to one CAIX monomer, was selected for modeling. Editing and structure quality checking of the protein structure were performed using the programs YASARA (www.yasara.org) and/or MOE (www.chemcomp.com), unless otherwise noted. Solvent molecules, co-crystallized ligands (except for Zn²⁺) and covalently linked carbohydrate moieties were removed from the structure. A mutated residue in the crystal structure (Cys41Ser) was replaced with its wild-type amino acid (Cys41). Missing atoms and side chains were identified and corrected. The CAIX model structure was then assigned CHARMM22 atomic charges and subjected to 2000 steps of conjugate-gradient minimization using the Dreiding force field, which contains parameters for zinc and easily handles the organic chemical compounds to be screened [8,9]. The choice of charges and force field was dictated by the virtual screening protocol, which uses a combination of Gasteiger atomic charges for ligands and CHARMM22 atomic charges for the protein, along with the Dreiding force field.

Structural validation tools including PROCHECK [10] and WHATCHECK [11] were used to check the quality of the resulting modeled structure. The CAIX model structure was compared to the experimental structure (PDB: 3IAI) to ensure that significant conformational changes were not induced by the energy minimization procedure. The model structure was aligned and superposed with 3IAI (Chain A) and all-atom and C-α RMSDs were calculated. This procedure provided a visual confirmation that residues proximal to the active site had retained their relative positions after modeling.

2.2. Multiple sequence alignment and binding site comparison

The amino acid sequences of 23 known CA isoforms were obtained from NCBI GenBank (<http://www.ncbi.nlm.nih.gov/genbank/>).

Table 1

The % sequence identity for residues within 5 Å of each binding region to the sequence alignment of 23 CA isoforms is shown, along with the distance from the catalytic site, volume and surface area of each site.

Region	Identity (%)	Distance from catalytic site (Å)	Volume (Å ³)	Surface area (Å ²)
1	14.6	21.3	337.52	354.98
2	67.2	n/a	378.45	375.62
3	17.4	13.8	423.58	499.69
4	29.2	11.7	461.53	565.70

<http://www.ncbi.nlm.nih.gov/genbank/>). A multiple sequence alignment was constructed using ClustalX 2.1 [12] with the following parameters: Gonnet matrix series, gap opening: 10, gap extension: 0.2, delay divergent sequences: 30%.

Putative binding pockets within the CAIX model structure were identified using PASS which is a rapid and efficient program that uses a geometry-based algorithm to identify putative binding pockets based their shape, size, and depth relative to the protein surface [13]. PASS is freely available and generates output that is generally compatible with applications used in later steps, such as HierVLS. Putative binding pocket centers were used to construct a graphical representation of each site within the model structure using MOE software (www.chemcomp.com). Residues with atoms between 0 and 5 Å of each predicted site center were identified, and their position was annotated with respect to the multiple sequence alignment of CA isoforms. Percentage of amino acid identity to CAIX at each marked position was determined, and each site was scored according to its uniqueness to the CAIX isoform (Table 1).

2.3. Known inhibitors library for structure validation

A literature review was conducted to identify 32 small molecule inhibitors of CAIX with reported K_i values in the 14–305 nM range [6]. Structures for these molecules were obtained either from PubChem (<http://pubchem.ncbi.nlm.nih.gov/>), BindingDB (www.bindingdb.org), or were constructed using MOE software (www.chemcomp.com) according to the structure reported in the literature. Inhibitor structures were assigned Gasteiger atomic charges and energy-minimized using the MMFF94x force-field [14,15].

2.4. Decoy library for virtual screening

A decoy library of chemical compounds was created for validation purposes. No available decoy databases were found for CAIX in two commonly used decoy repositories, DUD (<http://dud.docking.org/>) and DUDE (<http://dude.docking.org/>). Several target proteins that are structurally dissimilar to each other and that are also unrelated in function and structure to carbonic anhydrase were selected and the structures of known high-affinity ligands for these targets were downloaded from a publicly available repository of ligand structures and binding data known as BindingDB (www.bindingdb.org). Our assumption is that high affinity ligands for targets that are dissimilar to CAIX, will be unlikely to have any significant affinity for CAIX, and would therefore be suitable as decoy molecules. In addition, using ligands for multiple “targets” help ensure some chemical diversity in the decoy library. Specific inhibitors with K_i or IC_{50} values below 10 nM were obtained for the following targets (number of inhibitors in parentheses): neuraminidase (25), metabotropic glutamate receptor isoforms 1–8 (213), monoamine oxidase A and B (199), VEGFR (44), and EphB4 (39) [16–19]. 3D structures for these decoy ligands were generated using MOE (www.chemcomp.com) from available 2D SDF structures or canonical SMILES (simplified molecular-input line-entry system) strings. After eliminating 42 redundant molecules

