JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

# ColBioS-FlavRC: A Collection of Bioselective Flavonoids and Related Compounds Filtered from High-Throughput Screening Outcomes

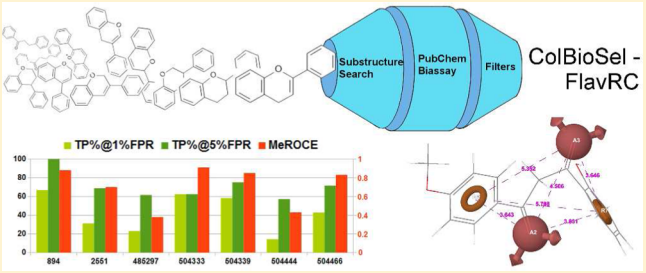
Sorin I. Avram,[†] Liliana M. Pacureanu,[†] Alina Bora,[†] Luminita Crisan,[†] Stefana Avram,[‡] and Ludovic Kurunczi*[,†,§]

[†]Department of Computational Chemistry, Institute of Chemistry Timisoara of Romanian Academy, Mihai Viteazul Avenue, 24, Timisoara, 300223, Romania

[‡]Faculty of Pharmacy, Department Pharmacy II, Discipline of Pharmacognosy and [§]Faculty of Pharmacy, Department Pharmacy I, Discipline of Physical Chemistry, University of Medicine and Pharmacy "Victor Babes", Eftimie Murgu 2, Timisoara, 300041, Romania

Supporting Information

**ABSTRACT:** Flavonoids, the vastest class of natural polyphenols, are extensively investigated for their multiple benefits on human health. Due to their physicochemical or biological properties, many representatives are considered to exhibit low selectivity among various protein targets or to plague high-throughput screening (HTS) outcomes. The aim of this study is to highlight reliable, bioselective compounds sharing flavonoidic scaffolds in HTS experiments. A filtering scheme was applied to remove undesired flavonoids (and related compounds) from confirmatory PubChem bioassays. A number of 433 compounds addressing various protein targets form the core of the collection of bioselective flavonoids and related compounds (ColBioS-FlavRC). With an additional set of 2908 inactive related compounds, ColBioS-FlavRC offers the grounds for method optimization and validation. We exemplified the use of ColBioS-FlavRC by pharmacophore modeling, subsequently (externally) validated for virtual screening purposes. The early enrichment capabilities of the pharmacophore hypotheses were measured by means of the median exponential retriever operating curve enrichment (MeROCE), a suited metric in comparative evaluations of virtual screening methods. ColBioS-FlavRC is available in the Supporting Information and is freely accessible for further studies.




## INTRODUCTION

Flavonoids are polyphenols widely spread in nature, with more than eight thousand natural derivatives indexed in 2006.[1] Structurally, flavonoids share the diphenylpropane (C6−C3−C6) skeleton as exemplified by the commonly known phenylbenzopyran structure. Depending on the position of the phenyl ring, flavonoids are classified as common flavonoids (2-phenylbenzopyrans; CF), isoflavonoids (3-phenylbenzopyrans; IF), neoflavonoids (4-phenylbenzopyrans; NF). Minor flavonoids (MF) established themselves as a fourth group, comprising chalcones and aurones (2-benzilidenecoumaranone).[2,3] The degree of oxidation and saturation of the heterocyclic C-ring defines subclasses of flavonoids.[3] The chemical structures represented in the left side of Figure 1 are main chemical skeletons of natural and semisynthetic flavonoids.

During the past decade, vast chemical libraries have been tested for biological activity in high-throughput screening (HTS) assays.[4] A series of papers, concerning the reliability of HTS outcomes, highlighted sources of false positives related to physicochemical properties of small molecules.[5,6] Nonspecific bioactivity due to colloidal aggregation of small molecules has been reported to be a major source of promiscuity in biochemical assays.[6] The underlying mechanism of aggregate-based inhibition proceeds via partial protein unfolding, when bound to an aggregate particle,[7] or reversible sequestration of the enzyme,[8] resulting in apparent inhibition. High biological reactivity (usually due to the presence of highly reactive groups),[9] autofluorescence and luciferase inhibition[6] represent other causes of promiscuity and readout artifacts present in HTS experiments. The largest, freely accessible, depository of HTS results is PubChem BioAssays (http://www.ncbi.nlm.nih.gov/pcassay), containing thousands of bioassay depositions and tenths of millions of biological activity outcomes.[10] During recent years, the number of depositions raised significantly offering an attractive source of activity data for retrospective studies.

Flavonoids are generally known to be highly bioactive and show a series of benefic effects on human health.[11,12] However, this class of polyphenols exhibits vicious physicochemical and biological properties that give rise to numerous false positives in HTS.[8,13−15] As one of the most representative flavonoid,

**Received:** May 4, 2014
**Published:** July 15, 2014

   dx.doi.org/10.1021/ci5002668 | J. Chem. Inf. Model. 2014, 54, 2360−2370

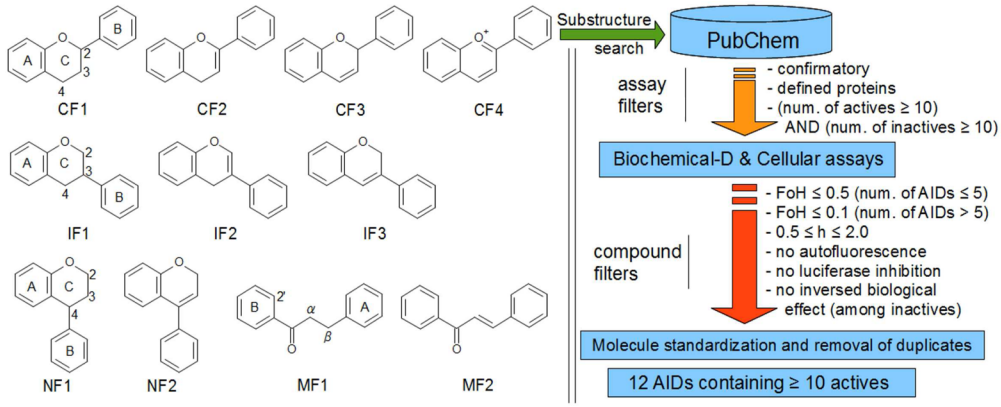



**Figure 1.** (left side) Common skeletons of natural and semisynthetic flavonoids used in this work. (right side) Workflow used to search in PubChem (http://pubchem.ncbi.nlm.nih.gov/) and to assemble the ColBioS-FlavRC data sets. Notations: CF common flavonoids; IF isoflavonoids; NF neoflavonoids; MF minor flavonoids; Biochemical-D assays biochemical assays containing detergents; FoH frequency of hits; AID PubChem Bioassay identification number; h Hill slope value.

quercetin has been repeatedly reported as an example of promiscuous compound.[6,8,13,16]

In 2006, Kinoshita et al.[17] assembled a collection of bioactive flavonoid derivatives, searching for molecules sharing skeleton CF1 and IF1 (see Figure 1) in the literature and several freely and commercially available databases. The integrated database resulted in a number of 7752 compounds as reported by the authors. Only a relatively small number of members showed biological activities against individual targets (e.g., at most 16 compounds active against lipooxigenase).[17]

Such efforts can support the search for other valuable flavonoids in large chemical libraries by means of high-throughput virtual screening (VS) methods. Pharmacophore modeling has been applied successfully and extensively in VS, de novo design and lead optimization, and is considered an important tool in drug discovery.[18] A pharmacophore (or pharmacophoric pattern) is defined as "the ensemble of steric and electronic features that is necessary to ensure the optimal supramolecular interactions with a specific biological target structure and to trigger (or to block) its biological response".[19] In ligand-based 3D pharmacophore modeling, a search algorithm maps pharmacophores common to a number of known ligands. The pharmacophore models (hypotheses) can be further used to screen chemical libraries. Many successful pharmacophore approaches have been developed in the past decade, e.g., PHASE, MOE, Catalyst etc., only to name a few. The efficiency of pharmacophore models can be assessed in VS scenarios, using sets of molecules with experimentally determined biological activities classified as actives and inactives (or decoys) against the protein target of interest. Furthermore, the choice of unbiased evaluation metrics to measure the early recovery of actives (true positives) is essential in comparative VS studies.[20−22]

In this study, we explored the PubChem Bioassay library focusing on bioactive molecules sharing flavonoidic chemical skeletons. We will refer to such compounds as to flavonoids and related compounds (FlavRC). We sought to filter out false positives and to identify biologically selective molecules. Thereby, we established the Collection of BioSelective Flavonoids and Related Compounds (ColBioS-FlavRC), comprising sets of active and inactive compounds (further referred to as actives and inactives) against 12 protein targets. In order to illustrate the practical use of the data assembled herein, we performed extensive ligand-based pharmacophore modeling. We tested the early enrichment performance in VS conditions and describe the results using newly proposed, unbiased, evaluation metrics.