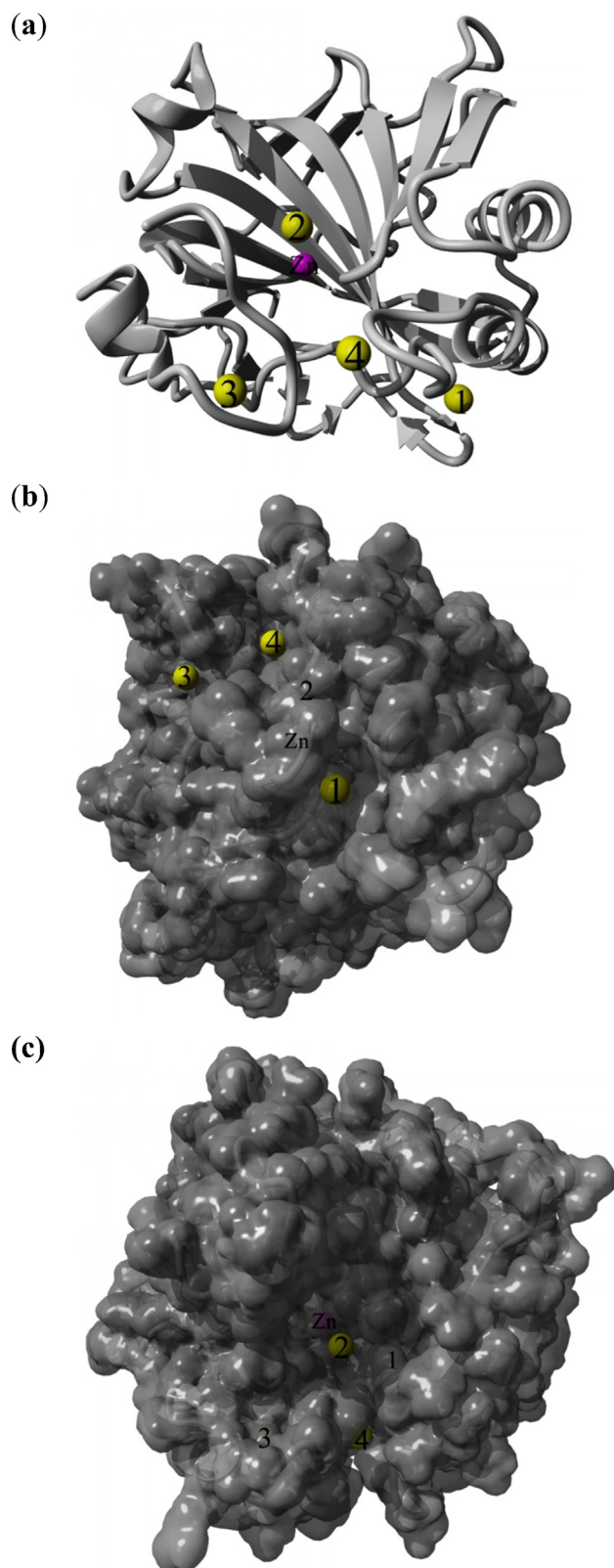


Fig. 1. (a) Cartoon representation of the CAIX model structure with the zinc cation shown in the center of a large binding cavity. The centers of putative binding sites 1, 3 and 4, and catalytic site 2 are shown as labeled spheres. Surface representation of the CAIX model with the centers of (b) sites 1, 3, and 4 and (c) site 2 (with Zn^{2+} shown), shown as spheres. (b) and (c) are rotated images from (a) to better show the cavities. Based on site conservation among CA isoforms, site 1 is the most preferable for binding, 3 the second most, and 4, the least.

from the database of 520 decoys, 478 unique entries remained. Decoys were cross-referenced for any documented activity against any carbonic anhydrase isoform in secondary BindingDB queries. Various chemical descriptors such as molecular weight, number of hydrogen bond donors and acceptors, and Log *P* were calculated using MOE (www.chemcomp.com) for the decoy set, and their distributions were compared to those of the inhibitor library. The goal was to ensure that the compounds in the decoy library would reasonably represent the compounds contained in the inhibitor library such that they could serve as plausible decoys. Molecule structure files were prepared for screening by removing counter ions and solvent molecules and eliminating molecules containing unparameterized atoms. Ligand structures were energy-minimized using the MMFF94x force-field [14,15]. Various molecular descriptors were calculated for both the inhibitor and decoy set using MOE software (www.chemcomp.com) in order to compare the distribution of properties between the groups.

2.5. Validation screening

The combined library of 478 decoys and 32 known CAIX inhibitors (510 compounds total) was computationally screened against the 3D model of CAIX. PASS [13] was used for putative binding site identification and HierVLS for molecular docking and force field-based scoring. The steps and parameters of the HierVLS docking and scoring scheme have been described previously [5]. HierVLS is a hierarchical approach to virtual ligand screening that tests the largest number of bound conformations of each ligand using the least computationally expensive methods (course-grain conformational search and protein-fixed minimization), and proceeds to use more computational resources to obtain more accurate predictions of binding energy for the most promising conformers (all-atoms energy minimization with solvation model) [5]. Screening was initiated using a graphical user interface (GUI) known as Cassandra, and calculations were run using the SHARCNET high-performance computing network (www.sharcnet.ca) [20]. For each ligand, a maximum number of 10,000 docked conformations were generated in level 0 (course-grain conformational search), with the 1500 best conformers passed to the filter step. Conformations were eliminated by the filter if they had less than 70% buried surface area. 150 conformations per ligand were passed from the filter step through to level 1 (protein-fixed minimization). The best 3 docked conformations for each ligand were subject to an all-atoms minimization (level 2), the most computationally expensive step of the process. Final force field-based binding scores (kcal/mol) consider

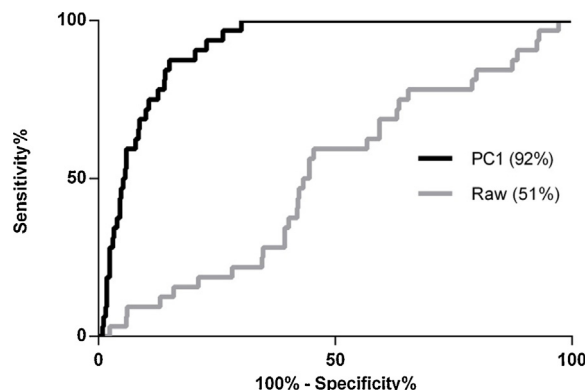


Fig. 2. The ROC curves for the PCA (gray) and raw (force field-based; gray) scoring schemes used for the validation screen of CAIX. The area under each curve (AUC), a measurement of predictive power, is shown for each scoring scheme in parentheses. At 92% AUC, PCA significantly outperforms raw (force field-based) scores (51% AUC), and it is effective at identifying binders from non-binders.