## MATERIALS AND METHODS

**Substructure Search.** We selected 11 chemical scaffolds (see Figure 1) representative for natural flavonoids, e.g., flavanones (derivatives of 2-phenyl-2,3-dihydrochromen-4-one), flavones (derivatives of 2-phenylchromen-4-one), 3-flavenes (derivatives of 2-phenyl-2*H*-chromene), anthocyans (derivatives of 2-phenylchromenylium), isoflavanones (derivatives of 3-phenyl-2,3-dihydrochromen-4-one), isoflavones (derivatives of 3-phenylchromen-4-one), coumarins (3-phenylchromen-2-one), 4-arylcoumarins (derivatives of 4-phenyl-3,4-dihydrochromen-2-one and 4-phenylchromen-2-one), derivatives of chalcones, dihydrochalcones and aurones, as well as other intermediates or synthetic compounds. These structures were used for substructure search in PubChem Compound database (PubChem Substructure Search module: http://pubchem.ncbi.nlm.nih.gov/search/#; accessed on July 11, 2011).

**Assay Filters.** We retained only confirmatory assays with defined protein targets (as reported in PubChem BioAssay) and filtered out assays with a relatively small amount of bioactivity data, i.e., AIDs containing less than 10 actives and less than 10 inactives (compounds declared inactive in the assays).

Target binding due to aggregation of small molecules, in certain assay conditions, has been reported as the major source of nonspecific biological activity.[5,8,13] The inclusion of detergents (e.g., 0.01−0.1% Triton X-100) in the assay medium (if the stability of the enzyme remains unaffected) has been demonstrated to significantly reduce high rates of false positives (up to 90−95%, depending on the assay).[5,6,8,23,24] Thus, we retained only biochemical assays containing detergents and cellular assays.

**Compound Filters.** In order to remove compounds that show up as active in multiple assays regardless of the underlying mechanism (highly reactive groups, colloidal aggregation or interference with signal detection system) we first applied the frequency of hits filter, which we detail in the next subsection.


dx.doi.org/10.1021/ci5002668 | J. Chem. Inf. Model. 2014, 54, 2360−2370

Molecules acting as competitive single-site inhibitors tend to exhibit dose−response curves with Hill slope[25] close to 1. Increased Hill slope values were associated with aggregation-based or, less frequently, potent covalent inhibition.[24] In order to remove harbingers of unspecific activity, we retained compounds associated with Hill slope values within the range of 0.5−2, in accordance with previous studies and recommendations.[24,26]

Chromophoric and fluorophoric properties of small molecules (and impurities) plague HTS assay readouts leading to false positives and false negatives.[6,27] Recently, the outcomes of a series of counter screens were deposited in PubChem BioAssay (AIDs 587−594), profiling spectral properties of a large number of compounds across spectral regions commonly utilized in HTS.[27] These results were complemented by a supplementary list of 507 compounds categorized as detergent-insensitive autofluorescent false positives in the coumarin light detection region.[6] FlavRCs tested in assays in which the detection signal was measured in lower light detection region (Em/Ex: 340/450) were removed according to AID 589, AID 590, and the list of 507 compounds. Fluorescent compounds in AID 587, AID 588, and AID 591−594 were used to identify and remove FlavRCs from assays where the activity was measured at higher wavelengths.

Luciferases from organisms that yield bright bioluminescence have been adapted for use as reporters in HTS. Thus, many cellular assays rely on the ability of luciferase to produce light in the presence of luciferine and ATP (adenosine 5′-triphosphate).[15] Small molecules are able to interfere with this reaction generating false positives in HTS experiments.[15,28] The results of several counter screenings for the identification of luciferase inhibitors were deposited in PubChem Bioassays (AID 411, AID 1379, and AID 773). We used these positive outcomes to identify and remove readout artifacts in the luciferase-based assays evaluated herein.

Compounds denoted as "inconclusive" or "unspecified" were omitted as well as compounds denominate "inactive" but reported with some activity (probably due to an opposite effect compared to the one sought, i.e., in assays searching for inhibitors such a compound would activate the enzyme-target).

**Frequency of Hits (FoH).** Compounds that show up as hits in many different biological assays, covering a wide range of targets, are referred to as frequent hitters (FHs).[29] The frequency of hits (FoH) parameter is calculated as the ratio of the number of assays in which molecule $m$ was found active and the total number of assays in which molecule $m$ was tested. Considering that compounds may act likewise on closely related targets, Rohrer and Baumann[26] applied a weighting function to split the contribution of highly similar proteins. In eq 1 the FoH is computed considering also protein diversity.

$$\mathrm{FoH} = \frac{1}{T}\sum_{(i=1)}^{(N)}\frac{1}{n_i}\sum_{(j=1)}^{(n_i)} w_{ij} \qquad (1)$$

where $T$ is the total number of assays in which molecule $m$ was tested; $N$ is the total number of assay clusters, each containing similar targets; $n_i$ is the size of the $i$th assay cluster; $w_{ij}$ is 1 if molecule $m$ was reported active in assay $j$, part of cluster $i$, otherwise 0. A molecule active against a protein target is weighted in inverse proportion to the size of the cluster containing the target. Assays containing similar targets (according to a target sequence identity score higher than 0.8, as computed by the Pubchem BioAssay Structure−Activity module) were clustered together. The FoH filter was employed to remove compounds showing low biological selectivity and possible artifacts from HTS results (see Figure 1).

**Molecular Descriptors and Bioavailability Filters.** The number of hydrogen bond acceptors (HBA), hydrogen bond donors (HBD), rotatable bonds (RB), as well as molecular weight (MW), XLogP,[30] and 2D polar surface area[31] (PSA) were computed with FILTER (version 2.0.2, www.eyesopen.com).[32] The same program was employed for Lipinski rule of five,[33] Pharmacopia,[34] Veber[35] and aggregators[8,14] filtering, and also for solubility[32] calculations. The results were used to describe the distributions of the actives and inactives, comprised in ColBioS-FlavRC, and to assess properties as bioavailability and aggregation.

**Similarity Search.** Two-dimensional similarity search was performed on molecular 166-bit string MACCS fingerprints[36] computed with PaDEL software (version 2.7).[37] The Tanimoto coefficient was calculated using our in-house software SSTICIv1.[22] We employed similarity search for the selection of the inactive compounds included in the pharmacophore elucidation set and, further, to measure the mean similarity between first, the actives in the test set (for pharmacophore validation) against the inactives of the same set and, second, against the entire training set.

**Data Set Preparation for Pharmacophore Modeling and VS Validation.** Three-dimensional pharmacophore modeling relies on the spatial arrangement of atoms and fragments for known biologically active ligands. Many optically active compounds deposited in PubChem show poorly specified chiral characteristics and are not appropriate for three-dimensional modeling. Thus, before proceeding to ligand-based pharmacophore elucidation, we removed, from the ColBioS-FlavRC sets, compounds with unspecified stereochemistry. Further, we split the remaining set of actives into a pharmacophore elucidation (training) set and a pharmacophore validation (test) set, by including every third compound, in order of increasing activity, in the validation set. Using the actives in the training set as reference molecules, we extracted an equally sized set of inactives showing the highest Tanimoto similarity. The rest of the inactive molecules, together with the remaining 1/3 actives, form the test set (Figure 2). Thereby, the number of actives in the validation pool of compounds is (much) smaller compared to the number of inactives, a typical situation encountered in virtual screening scenarios.