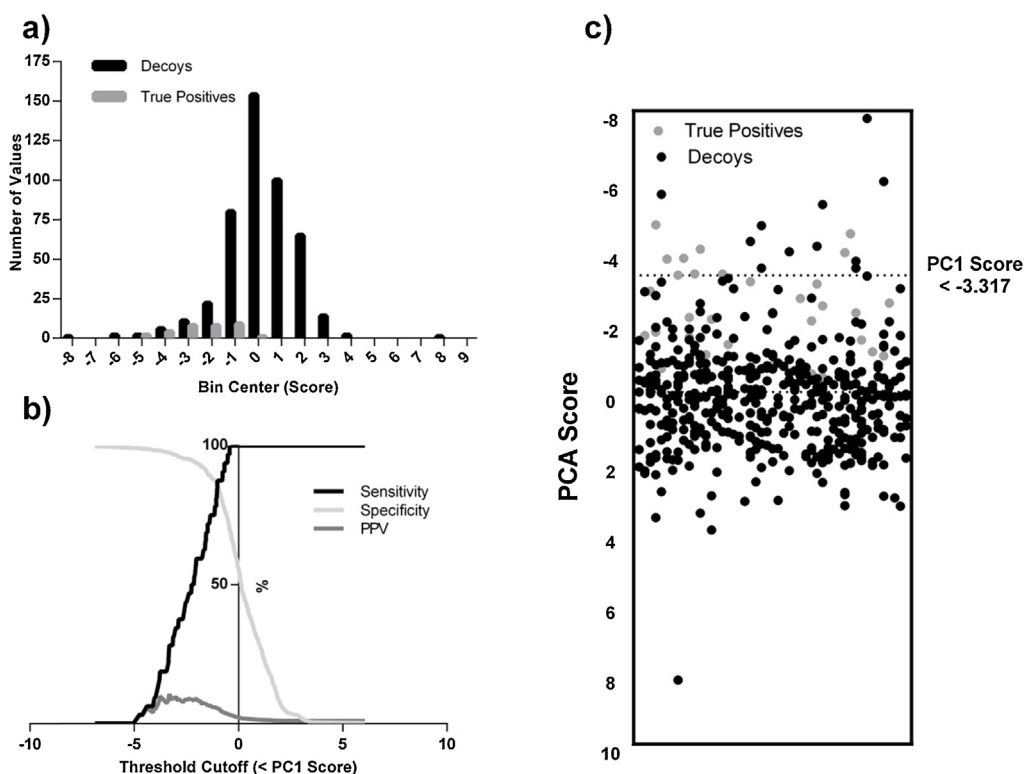


Fig. 3. (a) Frequency histogram of PC1 scores for decoys and true positives in the validation library. (b) Specificity, sensitivity, and positive predictive value (PPV) of the PCA scheme determined using the validation library. The highest possible PPV value, assuming a 1% prevalence of hits in a database, occurs at a threshold of -3.32 . Relatively few compounds in the validation library scored below the threshold value of -3.32 (c).

the desolvation effects of the protein/ligand complex by applying the Analytical Volume Generalized Born solvation model to the complex, and the protein and ligand alone. Force field-based (raw) binding scores were collected for each ligand docked to each of the four identified sites. Raw binding scores for each ligand in all regions were subjected to principal components analysis (PCA) using PAST statistical software (<http://folk.uio.no/ohammer/past/>). In PCA, the number of variables in a dataset is transformed into orthogonal principal components (PC) that retain most of the information observed in the dataset. Each PC is a linear combination of the original force field-based binding energies. The first principal component (PC1) usually carries most of the variance in the dataset, and it is used in our approach as a predictive binding score. Ligands that failed the buried surface cut-off filter in HierVLS for any of the 4 sites were excluded from analysis (18 decoys), leaving 460 decoys and 32 true positives scored. PCA was performed using a correlation matrix type with singular value decomposition (SVD) enabled and iterative imputation to handle missing values with the following other parameters: Boot $N=0$, Jolliffe cut-off=0.7. Analysis of the test characteristics for PCA was performed using PRISM [21] and Microsoft Excel 2010. The correlation matrix option normalizes all binding energy values by their standard deviations and, hence, the PC1 scores are not on the same scale as the original binding energies (raw scores).

2.6. CAIX virtual screening

The CAIX model was screened against two compound libraries. One library contains 14,862 primary-amine containing compounds that were identified from Molecular Libraries Initiative Small Molecule Repository (MLSMR) and tagged virtually with Atto680 NHS-ester (database available with supplementary material). Atto680 is a near-IR fluorophore that is commercially

available in a reactive N-hydroxysuccinimidyl ester form [22]. The preparation of this database is described in a separate manuscript [23]. An additional subset of the MLSMR, was identified using PubChem BioAssay data. Molecules with reported fluorescence activity in a published series of high-throughput assays for spectroscopic profiling (PubChem Bioassay IDs: 587–594) [19] were selected, totalling 5998 compounds. These are compounds found to exhibit fluorescence activity in spectral regions common to well-known dyes such as fluorescein. The overall number of compounds screened against CAIX was 20,860. The parameters used for screening were identical to those used during the validation procedure with the exception of the buried surface area cut-off for docked ligands, which was adjusted from 70% to 40% when screening the fluorescently-labeled compounds to reflect their larger overall size.

“Hit” candidates were identified based on PCA scoring scheme and binding thresholds as described in the validation screening. Best conformers for the binding candidates were examined for favorable protein–ligand interactions using MOE software (www.chemcomp.com). Candidates were also assessed for their binding site preference by comparing raw scores between specific sites relative to each site’s mean scores and score standard deviation for that database.

3. Results and discussion

3.1. Preparation of the CAIX model structure

A model structure was prepared from the 3IAI PDB X-ray crystal structure of CAIX. A Ramachandran plot was generated, and PROCHECK/WHATCHECK quality analyses performed. Psi-phi bond angles (Ramachandran) appeared to be consistent with a reasonable structure. Quality analyses were used to check to ensure that

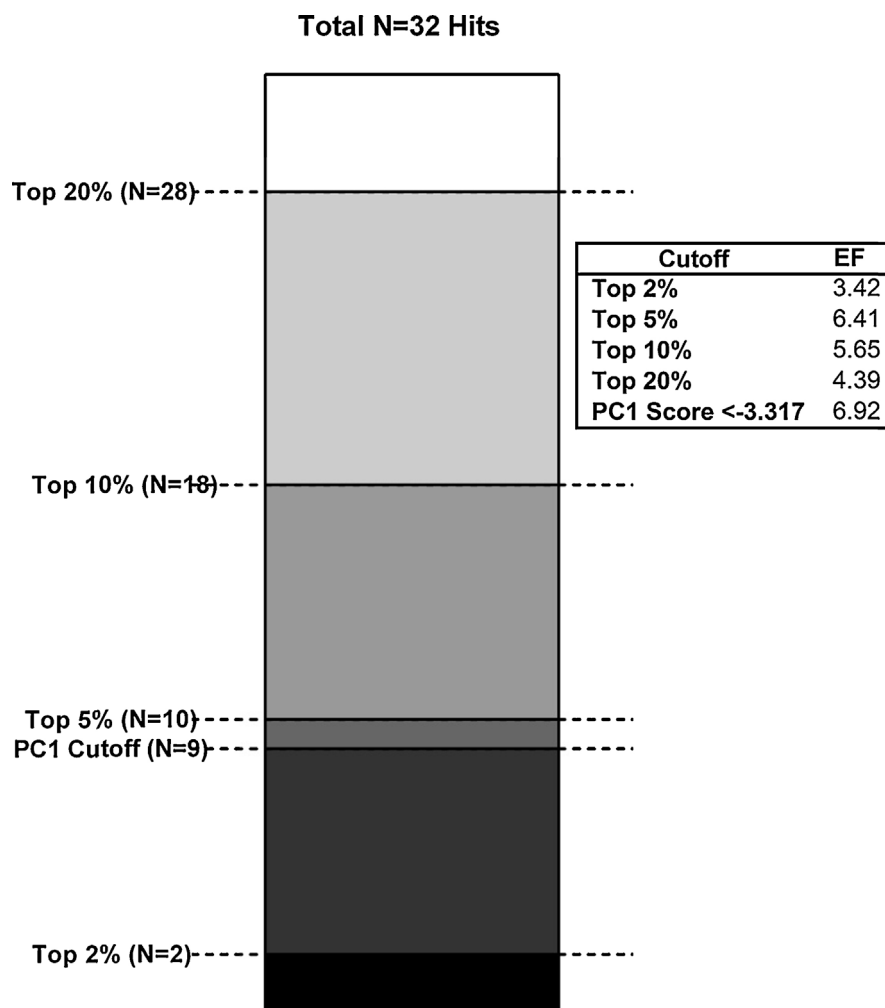


Fig. 4. The number of true positives found (N) using each indicated threshold (i.e., top $x\%$ of validation library screened) as a fraction of the total number of true positives present in the validation library ($N=32$). Enrichment factors (EFs) were calculated for each of the indicated threshold values and shown in tabulated format. The PC1 score threshold that was selected using ROC analysis gave the highest EF value, 6.92.

missing side chains (and atoms) had been added during the modeling process, and that the starting and final structures did not have any serious structural abnormalities. The model was compared to the crystal structure by calculating the all-atoms RMSD which was found to be 0.354 Å. When a α C RMSD was performed, results also showed minimal changes in the structure (0.19 Å mean RMSD). Only 10 residues had an RMSD greater than 2 standard deviations from the mean: G9 (0.48 Å), E48 (0.47 Å), G63 (0.46 Å), D132 (0.49 Å), G136 (0.039 Å), E153 (0.43 Å), E173 (0.47 Å), G178 (0.39 Å), N193 (0.38 Å) and P261 (0.56 Å). The total number of residues in the model is 254 and the maximum α C RMSD was 0.56 Å for P261, which is a terminal residue. Combined with the quality results, these comparisons gave us confidence that our model would be representative of the actual structure. A ribbon representation of the CAIX model is shown in Fig. 1(a), with the centers of the four sites identified by PASS shown as spheres labeled 1 to 4. Putative binding sites 1, 3, and 4 are shown as surfaces in Fig. 1(b). Site 2, shown as surface in Fig. 1(c), corresponds to the catalytic site (Zn^{2+} shown) and has the most buried surface of the four sites.

3.2. Multiple sequence alignment and binding site comparison

A multiple sequence alignment of 23 human isoforms was performed. All residues within 5 Å of each (putative or catalytic) binding site of CAIX were compared for identity using the sequence

alignment. The distances from the center of each putative binding site to the catalytic site (site 2) were also calculated. A summary of % homology and distance to the catalytic site for all sites is shown in Table 1. Based on this analysis, hits will be prioritized for experimental testing if they show selectivity for sites 1 or 3, which show the least homology to other CAIX isoforms, and largest distance from the catalytic site. This will increase the chances of finding molecular probes that are selective for CAIX. For comparison, the molecular surface area and enclosed volumes of each site is included in Table 1. Molecular surfaces and volumes were calculated using the program Yasara (www.yasara.org) based on the spheres generated in the first step of HierVLS to fill each site. The smallest site is 1, whereas the largest is site 4.

3.3. Validation screening

Principal component analysis was used to transform raw scores (force field-based binding energies) corresponding to each ligand bound to each of 4 sites into normalized, orthogonal scores, the principal components 1–4. The first principal component (PC1) was selected for analysis because it carries the most variance (66%) in the calculated binding affinity it represents. PC1 scores were analyzed for a test set of 32 known CAIX inhibitors [6] and 460 decoys. Raw (force field-based) scores from region 2, which

corresponds to the active site, were also analyzed because 32 inhibitors in the validation site bind to the active site of the protein and, hence, their binding affinities for that site may correlate with experimental values. Fig. 2 represents the sensitivity and specificity of the PC1 scheme relative to the raw scoring scheme using a receiver–operating curve (ROC). ROCs provide a graphical way to compare the true positive rate (sensitivity) and the false positive rate (1-specificity) of different prediction schemes. They are commonly used in medicine to assess the accuracy of clinical tests and determine a cutoff value that best discriminates normal from abnormal results [24]. Here, we use ROCs to assess and compare the ability of the two different scores (PC1 and raw) in discriminating true from false binders. Sensitivity is calculated as the number of true positives divided by the number of all known positives in the set. Specificity corresponds to the number of true negatives divided by the total number of true negatives which, in our test set, are the decoy ligands. With respect to the ROC curve, increasing sensitivity and specificity values are associated with a shift in the curve toward the topmost and leftmost parts of the graph, respectively.