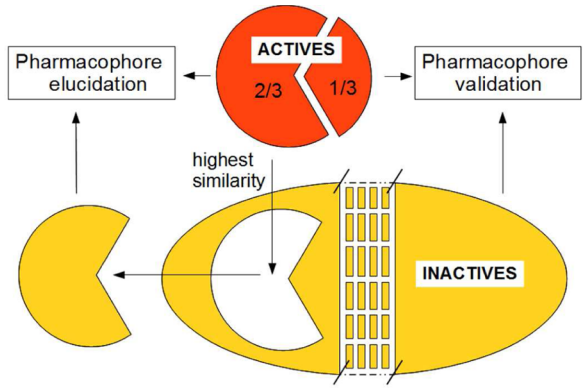



**Figure 2.** Preparation of data sets for pharmacophore elucidation and validation.


dx.doi.org/10.1021/ci5002668 | J. Chem. Inf. Model. 2014, 54, 2360−2370

**Settings for Pharmacophore Hypothesis Elucidation.** Three-dimensional pharmacophore mapping relies on the conformational analysis of the ligands. According to several studies, OMEGA (version 2.4.6, www.eyesopen.com)[38−40] is effectively able to reproduce the bioactive conformation of the ligands and recommended by Perola and Charifson[41] for conformational analysis on large chemical databases. Thus, conformers generated with OMEGA, using default parameters, (i.e., ewindow 10 kcal/mol, rms 0.8), were passed to PHASE[42−44] for 3D pharmacophore mapping. Given a set of ligands, PHASE identifies (by default) six types of pharmacophore sites: hydrogen bond acceptor, hydrogen bond donor, aromatic ring, hydrophobic, negative ionizable and positive ionizable. The site definitions of the first three include also a vector attribute to each idealized binding axes. Next, the algorithm finds *k*-point pharmacophores spatially shared by a user-defined minimum number of actives (and other tolerance criteria), by means of a tree-based partitioning technique.[43] Each alignment is measured in terms of the root-mean-square deviation (RMSD) in the site point positions ($S_{site}$), the average cosine of the angles formed by corresponding pairs of vector features (acceptors, donors, and aromatic rings) in the aligned structures ($S_{vector}$), and the overlap of van der Waals models of the non-hydrogen atoms in each pair of structures ($S_{volume}$). The so-called *active survival score*, attributed to each pharmacophore hypothesis (pharmacophore model), consists of the weighted sum of the previously enumerated terms in addition to a selectivity score (an empirical estimate of the rarity of a hypothesis; $S_{sel}$), a reference ligand energy score (a penalty for high-energy structures; $\Delta E$) and a reference ligand activity score (the biological activity; $A$). Another term regards the number of actives that match the hypothesis ($w^{M-1}$).[43]

The active survival score ranks pharmacophore hypotheses and, by default, attributes unitary weights to $S_{site}$, $S_{vector}$, $S_{volume}$ ($S_{sel}$, $\Delta E$, and $A$ are neglected). In this study, we employed pharmacophore search only for the purpose of VS and have not considered the activity values of the training molecules for pharmacophore hypotheses scoring. Because we searched for pharmacophores that would match at least 2/3 of the actives in the training set, we modified the $w^{M-1}$ term from 1 (default) to 1.1, in order to favor pharmacophore hypothesis matched by a larger number of ligands (as required in VS). The resulted pharmacophore hypotheses were clustered according to the complete linkage algorithm incorporated in PHASE. From each cluster, we kept only the hypothesis showing the highest active survival score, so that more diverse hypotheses would be evaluated. Using the equally sized set of inactives (in the pharmacophore elucidation set) we computed the so-called *adjusted survival score*, i.e., a multiple of the survival score of the inactives subtracted from the active survival score.[43]

**Evaluation Metrics.** Some of the commonly used parameters to evaluate VS methods have only limited applicability.[22,45] The exponential receiver operating curve enrichment (eROCE) is a robust metric, unbiased by the size of the data set, and is indicated in comparative studies.[22] The eROCE parameter is defined as the mean of the exponential weights, $\varepsilon_i$ (see Table 1), attributed to every active $i$ in the list (as ranked by the evaluated method), function to the corresponding false positive rate. Hence, the true positives recovered at the beginning of the list receive higher weights (starting from 1) compared to the upcoming ones (tending to zero).[22] The steepness of the exponential is adjusted by means of the $\alpha$ value.[22] It is often the case that outliers occur or the distribution of the actives is skewed and the mean might fail to measure the central tendency of $\varepsilon_i$. Hence, the median of the exponential weights $\varepsilon_i$ (further denoted MeROCE) would be recommended (in normal distributions the two measures provide close results).

**Table 1. eROCE and MeROCE Parameter Calculation Formula**

| $\varepsilon_i$[a] | eROCE | MeROCE |
|---|---|---|
| $\exp(-\alpha \mathrm{FPR}_i)$ | $\mathrm{mean}(\varepsilon_i)$ | $\mathrm{median}(\varepsilon_i)$ |

[a]Exponential weightings of every $i$th active in the ranking list found at false positive rate ($\mathrm{FPR}_i$); in this study $\alpha = 20$.

In this study, significant differences, in terms of MeROCE, between the evaluated pharmacophore models were assessed using the two-sample Mann−Whitney U test. We tested the equality of the distributions of exponential weight ($\varepsilon_i$) in each two of our pharmacophore model results against right and left shift alternatives by means of package "coin" (version 1.0−21)[46,47] available in R (version 2.14.2).[48] Exact conditional distributions were approximated by 99999 Monte Carlo sampling.

Because eROCE and MeROCE values lack in probabilistic interpretation, we characterized the VS results also in terms of more suggestive metrics, i.e., the percentage of true positives (TP%) at 1% and 5% false positive rate (FPR)[45] and the area under the receiver operating curve[49] (AUC). The first metric characterizes the early enrichment capabilities of VS methods, and the latter is a well-established measure of the discriminative power along the entire ranking list.

## RESULTS AND DISCUSSION

**Assembling ColBioS-FlavRC.** The efficient identification of interesting compounds, showing reasonable biological selectivity and stoichiometric, site-specific interaction with the protein target (1:1), is of major interest for lead-development.[50] The workflow employed to assemble ColBioS-FlavRC is represented in the right side of Figure 1. The substructure search in PubChem Compounds database resulted in about 306 000 CIDs (compound ID, as denoted in PubChem Compound), from which almost 23 000 were comprised in approximately 16 000 AIDs (assay IDs, as denoted in PubChem BioAssay) representing 2700 protein-targets. The application of the assay filters resulted in the analysis of 83 assays, from which we retained 53 biochemical (containing detergents) and cellular assays. Counter screens and summary reports were omitted. The retained 53 assays contain 3683 CIDs.

Compound filters aiming to remove biologically unselective molecules and false positives succeeded the assay filters. In the attempt to efficiently filter out promiscuous active-declared compounds, we applied a slightly modified version of the "Assay Artifacts Filter" used by Rohrer and Baumann[26] (right side of Figure 1). We computed the FoH parameter, evaluated the concentration−response Hill slope values (as reported by the assay depositors) and removed compounds according to the assembled lists of known autofluorescence and luciferase inhibitors (see Compound Filters section in Materials and Methods).

The FoH parameter of a molecule is defined as the ratio of the number of assays in which it was declared active and the total number of assays in which it was tested.[26] The distribution of the actives according to the FoH values (Figure 3a) indicates that most of the molecules show FoH $\leq$ 0.1. In Figure 3b one


dx.doi.org/10.1021/ci5002668 | J. Chem. Inf. Model. 2014, 54, 2360−2370

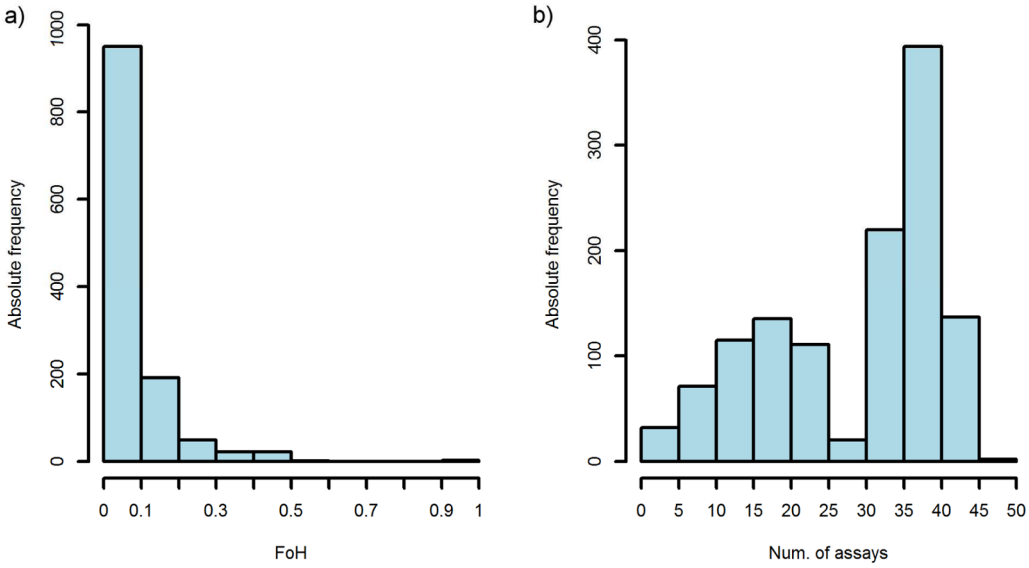



**Figure 3.** Histograms showing the number of (a) compounds (before applying compound filtering) according to the FoH values and (b) the number of assays in which they were tested.

**Table 2. Cellular and Biochemical-D Data Sets Comprised in ColBioSel-FlavRC**

| assay type | assay ID[a] | acivity range[b] | num. of actives | num. of inactives | protein target | biological effect | Uniprot ID[c] | num. of actives removed[d] | num. of inactives removed[e] |
|---|---|---|---|---|---|---|---|---|---|
| cellular | 1458 | 1.12–15.85 | 10 | 1566 | SMN2[f] | activation | Q16637 | 29/52/1 | 197 |
| | 1461 | 3.16–12.59 | 19 | 1297 | NPSR1[g] | inhibition | Q6W5P4 | 11/18/0 | |
| | 2551 | 0.22–19.95 | 58 | 1905 | ROR-$\gamma$[h] | inhibition | P51450 | 43/78/1 | |
| | 485297 | 0.5–5.01 | 43 | 1679 | Rab-9A[i] | activation | P51151 | 37/29/0 | 18 |
| | 504444 | 0.29–25.93 | 29 | 1874 | Nrf2[j] | inhibition | Q16236 | 23/36/0 | |
| | 504466 | 1.46–23.11 | 23 | 1557 | ATAD5[k] | inhibition | Q96QE3 | 28/20/1 | 17 |
| biochemical | 893 | 2.51–31.62 | 16 | 430 | 17$\beta$-HSD 10[l] | inhibition | Q99714 | 28/51/0 | |
| | 894 | 1.12–44.67 | 38 | 1127 | 15-PGDH[m] | inhibition | P15428 | 69/25/0 | |
| | 1030 | 0.11–31.62 | 111 | 989 | ALDH1A1[n] | inhibition | P00352 | 47/66/1 | 27 |
| | 2147 | 1–31.62 | 14 | 1347 | JMJD2E[o] | inhibition | B2RXH2 | 66/27/0 | 8 |
| | 504333 | 4.47–56.23 | 61 | 1875 | BAZ2B[p] | inhibition | Q9UIF8 | 89/56/0 | 59 |
| | 504339 | 8.91–63.1 | 73 | 1357 | JMJD2A[r] | inhibition | O75164 | 95/55/0 | 67 |

[a]Pubchem Bioassay identification. [b]Biological activity range measured as $IC_{50}$ ($\mu$M). [c]UniProtKB Accession Number (http://www.uniprot.org/).[52] [d]The number of actives removed by means of FoH/h/(autofluorecence and luciferase inhibitors). [e]The number of compounds declared inactive but showing activity concentrations (possible due to an inverse biological effect). [f]Survival of motor neuron protein 2, centromeric isoform d. [g]Neuropeptide S receptor isoform A. [h]Nuclear receptor ROR-gamma (*Mus musculus*). [i]Ras-related protein Rab-9A. [j]Nuclear factor erythroid 2-related factor 2. [k]ATPase family AAA domain-containing protein 5–DNA-replication. [l]17$\beta$-Hydroxysteroid Dehydrogenase Type 10. [m]15-Hydroxyprostaglandin dehydrogenase [$NAD^{(+)}$]. [n]Aldehyde dehydrogenase 1 family, member A1. [o]Jumonji domain containing 2E demethylase (Lysine-specific demethylase 4E). [p]Bromodomain adjacent to zinc finger domain 2B. [r]Lysine-specific demethylase 4A.

can observe that more than half of the actives were tested in more than 30 assays and only a small number in less than five. Of course, the higher the number of protein targets a molecule is tested against, the better the selectivity and/or promiscuity is reflected by FoH. For the purpose of this study, we retained molecules characterized by FoH $\leq$ 0.1, if the corresponding number of assays exceeded five. A number of 20 molecules were found active in one or two out of five assays (FoH $\leq$ 0.5). Considering this small number of compounds (compared to the entire set, i.e., 3683) and the insufficient bioactivity data to classify them as FHs, we supplied these compounds to the next filters (the Hill slope and detection interference).

Molecules which passed the compound filters were standardized (removal of salts, ionization at pH =7.4) and the eventual duplicates were eliminated. Twelve assays, each containing $\geq$10 actives, were retained. A number of 433 unique molecules, active in at least one of the 12 assays sets establish the ColBioS-FlavRC. Moreover, a number of 2908 inactives were retained and included in ColBioS-FlavRC to provide additional information and functionality to the collection.


dx.doi.org/10.1021/ci5002668 | J. Chem. Inf. Model. 2014, 54, 2360–2370

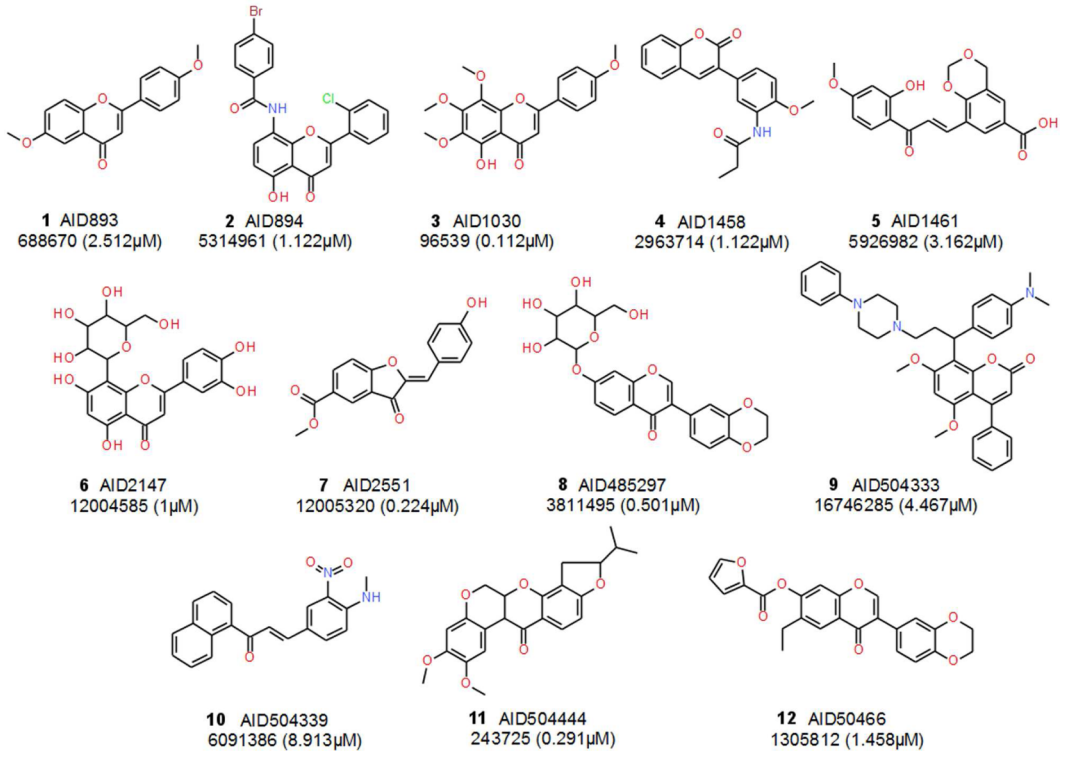



**Figure 4.** Chemical structures (compound number, assay ID, compound ID, and activity) of the most active compounds in the 12 ColBioS-FlavRC assays.