The accuracy of each scoring scheme can be measured by their respective area under the ROC curve (AUC). AUC is a commonly used metric for evaluating diagnostic test performance, and signifies the probability that a random true positive will be ranked better than a random true negative [25]. The PC1 scoring scheme performed the best with an AUC of 0.92, compared to an AUC of 0.51 obtained for the raw scores. We concluded that PCA greatly improved the discrimination between true binders and decoys relative to the raw scores. Principal component analysis of binding energies for multiple binding sites within a target protein is a novel approach for virtual screening hit selection, and these results are very encouraging.

A frequency histogram of PC1 scores for both decoys and true positives is shown in Fig. 3a. It is evident from this distribution that the PC1 scores for true positives lay toward the negative side of the distribution. In order to choose a threshold below which candidates will be selected for experimental testing, the positive predictive value (PPV) of the PCA scheme was examined. The PPV (Eq. (1)) signifies the probability that a compound with a score that is better than the threshold will be a true binder. The PPV depends on the sensitivity and specificity of the selected threshold, as well as the prevalence of true positives in the database being screened.

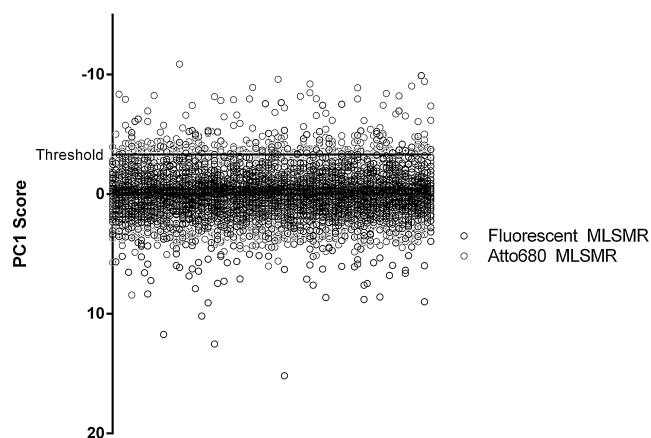


Fig. 5. A distribution of scores for molecules screened in the Atto680-tagged (light gray) and Fluorescent MLSMR (dark gray) databases. The threshold PC1 score -3.32 is marked. The top 20 compounds passing this threshold for each library were further analyzed, and a final list of 10 compounds is suggested for experimental testing.

The positive predictive value (PPV) is the fraction of all positives, true (TP) and false (FP), that are true positives.

$$PPV = \frac{TP}{TP + FP} \quad (1)$$

To maximize the chances of finding true positives among hits selected for experimental testing, we want to identify a threshold score that yields a high PPV. This can be achieved by enhancing the number of true positives (*i.e.*, increasing sensitivity) or by reducing the number of false positives (*i.e.*, increasing specificity). It is acceptable if our selected threshold has a relatively low sensitivity, as it is unnecessary for us to successfully identify *all* true positives within a database. For a virtual screening campaign to be considered successful, it is sufficient to identify *some* true positives from a large database. It is, thus, desirable to adjust our threshold for better selectivity. It should be noted, however, that if we believed

Table 2

Top 20 ranked compounds screened from both the Atto680-tagged and Fluorescent MLSMR databases.

Fluorescent MLSMR		Atto680 MLSMR	
PubChem CID	PC1 score	PubChem CID ^a	PC1 score
CID652931	−9.93	CID16746148	−10.89
CID5154	−7.67	CID5939529	−9.61
CID1256741	−7.58	CID666466	−9.44
CID2999504	−7.54	CID5939530	−9.23
CID665212	−7.44	CID20846957	−9.06
CID854591	−6.97	CID666309	−8.93
CID446849	−6.29	CID667134	−8.91
CID649288	−6.10	CID665941	−8.50
CID2368420	−5.35	CID535684	−8.44
CID441975	−5.33	CID666821	−8.38
CID656158	−5.32	CID3153996	−8.26
CID648233	−5.31	CID1939105	−8.23
CID5389398	−5.26	CID3104972	−8.20
CID662206	−5.25	CID25163910	−8.16
CID3243178	−5.24	CID6916824	−7.99
CID1301431	−5.22	CID8795	−7.98
CID666072	−5.15	CID666424	−7.98
CID4849	−5.10	CID1491128	−7.95
CID16231	−5.09	CID667344	−7.91
CID5546	−5.06	CID1853629	−7.85

^a PubChemCID given for parent compound of Atto680-tagged conjugate.

Table 3

Best five compounds with suggested binding preference for region 1 from the Fluorescent MLSMR database. Raw scores were examined relative to mean, and expressed as $N \times$ of standard deviations from the mean score.

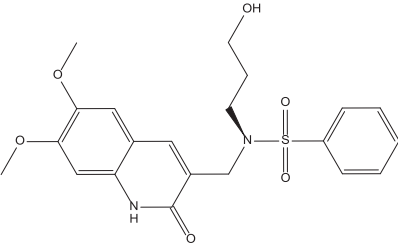
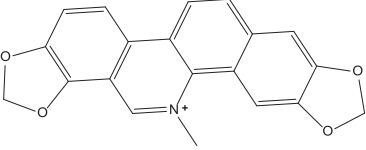
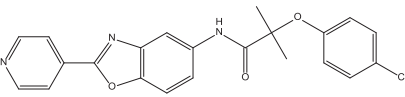
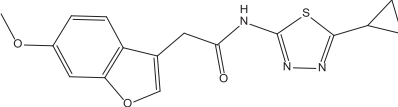
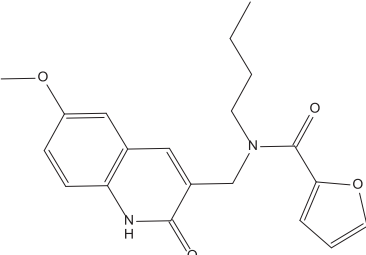
PubChem CID	Raw scores	$N \times$ std. dev. above mean raw site 1 score	$N \times$ std. dev. above mean raw site 2 score
CID652931	−60.88	3.7	0.3
CID5154	−56.82	3.1	0.5
CID1256741	−51.69	2.3	−0.5
CID2999504	−50.75	2.2	−0.7
CID665212	−51.67	2.3	−0.4

Table 4

Best five compounds with suggested binding preference for region 1 from the Atto680-tagged MLSMR database. Raw scores were examined relative to mean, and expressed as $N \times$ of standard deviations from the mean score.