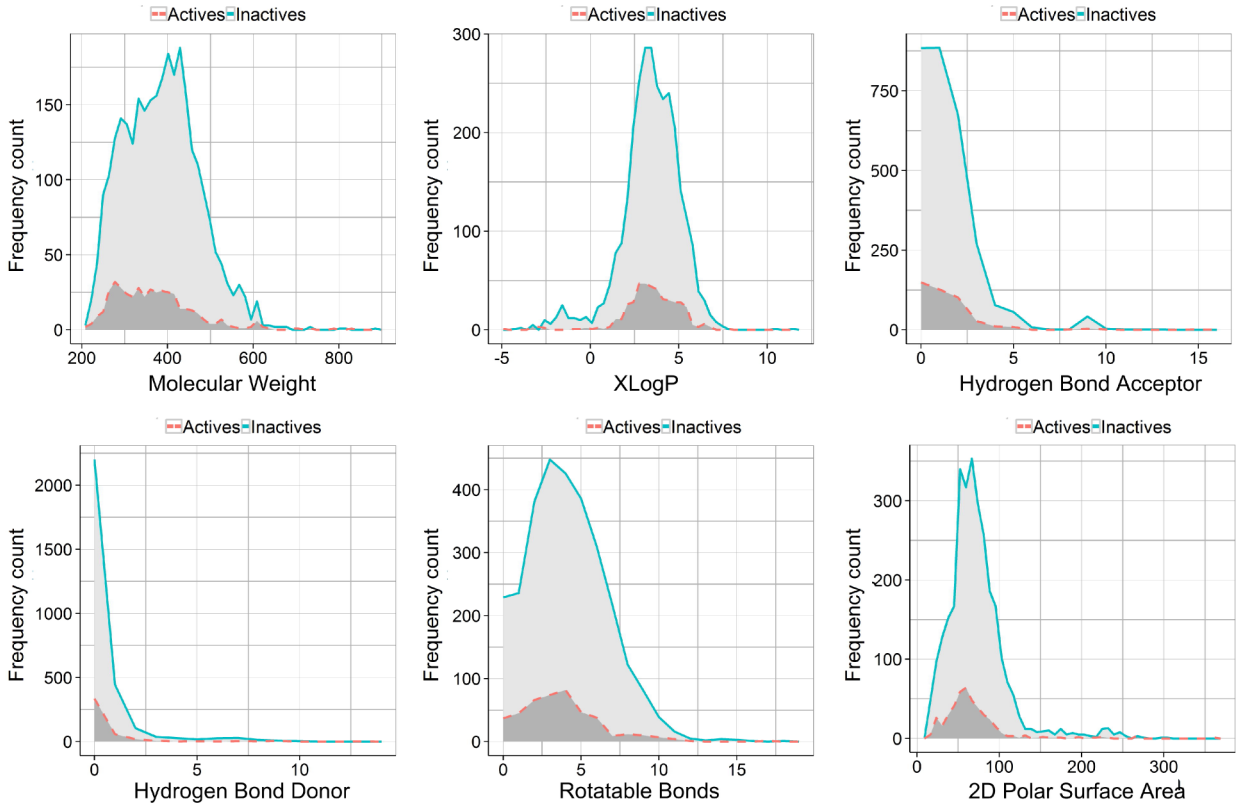



**Figure 5.** Histograms showing the distributions of the active (red colored dashed line) and the inactive (blue colored continuous line) compounds comprised in ColBioS-FlavRC, as described by six basic physical descriptors: molecular weight, XLogP, 2D polar surface area (computed on 50 bins), hydrogen bond acceptor, hydrogen bond donor, and rotatable bonds descriptors (bin width set to one). The maximum area described by both distributions is represented in "light gray" and the overlapping area in "dark gray".


dx.doi.org/10.1021/ci5002668 | J. Chem. Inf. Model. 2014, 54, 2360−2370

Recent studies argue not to ignore so-called "undesirable" functional groups (e.g., electrophilic reactive groups, particularly aliphatic esters etc.) in chemical screening libraries due to a large prevalence shared by commercially available drugs.[51] In our approach, we have not considered a filter focusing specifically on removing highly reactive groups. The FoH filter removes promiscuous compounds overseeing the underlying mechanism.

The compound filters (Figure 1) removed 1082 compounds from the final 12 sets. The FoH and the Hill slope filters were the most effective, in contrast to the few autofluorescent compounds or luciferase inhibitors left (Table 2). In the case of seven assays, both cellular and biochemical, we identified and removed inactives for which, in spite that, activity data were reported (Table 2).

**Brief Description of ColBioS-FlavRC.** The collection of FlavRC contains 12 sets of active and inactive compounds as described in Table 2. A number of 433 compounds are active against at least one protein target, from which 34.41% only in cellular assays, 60.97% only in biochemical assays and 4.62% in both types. From the 495 activity ($IC_{50}$) values, 5.86% are <1 $\mu$M and 14.34% are >30 $\mu$M. In Figure 4, we illustrate the chemical structures of the most potent member of each assay. Representatives from all four flavonoid classes (see Figure 1) are present, including two C- and O-glycosylflavonoids (i.e., compounds **6** and **8**). A number of 2908 molecules display no activity against neither of the protein targets.

Six molecular properties reflecting basic physicochemical properties were computed for the 433 actives and 2908 inactives in ColBioS-FlavRC. The histograms in Figure 5 describe the distributions of the actives against the inactives in terms of MW, XLogP, HBA, HBD, RB, and PSA. We found that approximately 90% of the actives obey simultaneously the following rules: MW < 550, 1 < XLogP < 6, HBA < 6, HBD < 4, RB < 11, PSA < 130. The shapes of the histograms described by the two classes are similar and display peaks at about the same property values. One can observe that the actives are entirely "buried" in the area described by the span of the inactives, suggesting that (i) the two classes share very similar physicochemical characteristics and (ii) the number of inactives exceeds the number of actives along the value ranges of the descriptors.

Recent studies have drawn attention toward the varying bioavailability of dietary flavonoid derivatives[53−56] and others have highlighted the promiscuity of these compounds in HTS studies.[6,8,13,15,57,58] Using physicochemical properties of drugs and drug candidates several filters have been developed to facilitate the identification of orally bioavailable molecules. We found that between 89.96% and 93.07% of the ColBioS-FlavRC molecules pass the Pharmacopia and Veber filters and about three-quarters obey Lipinski's rule of five with no violation. Interestingly, a higher percentage of inactives (69.60%) indicate superior solubility in water[32] compared to the actives (62.12%) (Table S1 in the Supporting Information).

All actives and 99.86% of the inactives (ColBioS-FlavRC compounds) pass the "aggregators filter", which has been set up according to the list of aggregators established by the Shoichet group.[8] Moreover, a percentage of 49.88% of the actives and 42.26% of the inactives pass the aggregator prediction model developed by the same group.[14] However, aggregation (and molecular solubility) is very condition-dependent (pH, salt, detergent, protein target)[6,50] and varies from one assay to another. The influence of assay conditions upon nonspecific bioactivity is not fully understood, but actives found in detergent-based biochemical and cellular confirmatory assays, showing reasonable bioselectivity (here measured as FoH) and steepness of the dose—response curve, increase the certainty of HTS outcomes.

The activities highlighted by our study extend the current knowledge of flavonoid derivatives toward high impact targets and biological selectivity. Only a small number of compounds (implying flavonoid derivatives) indexed in BindingDB[59] (http://www.bindingdb.org; accessed on March 4, 2013) were found to exert biological activity against the ColBioS-FlavRC protein targets. BindingDB collects activity data published almost exclusively via the scientific journals. We searched the database according to the Uniprot IDs shown in Table 2 (none of the PubChem assays present in ColBioS-FlavRC were indexed in BindingDB) and found bioactive molecules only for seven of the ColBioS-FlavRC protein targets (the number of active compounds is given in parentheses): NPSR1 (49), 17$\beta$-HSD10 (236), HPDG (39), ALDH1A1 (1), JMJD2E (36), BAZ2B (1), and JMJDA2 (31). Hence, ColBioS-FlavRC can provide valuable information for rationalized bioactivity determinations of other natural flavonoid derivatives.

**Validation of Pharmacophore Hypothesis for VS.** After removing optically active compounds with incomplete specified stereoisomerisms, we split the data sets into pharmacophore hypothesis elucidation and validation subsets for each target (see section Data Set Preparation for Pharmacophore Modeling and VS Validation in Materials and Methods). Due to a small number of remaining actives, four assay sets were left out, i.e., AID 893, AID 1458, AID 1461, and AID 2147. The remaining data sets (Table 3) were employed for ligand-based pharmacophore hypothesis generation and subsequent database search. These sets qualify for external validation as revealed by the low Tanimoto molecular similarity values, i.e., between 0.43 and 0.5, measured between the actives and the inactives in the validation set. Analogous Tanimoto values were attained when the same set of actives was compared to the actives in the training set (see Table S2 in the Supporting Information). The reasonable structural diversity, the physicochemical similarity (see the distributions in Figure 5) and the small number of actives compared to the inactives (see Table 3), qualify ColBioS-FlavRC sets for method optimization and validation. These criteria constitute the basis for validation sets[60] and well-known benchmarking sets for VS approaches.[61]

**Table 3. ColBioS-FlavRC Data Sets Used in Pharmacophore Modeling and Validation**

| | elucidation | | validation | | |
|---|---|---|---|---|---|
| assay ID[a] | num. of actives | num. of inactives[b] | num. of actives | num. of inactives | active to inactive ratio |
| 894 | 20 | 20 | 9 | 628 | 1:70 |
| 1030 | 38 | 38 | 18 | 588 | 1:33 |
| 2551 | 34 | 34 | 16 | 1120 | 1:70 |
| 485297 | 28 | 28 | 13 | 942 | 1:72 |
| 504333 | 16 | 16 | 8 | 1166 | 1:146 |
| 504339 | 25 | 25 | 12 | 866 | 1:72 |
| 504444 | 16 | 16 | 7 | 1031 | 1:147 |
| 504466 | 14 | 14 | 7 | 840 | 1:120 |

[a]Pubchem Bioassay identification. [b]The number of inactives equals the number of actives in the training set (see Figure 2).


dx.doi.org/10.1021/ci5002668 | J. Chem. Inf. Model. 2014, 54, 2360−2370

**Table 4. Early Enrichment Performance of Three- and Four-Point Pharmacophore Hypotheses**

| | three-point pharmacophore hypotheses | | | | four-point pharmacophore hypotheses | | | |
|---|---|---|---|---|---|---|---|---|
| assay ID[a] | Php hypo[b] | MeROCE[c] | TP% at 1% FPR | TP% at 5% FPR | Php hypo[d] | MeROCE | TP% at 1% FPR | TP% at 5% FPR |
| 894 | ARR.12 | 0.88 | 66.67 | 100 | AARR.14 | 0.73 | 33.33 | 100 |
| 1030 | ARR.62 | 0 | 0 | 0 | na[e] | na | na | na |
| 2551 | ARR.43 | 0.70 | 31.25 | 68.75 | AARR.27 | 0.62 | 25 | 81.25 |
| 485297 | ARR.28 | 0.38 | 23.08 | 61.54 | AARR.87 | 0.34 | 0 | 46.15 |
| 504333 | ARR.39 | 0.91 | 62.5 | 62.5 | AHRR.13 | 0.33 | 25 | 50 |
| 504339 | ARR.18 | 0.85 | 58.33 | 75 | AHRR.50 | 0.47 | 25 | 50 |
| 504444 | ARR.46 | 0.43 | 14.29 | 57.14 | AARR.142 | 0.2 | 0 | 42.86 |
| 504466 | ARR.19 | 0.83 | 42.86 | 71.43 | AARR.34 | 0.59 | 28.57 | 100 |

[a]Pubchem Bioassay identification. [b]Pharmacophore hypothesis (A hydrogen bond acceptor, H hydrophobic, R aromatic ring, and ID number attributed by the pharmacophore hypothesis elucidation algorithm). [c]Exponential receiver operation curve enrichment, $\alpha$ = 20.0. [d]Pharmacophore hypothesis. [e]Not available (see text).

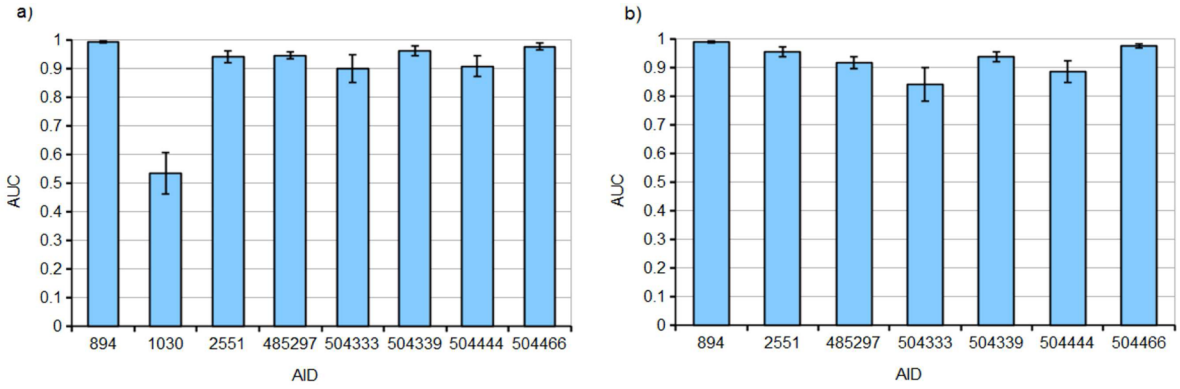



**Figure 6.** Area under the receiver operating curve (AUC) values (and standard deviation) achieved by the three- (a) and four-point (b) pharmacophore models reported in Table 4.

From each data set, we evaluated the first five hypotheses showing the highest adjusted survival scores (see section Settings for Pharmacophore Hypothesis Elucidation in Materials and Methods), giving rise to a number of 59 pharmacophore hypotheses assessed for VS capabilities (using the validation sets described in Table 3). We found three- and four-point pharmacophore hypotheses for every target set except AID 1030 (see Table 4). The molecules were ranked according to the *fitness score* (i.e., $S_{site} + S_{vector} + S_{volume}$) and the early recovery of actives was measured by the median exponential ROC enrichment ($\alpha$ = 20.0) parameter (see Materials and Methods section Evaluation Metrics).

The one-sided Mann−Whitney U test was applied to determine significant differences ($p < 0.05$) in MeROCE between pharmacophore searches. The results are graphically displayed in Figure S1 in the Supporting Information. In many cases pharmacophore models did not perform significantly different from each another.

In Table 4, we report the results obtained by the three- and four-point pharmacophores achieving the highest MeROCE per target set. MeROCE and TP% at 1% and 5% FPR describe the early recovery of actives and the discriminative performances throughout the entire ranked list is measured by AUC (Figure 6). In Table S3, in the Supporting Information, additional evaluation results using other metrics commonly employed in VS are available for all pharmacophore searches performed in this study.

All three-point pharmacophore hypotheses consist of two aromatic rings and one hydrogen bonding acceptor. Most of the four-point hypotheses exert an additional H-bond acceptor group, apart from assays 504339 and 504333, which require a hydrophobic site (the pharmacophore hypotheses are described in Figure S2 to S9 in the Supporting Information). These inherent simple models account for essential features common to FlavRCs as reflected by the generally high enrichment values. The most successful models were found in the case of AID 894, AID 504333, and AID 504339, where more than 50% of the actives were retrieved among the top 1% inactives. Although, three-point pharmacophores are more efficient, compared to four-point pharmacophores, in the cases of AID 2551 and AID 504466 (Table 4), four-point pharmacophores recovered more actives at 5% FPR. The high AUC values (generally higher than 0.9—see Figure 6) support the good performances achieved by the pharmacophore models. In the case of AID 1030, in spite of the many submicromolar activities available, we were not able to find an efficient pharmacophore hypothesis for VS (the best model resulted in a MeROCE score of 0.004 and an AUC score of 0.53 ± 0.07, suggesting no significant enrichment).

The development of prediction models to discriminate between actives and inactives might depend upon the difference in activity range between the two classes. For such purposes, a common practice is to establish activity thresholds to define class membership (e.g., $IC_{50} \leq 100$ nM, $IC_{50} \leq 1$ $\mu$M or $\leq$10 $\mu$M are commonly used to define classes of actives).[60,62] In assembling the ColBioS-FlavRC we adopted class labels as provided by the depositors of the assays, and further for pharmacophore modeling we have not considered any activity cutoffs. In spite of micromolar range covered by most $IC_{50}$ values (80% of the actives exhibit $IC_{50}$ values between 1 and 30 $\mu$M) in the ColBioS-FlavRC sets, we were able to obtain simple


dx.doi.org/10.1021/ci5002668 | J. Chem. Inf. Model. 2014, 54, 2360−2370

pharmacophore models capable to discriminate effectively between the two classes. However, one should be aware of the various active/inactive thresholds throughout the ColBioS-FlavRC sets if the use of unanimous activity cutoffs for class labeling is intended.

The pharmacophore models proposed in this study can be further used to predict the activity of chemical libraries enriched in flavonoids (and related compounds) and to guide the identification of other valuable derivatives for the targets assessed herein (see Figure S2 to S9 in the Supporting Information). Moreover, for most of the ColBioS-FlavRC biological targets, crystallographic structures are available, which can facilitate the use of structure-based methods in future studies. The complete ColBioS-FlavRC data sets are available in the Supporting Information (ci5002668_si_002.xls).

**MeROCE.** We attempted to determine which of the two parameters are more suited for the evaluation of the $\varepsilon_i$ distributions resulted from the 59 pharmacophore searches: MeROCE, as the median $\varepsilon_i$, or eROCE, as the mean $\varepsilon_i$. We proceeded to test the 59 distributions for normality by applying the Shapiro-Wilk test.[63] A percentage of 66.1% of the distributions rejected the null hypothesis ($p < 0.05$), i.e., that the distribution would be normal. Thus, for the majority of the situations encountered herein, the mean is not a valid measure of centrality, and the use of the median would be a better choice. In order to highlight significant differences between the performances of the pharmacophore models, we applied the two-sample nonparametric Mann−Whitney U test. For the remaining 33.9%, i.e., 20 models, showing valid normal distributions, we applied the $t$ test. For these cases we sought differences between the results obtained by the two significance tests (applied one-sided; $p < 0.05$). We found only two cases where the sample tests disagreed. Thus, given the results of this last endeavor, we were encouraged to employ MeROCE.

## CONCLUSIONS

This study represents an effort to identify flavonoids with reasonable biological selectivity in high-throughput screening assays. We compiled a collection of 433 valuable, biologically selective, flavonoids and related compounds, by exploring HTS results, which are known to be plagued by promiscuous small molecules. ColBioS-FlavRC is grounded on confirmatory cellular as well as biochemical assays. Regarding the latter type, we selected assays containing a small amount of detergent (to remove small-molecule aggregation) and followed major guidelines to obtain compounds showing selectivity and stoichiometric protein−ligand interactions.

The structural diversity, the high physicochemical similarities and the reasonable high active to inactive ratios, which characterizes the class of ColBioS-FlavRC actives and inactives, recommend these sets for VS method optimization and validation. Accordingly, we found simple, but efficient, three- and four- point pharmacophore models with high early enrichment capacities. These performances were described by means of MeROCE, which we show to be (for the pharmacophore models evaluated herein) a better choice compared to eROCE.