PubChem CID ^a	Raw scores	$N \times$ std. dev. above mean raw site 1 score ^a	$N \times$ std. dev. above mean raw site 2 score
CID667134	−8.91	6.6	5.3
CID535684	−8.44	7.8	4.6
CID666821	−8.38	5.6	4.4
CID3104972	−8.20	5.9	2.8
CID8795	−7.98	5.1	4.4

Table 5
Chemical structures of the top 5 scoring fluorescent MLSMR compounds.

PubChem CID	Compound name(s)	Structure	MW (g/mol)	H-bond Donor (N)	H-bond Acceptor (N)	TPSA	XLogP	ADME/toxicology/data
652931	N-[(6,7-dimethoxy-2-oxo-1H-quinolin-3-yl)methyl]-N-(3-hydroxypropyl)benzenesulfonamide		432.49	2	7	114	1.6	Not found
5154	Sanguinarin; Dimethylene-dioxy benzphenanthridine		332.33	0	4	40.8	4.4	Yes
1256741	STK236667; 2-(4-chlorophenoxy)-2-methyl-N-[2-(pyridin-4-yl)-1,3-benzoxazol-5-yl]propanamide		407.85	1	5	77.2	4.4	Not found
2999504	N-(5-cyclopropyl-1,3,4-thiadiazol-2-yl)-2-(6-methoxy-1-benzofuran-3-yl)acetamide		329.37	1	6	106	2.4	Not found
665212	N-butyl-N-[(6-methoxy-2-oxo-1H-quinolin-3-yl)methyl]furan-2-carboxamide		354.40	1	4	71.8	2.9	Yes

that there were very few true positives in a database, or if we were using this type of screening for another application where identifying all actives were a priority, we could adjust the threshold for better sensitivity.

Assuming a prevalence of true positives of 1% in a hypothetical database, we graphed the PPV across a range of scoring thresholds, also including associated specificity and sensitivity values (Fig. 3b). The PC1 score threshold that gave the best PPV was −3.32. At this threshold and for a database containing 1% true positives, the PPV value is 10.6%. The corresponding sensitivity and specificity for this threshold are 28.1% and 97.6% respectively. A PPV of ~11% for the PC1 scores suggests that at least 10–15 top candidates should be

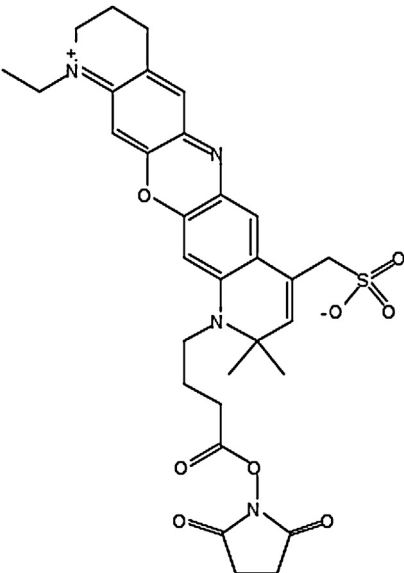
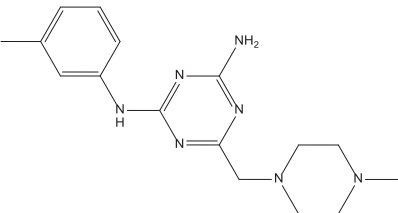
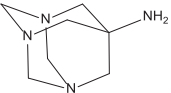
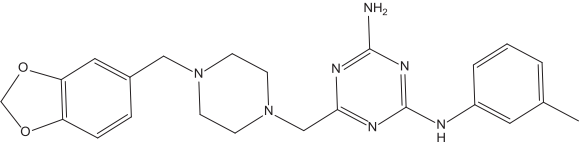
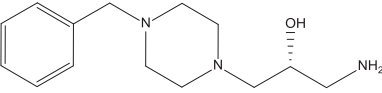
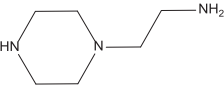
selected for experimental testing from screening experiments to increase the likelihood of identifying a true binder. Fig. 3c shows a distribution of the PC1 scores for the validation set, with decoys and true positives shown with respect to the chosen threshold.

The enrichment factor (EF) is a metric commonly used to describe the performance of virtual screening methods [26]. EF gives the ratio of true positives detected in the top x% of scored compounds to the number that would have been expected based on random selection.

The enrichment factor (EF) is given by the ratio of the number of hits (n) found above the threshold (top x%) to the expected number of hits, the proportion (P) of overall hits in the database multiplied

Table 6

Chemical structures of the top 5 scoring Atto680-tagged MLSMR compounds and selected properties computed using MOE software. Only the primary amine base moiety is shown. The Atto680 NHS ester is attached to the primary amine base moiety through an amide bond [23].

Base compound PubChem CID	Base compound name	Structure	MW (g/mol)	H-bond Donor (N)	H-bond Acceptor (N)	TPSA	SLogP
16218508	Atto680 NHS ester ^a (Labelling dye)		622.69	0	10	157 ^b	0.5 ^b
667134	STK578101; 2-N-(3-methylphenyl)-6-[(4-methylpiperazin-1-yl)methyl]-1,3,5-triazine-2,4-diamine		823.04	2	5	167.7	2.6
535684	1,3,5-triazatricyclo[3.3.1.1~3,7~]decan-7-amine		664.85	1	2	121.5	-2.0
666821	STK593016; 6-[[4-(1,3-benzodioxol-5-ylmethyl)piperazin-1-yl]methyl]-2-N-(3-methylphenyl)-1,3,5-triazine-2,4-diamine		943.14	2	7	186.2	4.2
3104972	1-Amino-3-(4-benzyl-piperazin-1-yl)-propan-2-ol; AC1MJ1CE; (2R)-1-amino-3-(4-benzylpiperazin-1-yl)propan-2-ol		758.99	2	3	137.3	1.3
8795	Aminoethylpiperazine; 2-piperazin-1-ylethanamine		638.83	1	2	129.2	0.13

^a Note that the Atto680 NHS ester is shown for reference (data obtained from PubChem entry) but only conjugated (base compound + Atto680 moiety) structures were screened.