The physicochemical profile of ColBioS-FlavRC actives suggest reasonable bioavailability, which is complemented by increased biological selectivity. These essential requirements for promising drug-candidate molecules recommend the 433 actives for further experimental and theoretical investigations. We limited this study to a brief description of the data sets and to retrospective pharmacophore-based virtual screenings providing ColBioS-FlavRC in the Supporting Information for future explorations. Additionally, in future work we will attempt to expand the application of the workflow described herein to the entire library of compounds tested in PubChem Bioassay high-throughput screenings. If successful, such efforts would provide valuable data for the identification of bioselective compounds.

## ASSOCIATED CONTENT

### Supporting Information

The ColBioS-FlavRC data sets (xls file containing molecules in SMILES format and the corresponding activity data), statistics regarding bioavailability and solubility of ColBioS-FlavRC compounds, properties of pharmacophore elucidation and evaluation sets, detailed pharmacophore evaluation, graphical representations of most successful three- and four-point pharmacophore models, and heatmaps reflecting statistically significant differences between the performances of the pharmacophore hypotheses evaluated. This material is available free of charge via the Internet at http://pubs.acs.org.

## AUTHOR INFORMATION

### Corresponding Author

*E-mail: kurunczi@umft.ro. Phone: +40-724-263-818.

### Notes

The authors declare no competing financial interest.

## ACKNOWLEDGMENTS

This work was supported by Project 1.2 of the Institute of Chemistry Timisoara of the Romanian Academy. We thank OpenEye for providing us an academic license for the OpenEye software package.

## ABBREVIATIONS

AID, PubChem Bioassay identification number; CID, PubChem Compound identification number; eROCE, exponential retriever operating curve enrichment; FH, frequent hitters; FlavRC, flavonoids and related compounds; FoH, frequency of hits; h, Hill slope value; MeROCE, median exponential retriever operating curve enrichment

## REFERENCES

(1) Andersen, A. M.; Markham, K. R. *Flavonoids: chemistry, biochemistry, and applications*; CRC Press Taylor & Francis: Boca Raton, 2006.

(2) Ververidis, F.; Trantas, E.; Douglas, C.; Vollmer, G.; Kretzschmar, G.; Panopoulos, N. Biotechnology of flavonoids and other phenylpropanoid derived natural products. Part I: Chemical diversity, impacts on plant biology and human health. *Biotechnol. J.* **2007**, *2*, 1214−1234.

(3) Samanta, A.; Das, G.; Das, S. K. Roles of flavonoids in plants. *Int. J. Pharm. Sci. Nanotechnol.* **2011**, *6*, 12−13.

(4) Inglese, J.; Johnson, R. L.; Simeonov, A.; Xia, M.; Zheng, W.; Austin, C. P.; Auld, D. S. High-throughput screening assays for the identification of chemical probes. *Nat. Chem. Biol.* **2007**, *3*, 466−479.

(5) Thorne, N.; Auld, D. S.; Inglese, J. Apparent activity in high-throughput screening: origins of compound-dependent assay interference. *Curr. Opin. Chem. Biol.* **2010**, *14*, 315−324.

(6) Jadhav, A.; Ferreira, R. S.; Klumpp, C.; Mott, B. T.; P, C.; Inglese, J.; Thomas, C. J.; Maloney, D. J.; Shoichet, B. K.; Simeonov, A. Quantitative Analyses of Aggregation, Autofluorescence, and Reactivity Artifacts in a Screen for Inhibitors of a Thiol Protease. *J. Med. Chem.* **2010**, *53*, 37−51.

(7) Coan, K. E. D.; Maltby, D. a.; Burlingame, A. L.; Shoichet, B. K. Promiscuous aggregate-based inhibitors promote enzyme unfolding. *J. Med. Chem.* **2009**, *52*, 2067−2075.
(8) McGovern, S. L.; Helfand, B. T.; Feng, B.; Shoichet, B. K. A specific mechanism of nonspecific inhibition. *J. Med. Chem.* **2003**, *46*, 4265−4272.
(9) Rishton, G. M. Reactive compounds and in vitro false positives in HTS. *Drug Discovery Today* **1997**, *2*, 382−384.
(10) Wang, Y.; Xiao, J.; Suzek, T. O.; Zhang, J.; Wang, J.; Bryant, S. H. PubChem: a public information system for analyzing bioactivities of small molecules. *Nucleic Acids Res.* **2009**, *37*, W623−633.
(11) Tapas, A. R.; Sakarkar, D. M.; Kakde, R. B. Flavonoids as Nutraceuticals: A Review. *Pharm. Res.* **2008**, *7*, 1089−1099.
(12) Nijveldt, R. J.; van Nood, E.; van Hoorn, D. E.; Boelens, P. G.; van Norren, K.; van Leeuwen, P. a. Flavonoids: a review of probable mechanisms of action and potential applications. *Am. J. Clin. Nutr.* **2001**, *74*, 418−425.
(13) McGovern, S. L.; Caselli, E.; Grigorieff, N.; Shoichet, B. K. A common mechanism underlying promiscuous inhibitors from virtual and high-throughput screening. *J. Med. Chem.* **2002**, *45*, 1712−1722.
(14) Seidler, J.; McGovern, S. L.; Doman, T. N.; Shoichet, B. K. Identification and prediction of promiscuous aggregating inhibitors among known drugs. *J. Med. Chem.* **2003**, *46*, 4477−4486.
(15) Auld, D. S.; Southall, N. T.; Jadhav, A.; Johnson, R. L.; Diller, D. J.; Simeonov, A.; Austin, C. P.; Inglese, J. Characterization of chemical libraries for luciferase inhibitory activity. *J. Med. Chem.* **2008**, *51*, 2372−2386.
(16) King, O. N. F.; Li, X. S.; Sakurai, M.; Kawamura, A.; Rose, N. R.; Ng, S. S.; Quinn, A. M.; Rai, G.; Mott, B. T.; Beswick, P.; Klose, R. J.; Oppermann, U.; Jadhav, A.; Heightman, T. D.; Maloney, D. J.; Schofield, C. J.; Simeonov, A. Quantitative high-throughput screening identifies 8-hydroxyquinolines as cell-active histone demethylase inhibitors. *PLoS One* **2010**, *5*, e15535.
(17) Kinoshita, T.; Lepp, Z.; Kawai, Y.; Terao, J.; Chuman, H. An integrated database of flavonoids. *Biofactors* **2006**, *26*, 179−188.
(18) Yang, S.-Y. Pharmacophore modeling and applications in drug discovery: challenges and recent advances. *Drug Discovery Today* **2010**, *15*, 444−450.
(19) Wermuth, C. G.; Ganellin, C. R.; Lindberg, P.; Mitscher, L. A. Glossay of Terms used in Medicinal Chemistry. *Pure Appl. Chem.* **1998**, *70*, 1129−1143.
(20) Hawkins, P. C. D.; Warren, G. L.; Skillman, a. G.; Nicholls, A. How to do an evaluation: pitfalls and traps. *J. Comput.-Aided Mol. Des.* **2008**, *22*, 179−190.
(21) Nicholls, A. What do we know and when do we know it? *J. Comput.-Aided Mol. Des.* **2008**, *22*, 239−255.
(22) Avram, S. I.; Crisan, L.; Bora, A.; Pacureanu, L. M.; Avram, S.; Kurunczi, L. Retrospective group fusion similarity search based on eROCE evaluation metric. *Bioorg. Med. Chem.* **2013**, *21*, 1268−1278.
(23) Feng, B. Y.; Shoichet, B. K. A detergent-based assay for the detection of promiscuous inhibitors. *Nat. Protoc.* **2006**, *1*, 550−553.
(24) Feng, B. Y.; Simeonov, A.; Jadhav, A.; Babaoglu, K.; Inglese, J.; Shoichet, B. K.; Austin, C. P. A high-throughput screen for aggregation-based inhibition in a large compound library. *J. Med. Chem.* **2007**, *50*, 2385−2390.
(25) Hill, A. V. The possible effects of the aggregation of the molecules of haemoglobin on its dissociation curves. *J. Physiol.* **1910**, *40*, iv−vii.
(26) Rohrer, S. G.; Baumann, K. Maximum unbiased validation (MUV) data sets for virtual screening based on PubChem bioactivity data. *J. Chem. Inf. Model.* **2009**, *49*, 169−184.
(27) Simeonov, A.; Jadhav, A.; Thomas, C. J.; Wang, Y.; Huang, R.; Southall, N. T.; Shinn, P.; Smith, J.; Austin, C. P.; Auld, D. S.; Inglese, J. Fluorescence spectroscopic profiling of compound libraries. *J. Med. Chem.* **2008**, *51*, 2363−2371.
(28) Auld, D. S.; Zhang, Y.-Q.; Southall, N. T.; Rai, G.; Landsman, M.; MacLure, J.; Langevin, D.; Thomas, C. J.; Austin, C. P.; Inglese, J. A basis for reduced chemical library inhibition of firefly luciferase obtained from directed evolution. *J. Med. Chem.* **2009**, *52*, 1450−1458.
(29) Roche, O.; Schneider, P.; Zuegge, J.; Guba, W.; Kansy, M.; Alanine, A.; Bleicher, K.; Danel, F.; Gutknecht, E.-M.; Rogers-Evans, M.; Neidhart, W.; Stalder, H.; Dillon, M.; Sjögren, E.; Fotouhi, N.; Gillespie, P.; Goodnow, R.; Harris, W.; Jones, P.; Taniguchi, M.; Tsujii, S.; von der Saal, W.; Zimmermann, G.; Schneider, G. Development of a virtual screening method for identification of "frequent hitters" in compound libraries. *J. Med. Chem.* **2002**, *45*, 137−142.
(30) Wang, R.; Fu, Y.; Lai, L. A New Atom-Additive Method for Calculating Partition Coefficients. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 615−621.
(31) Ertl, P.; Rohde, B.; Selzer, P. Fast calculation of molecular polar surface area as a sum of fragment-based contributions and its application to the prediction of drug transport properties. *J. Med. Chem.* **2000**, *43*, 3714−3717.
(32) *FILTER*, version 2.0.2 (OMEGA 2.5.1.4); OpenEye Scientific Software: Santa Fe, NM, USA, 2009; www.eyesopen.com.
(33) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Delivery Rev.* **1997**, *46*, 3−25.
(34) Egan, W. J.; Merz, K. M.; Baldwin, J. J. Prediction of Drug Absorption Using Multivariate Statistics. *J. Med. Chem.* **2000**, *43*, 3867−3877.
(35) Veber, D. F.; Johnson, S. R.; Cheng, H.-Y.; Smith, B. R.; Ward, K. W.; Kopple, K. D. Molecular properties that influence the oral bioavailability of drug candidates. *J. Med. Chem.* **2002**, *45*, 2615−2623.
(36) Steinbeck, C.; Han, Y.; Kuhn, S.; Horlacher, O.; Luttmann, E.; Willighagen, E. The Chemistry Development Kit (CDK): an open-source Java library for Chemo- and Bioinformatics. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 493−500.
(37) Yap, C. W. PaDEL-descriptor: An open source software to calculate molecular descriptors and fingerprints. *J. Comput. Chem.* **2011**, *32*, 1466−1474.
(38) *OMEGA*, version 2.4.6; OpenEye Scientific Software: Santa Fe, NM, USA, 2012; www.eyesopen.com.
(39) Hawkins, P. C. D.; Skillman, A. G.; Warren, G. L.; Ellingson, B. A.; Stahl, M. T. Conformer Generation with OMEGA: Algorithm and Validation Using High Quality Structures from the Protein Databank and Cambridge Structural Database. *J. Chem. Inf. Model.* **2010**, *50*, 572−584.
(40) Hawkins, P. C. D.; Nicholls, A. Conformer Generation with OMEGA: Learning from the Data Set and the Analysis of Failures. *J. Chem. Inf. Model.* **2012**, *52*, 2919−2936.
(41) Perola, E.; Charifson, P. S. Conformational analysis of drug-like molecules bound to proteins: an extensive study of ligand reorganization upon binding. *J. Med. Chem.* **2004**, *47*, 2499−2510.
(42) *PHASE*, version 3.2; Schrödinger, LLC: New York, NY, USA, 2010.
(43) Dixon, S. L.; Smondyrev, A. M.; Knoll, E. H.; Rao, S. N.; Shaw, D. E.; Friesner, R. a. PHASE: a new engine for pharmacophore perception, 3D QSAR model development, and 3D database screening: 1. Methodology and preliminary results. *J. Comput.-Aided Mol. Des.* **2006**, *20*, 647−671.
(44) Dixon, S. L.; Smondyrev, A. M.; Rao, S. N. PHASE: a novel approach to pharmacophore modeling and 3D database searching. *Chem. Biol. Drug Des.* **2006**, *67*, 370−372.
(45) Jain, A. N. Bias, reporting, and sharing: computational evaluations of docking methods. *J. Comput.-Aided Mol. Des.* **2008**, *22*, 201−212.
(46) Hothorn, T.; Hornik, K.; van de Wiel, M. A.; Zeileis, A. A Lego System for Conditional Inference. *Am. Stat.* **2006**, *60*, 257−263.
(47) Hothorn, T.; Hornik, K.; van de Wiel, M. A.; Zeileis, A. Implementing a Class of Permutation Tests: The coin Package. *J. Stat. Softw.* **2008**, *28*, 1−23.
(48) R Development Core Team. R: A Language and Environment for Statistical Computing; R Foundation for Statistical Computing: Vienna, Austria, 2012.