^b Properties for Atto680 NHS ester were obtained from available PubChem data (CID16218508).

by the number of compounds overall (N) above the threshold (top $x\%$).

$$EF = \frac{n_{\text{hits in top } x\%}}{P_{\text{hits}} N_{\text{in top } x\%}} \quad (2)$$

An enrichment factor of 1 would indicate that the method is no better at selecting true positives than random selection. The enrichment factors are given for the chosen threshold (PC1 score < -3.317) as well as for the top 2%, 5%, 10%, and 20% of scored compounds in the validation library in Fig. 4. This figure also displays the true positives (hits) identified at each threshold as a proportion of the total number of true positives ($N=32$) in the validation library. The PC1 threshold score (< -3.32) gave the best enrichment factor, 6.92. The interpretation of this EF is that the probability of a true positive occurring among compounds selected using the chosen threshold is increased relative to random selection within the overall database by a factor of almost 7. When systematically assessed for performance with test ligand sets across multiple targets, 11 different docking programs were shown to have enrichment factors between 0.0 and 10.0 for the top 10% of their databases [27]. For the PCA scoring scheme, $EF=5.65$ for the top 10% of PCA scores, indicating that our observed EFs agree reasonably well with values reported for other screening protocols. Comparison of enrichment factors with those obtained in the literature for other docking and scoring methods gives us confidence that our scoring scheme has significant predictive value for selecting binding candidates for CAIX.

3.4. CAIX screening

A total of 20,860 compounds were screened against the CAIX model. Scores were adjusted using principal components analysis (PCA) which was determined in the validation step to be a highly predictive scoring scheme. APC1 score threshold of -3.32 was used, below which molecules were considered candidates for binding. A distribution of the PC1 scores for molecules screened is shown in Fig. 5. Table 2 summarizes 20 top scoring binding candidates passing the PC1 threshold, for both the Atto680-tagged and Fluorescent MLSMR libraries. These compounds are likely to bind experimentally to CAIX. However, the scoring method used for ranking the compounds, PCA, does not retain information about binding site preferences. In order to identify candidates preferentially binding to isoform-selective sites among those with above threshold PC1 score, we turn our attention back to raw scores. For each of the candidates in Table 2, best scoring conformers from each (putative or catalytic) binding site were examined, and site-specific raw scores compared. Scores were examined for number of standard deviations above the mean score for that site (in a particular library). Putative binding site 1 had been selected previously as the most preferred site based on lack of homology between other isoforms, and distance from catalytic site. The site specific scores relative to the mean are summarized in Tables 3 and 4 for the best 5 compounds in each of the libraries. Each of these 5 ligands (or in the case of the Atto680-tagged library, the parent unlabelled primary amine) is shown in Tables 5 and 6.

The top 10 candidates shown in Tables 5 and 6 were, thus, selected by a two-step approach: (1) candidates are selected based on PC1 scores; (2) raw scores are used to identify which among these candidates prefer to bind to isoform-selective sites. This two-step method is useful as a rough approximation of isoform selectivity. An alternate approach could include developing homology models for the most common isoforms, and cross-screening promising candidates against them in order to estimate isoform specificity. Such an approach would require additional time and computational resources, but would be recommended as part of a subsequent investigation prior to experimental testing.

The compounds suggested for experimental testing are either fluorescent (MLSMR library candidates) or can be synthesized by conjugation of a commercially available reactive fluorophore to also commercially available base chemicals (Atto680-tagged candidates). Binding and isoform selectivity of these compounds can be tested using well-established assay platforms such as fluorescence polarization [28]. Compounds found to bind selectively to CAIX and possessing good spectrophotometric characteristics may be used as molecular probes in fluorescence-based assays for CAIX detection and quantitation.

4. Conclusions

A library of known CAIX ligands and representative decoy molecules was used to validate a scoring scheme based on a principal component analysis of force field binding energies obtained for multiple docking sites within CAIX. The first principle component of this analysis (PC1) achieved a 92% probability of correctly attributing a better (lower, in our case) score to a true positive compared to non-binder in a ROC analysis. A score threshold corresponding to the maximum positive predictive value was identified. The enrichment factor (EF) for screening the validation database was 5.65 for the top 10% of compounds, meaning that the probability of finding a true hit was increased six times relative to random selection, a value comparable to other virtual screening studies. This gives us confidence in the ability of our methodology to predict the true binding of fluorescent probes to CAIX. The use of first component analysis of binding energies for multiple binding sites within a protein is a novel approach for hit selection. Although a larger validation must be performed, this methodology may be useful for the analysis of results from other virtual screening platforms.

The validated PCA scoring scheme and additional post-screening criteria were used to screen a virtual library of 20,860 fluorescent or fluorophore-tagged structures against a CAIX model. Ten promising probe candidates have been identified for experimental testing against CAIX and we predict that the post-screening criteria used for ranking of these compounds will impart additional selectivity against CAIX isoforms than would have otherwise been present.

Acknowledgements

This work was supported by funds from the Thunder Bay Regional Research Institute (TBRRI), the Natural Sciences and Engineering Research Council of Canada (NSERC), and Royal Bank of Canada Dr. M. Poznansky Mentorship Development Award. Virtual ligand screening calculations utilized computational resources provided by the Shared Hierarchical Academic Research Computing Network (SHARCNET; www.sharcnet.ca) and the Lakehead University's High Performance Computing Centre (LUHPCC).

References

- [1] P. Swietach, R.D. Vaughan-Jones, Regulation of tumor pH and the role of carbonic anhydrase 9, *Cancer Metastasis Rev.* 26 (2007) 299–310.
- [2] J.A. Lancaster, A.L. Harris, S.E. Davidson, J.P. Logue, R.D. Hunter, C.C. Wycoff, et al., Carbonic anhydrase (CA IX) expression, a potential new intrinsic marker of hypoxia: correlations with tumor oxygen measurements and prognosis in locally advanced carcinoma of the cervix, *Cancer Res.* 61 (2001) 6394–6399.
- [3] L. Woelber, K. Kress, J.F. Kersten, M. Choschzick, E. Kilic, U. Herwig, et al., Carbonic anhydrase IX in tumor tissue and sera of patients with primary cervical cancer, *BMC Cancer* 11 (2011) 12.
- [4] V. Ntziachristos, C. Bremer, R. Weissleder, Fluorescence imaging with near-infrared light: new technological advances that enable in vivo molecular imaging, *Eur. Radiol.* 13 (2003) 195–208.
- [5] W.B. Floriano, N. Vaidehi, G. Zamanakos, W.A. Goddard, HierVLS hierarchical docking protocol for virtual ligand screening of large-molecule databases, *J. Med. Chem.* 47 (2004) 56–71.
- [6] D. Vullo, M. Franchi, E. Gallori, J. Pastorek, A. Scozzafava, S. Pastorekova, et al., Carbonic anhydrase inhibitors: inhibition of the tumor-associated isozyme

- IX with aromatic and heterocyclic sulfonamides, *Bioorg. Med. Chem. Lett.* 13 (2003) 1005–1009.
- [7] V. Alterio, M. Hilvo, A. Di Fiore, C.T. Supuran, P. Pan, S. Parkkila, et al., Crystal structure of the catalytic domain of the tumor-associated human carbonic anhydrase IX, *Proc. Natl. Acad. Sci. U. S. A.* 106 (2009) 16233–16238.
- [8] A.D. MacKerell, D. Bashford, R.L. Dunbrack, J.D. Evanseck, M.J. Field, S. Fischer, et al., All-atom empirical potential for molecular modeling and dynamics studies of proteins, *J. Phys. Chem. B* 102 (1998) 3586–3616.
- [9] S.L. Mayo, B.D. Olafson, W.A. Goddard, DREIDING: a generic force field for molecular simulations, *J. Phys. Chem.* 94 (1990) 8897–8909.
- [10] R.A. Laskowski, M.W. MacArthur, D.S. Moss, J.M. Thornton, PROCHECK: a program to check the stereochemical quality of protein structures, *J. Appl. Crystallogr.* 26 (1993) 283–291.
- [11] R. Hooft, G. Vriend, C. Sander, E. Abola, Errors in protein structures, *Nature* 381 (6580) (1996), 272–272.
- [12] M.A. Larkin, G. Blackshields, N.P. Brown, R. Chenna, P.A. McGettigan, H. McWilliam, et al., Clustal W and Clustal X version 2.0, *Bioinformatics* (Oxford, England) 23 (2007) 2947–2948.
- [13] G.P. Brady, P.F. Stouten, Fast prediction and visualization of protein binding pockets with PASS, *J. Comput. Aided Mol. Des.* 14 (2000) 383–401.
- [14] J. Gasteiger, M. Marsili, Iterative partial equalization of orbital electronegativity—a rapid access to atomic charges, *Tetrahedron* 36 (1980) 3219–3228.
- [15] T.A. Halgren, Merck molecular force field. I. Basis, form, scope, parameterization, and performance of MMFF94, *J. Comput. Chem.* 17 (1996) 490–519.
- [16] C. Bardelle, B. Barlaam, N. Brooks, T. Coleman, D. Cross, R. Ducray, et al., Inhibitors of the tyrosine kinase EphB4. Part 3: Identification of non-benzodioxole-based kinase inhibitors, *Bioorg. Med. Chem. Lett.* 20 (2010) 6242–6245.
- [17] C. Bardelle, T. Coleman, D. Cross, S. Davenport, J.G. Kettle, E.J. Ko, et al., Inhibitors of the tyrosine kinase EphB4. Part 2: Structure-based discovery and optimisation of 3,5-bis substituted anilino pyrimidines, *Bioorg. Med. Chem. Lett.* 18 (2008) 5717–5721.
- [18] C. Bardelle, D. Cross, S. Davenport, J.G. Kettle, E.J. Ko, A.G. Leach, et al., Inhibitors of the tyrosine kinase EphB4. Part 1: Structure-based design and optimization of a series of 2,4-bis-anilino pyrimidines, *Bioorg. Med. Chem. Lett.* 18 (2008) 2776–2780.
- [19] A. Simeonov, A. Jadhav, C.J. Thomas, Y. Wang, R. Huang, N.T. Southall, et al., Fluorescence spectroscopic profiling of compound libraries, *J. Med. Chem.* 51 (2008) 2363–2371.
- [20] Z.H. Ramjan, A. Raheja, W.B. Floriano, A cluster-aware graphical user interface for a virtual ligand screening tool, in: 30th Annual International Conference of the IEEE, Engineering in Medicine and Biology Society (EMBS), 2008, 2008, pp. 4102–4105.
- [21] GraphPad Software, PRISM. GraphPad Software, La Jolla, CA, 2013.
- [22] V. Buschmann, K.D. Weston, M. Sauer, Spectroscopic study and evaluation of red-absorbing fluorescent dyes, *Bioconjugate Chem.* 14 (2003) 195–204.
- [23] R.L. Kamstra, S. Dadgar, J. Wigg, W.B. Floriano, Creating and virtually screening databases of fluorescently-labeled compounds for the discovery of target-specific molecular probes, *J. Comput. Aided Mol. Des.* (2014), <http://dx.doi.org/10.1007/s10822-014-9789-0> [Epub ahead of print].
- [24] M.H. Zweig, G. Campbell, Receiver–operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine, *Clin. Chem.* 39 (1993) 561–577.
- [25] J. Kirchmair, P. Markt, S. Distinto, G. Wolber, T. Langer, 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.* 22 (2008) 213–228.
- [26] W.B. Floriano, N. Vaidehi, G. Zamanakos, W.A. Goddard 3rd., HierVLS hierarchical docking protocol for virtual ligand screening of large-molecule databases, *J. Med. Chem.* 47 (2004) 56–71.
- [27] G.L. Warren, C.W. Andrews, A.-M. Capelli, B. Clarke, J. LaLonde, M.H. Lambert, et al., A critical assessment of docking programs and scoring functions, *J. Med. Chem.* 49 (2005) 5912–5931.
- [28] T.J. Burke, K.R. Loniello, J.A. Beebe, K.M. Ervin, Development and application of fluorescence polarization assays in drug discovery, *Comb. Chem. High Throughput Screening* 6 (2003) 183–194.