(49) Hanley, J. A.; McNeil, B. J. The Meaning and Use of the Area under a Receiver Operating Characteristics (ROC) Curve. *Radiology* **1982**, *143*, 29−36.
(50) Coan, K. E.; Ottl, J.; Klumpp, M. Non-stoichiometric inhibition in biochemical high-throughput screening. *Expert Opin. Drug Discovery* **2011**, *6*, 405−417.
(51) Axerio-Cilies, P.; Castañeda, I. P.; Mirza, A.; Reynisson, J. Investigation of the incidence of ″undesirable″ molecular moieties for high-throughput screening compound libraries in marketed drug compounds. *Eur. J. Med. Chem.* **2009**, *44*, 1128−1134.
(52) The UniProt Consortium, Reorganizing the protein space at the Universal Protein Resource (UniProt). *Nucleic Acids Res.* **2012**, *40*, D71−D75.
(53) Shulman, M.; Cohen, M.; Soto-Gutierrez, A.; Yagi, H.; Wang, H.; Goldwasser, J.; Lee-Parsons, C. W.; Benny-Ratsaby, O.; Yarmush, M. L.; Nahmias, Y. Enhancement of Naringenin Bioavailability by Complexation with Hydroxypropoyl-$\beta$-Cyclodextrin. *PLoS One* **2011**, *6*, e18033.
(54) Reinboth, M.; Wolffram, S.; Abraham, G.; Ungemach, F. R.; Cermak, R. Oral bioavailability of quercetin from different quercetin glycosides in dogs. *Br. J. Nutr.* **2010**, *104*, 198−203.
(55) López-Lázaro, M. Distribution and biological activities of the flavonoid luteolin. *Mini-Rev. Med. Chem.* **2009**, *9*, 31−59.
(56) D'Archivio, M.; Filesi, C.; Varì, R.; Scazzocchio, B.; Masella, R. Bioavailability of the polyphenols: status and controversies. *Int. J. Mol. Sci.* **2010**, *11*, 1321−1342.
(57) Inglese, J.; Auld, D. S.; Jadhav, A.; Johnson, R. L.; Simeonov, A.; Yasgar, A.; Zheng, W.; Austin, C. P. Quantitative high-throughput screening: A titration-based approach that efficiently identifies biological activities in large chemical libraries. *Proc. Natl. Acad. Sci. U. S. A.* **2006**, *103*, 11473−11478.
(58) Sturm, N.; Desaphy, J.; Quinn, R. J.; Rognan, D.; Kellenberger, E. Structural insights into the molecular basis of the ligand promiscuity. *J. Chem. Inf. Model.* **2012**, *52*, 2410−2421.
(59) Liu, T.; Lin, Y.; Wen, X.; Jorissen, R. N.; Gilson, M. K. BindingDB: a web-accessible database of experimentally determined protein−ligand binding affinities. *Nucleic Acids Res.* **2007**, *35*, 198−201.
(60) Avram, S.; Pacureanu, L. M.; Seclaman, E.; Bora, A.; Kurunczi, L.; PLS-DA, - Docking Optimized Combined Energetic Terms (PLSDA-DOCET) Protocol: A Brief Evaluation. *J. Chem. Inf. Model.* **2011**, *51*, 3169−3179.
(61) Huang, N.; Shoichet, B. K.; Irwin, J. J. Benchmarking sets for molecular docking. *J. Med. Chem.* **2006**, *49*, 6789−6801.
(62) Schuffenhauer, A.; Brown, N.; Selzer, P.; Ertl, P.; Jacoby, E. Relationships between Molecular Complexity, Biological Activity, and Structural Diversity. *J. Chem. Inf. Model.* **2006**, *46*, 525.
(63) Royston, P. Remark AS R94: A remark on algorithm AS 181: The W-test for normality. *J. R. Stat. Soc. Ser. C (Appl. Stat.)* **1995**, *44*, 547−551.