Future Medicinal Chemistry

Mining large databases to find new leads with low similarity to known actives: application to find new DPP-IV inhibitors

Journal:	*Future Medicinal Chemistry*
Manuscript ID	FMC-2018-0597.R1
Manuscript Type:	Methodology
Keywords:	Dipeptidyl peptidase 4, Molecular fingerprints, Virtual screening

Mining large databases to find new leads with low similarity to known actives: application to find new DPP-IV inhibitors

Abstract

Aim: Fragment-based drug design or bioisosteric replacement is used to find new actives with low (or no) similarity to existing ones but requires the synthesis of non-existing compounds to prove their predicted bioactivity. Protein-ligand docking or pharmacophore screening are alternatives but they can become computationally expensive when applied to very large databases like ZINC. Therefore, fast strategies are necessary to find new leads in such databases.

Materials & methods: We designed a computational strategy to find lead molecules with very low (or no) similarity to existing actives and applied it to DPP-IV.

Results: The bioactivity assays confirm that this strategy finds new leads for DPP-IV inhibitors.

Conclusion: This computational strategy reduces the time of finding new lead molecules.

Keywords: CD26; Dipeptidyl peptidase 4; Diversifying molecular scaffolds; Expanding chemical space; Molecular fingerprints; Virtual molecular libraries; Virtual screening

1 . Introduction

Finding new leads is an essential step in projects to discover and develop new drugs [1–4]. There are two alternatives for achieving this goal: **(a)** experimentally testing compound libraries to find molecules that show the desired bioactivity against specific targets (a process known as *high throughput screening*; HTS) [5–8]; and **(b)** computationally predicting the bioactivity of interest in files containing molecular databases (known as *virtual screening*; VS) [9–12]. Obviously, the VS alternative is significantly cheaper than HTS, but the fact that, frequently, the former one relies on several sequential filters that use characteristics of known actives to find new leads (*e.g.*, constrained protein-ligand docking, ligand or structure-based pharmacophores and shape/electrostatics similarity) could be a major drawback because it could potentially lead to VS hits with high chemical similarity to the known actives and, therefore, in that cases, of limited interest.

There are a couple of computational strategies for finding new actives with low (or no) similarity to existing ones. The first consists of docking a fragment library at the target active site and then selecting the fragments with highest affinity and using linkers to join them (*i.e.*, *fragment-based drug design*) [13–15]. The second consists of replacing substructures involved in ligand-target interactions in a known active with other substructures that, although chemically different, can preserve equivalent intermolecular interactions with the target (*i.e.*, *bioisosteric replacement*) [16–18]. Unfortunately, both approaches involve the synthesis of non-existing compounds (with the corresponding difficulties associated to finding the proper synthetic plan and purifying the compound of interest) before it can be experimentally confirmed that the new compounds show the predicted bioactivity. Protein-ligand docking or pharmacophore screening are frequently used alternatives but they can become computationally very expensive if large databases like ZINC [19] need to be screened. Therefore, fast computational strategies are needed to find new lead molecules with very low (or no) similarity to existing actives in databases of purchasable compounds. This would have the advantage not only of eliminating the need to synthesize non-existing compounds before experimentally testing the bioactivity of the hits, but also, that if one of these hits has the desired activity, their synthesis and purification is described. Therefore, this can be of help for using this lead molecule for performing the corresponding structure-activity relationship studies by synthesizing new derivatives and finding which of them show improved bioactivity.

The goal of this paper is to design a computational strategy to find new lead molecules with very low (or no) similarity to existing actives in databases of purchasable compounds and to apply it to a target of pharmacological interest. The selected target is dipeptidyl peptidase IV (DPP-IV), whose inhibitors have been shown to be effective for the treatment of type 2 diabetes mellitus. Eleven DPP-IV inhibitors are now commercially available in different countries and there are many more in different stages of clinical development [20]. DPP-IV is a homodimeric transmembrane glycoprotein with a 22 residues long hydrophobic helix linking both subunits to the plasmatic membrane [21]. Each subunit has a large globular extracellular region formed by two different domains whose interface contains the corresponding catalytic center [22–25]. After the cleavage of the extracellular portion of DPP-IV from its transmembrane helix, DPP-IV is found in plasma and cerebrospinal fluid [25,26]. Although DPP-IV is secreted as a mature monomer, its

normal proteolytic activity is only possible after dimerization [27]. The DPP-IV binding site is highly druggable by small molecules with drug-like physicochemical properties [28,29].

The strategy used in the present work initially discarded those molecules in the database that show high similarity to known DPP-IV inhibitors by using an extremely fast fingerprint similarity search and applied a VS workflow to the remaining molecules (thus focusing computational resources only on those molecules that were of potential interest for finding new leads). The bioactivity assays performed with the VS hits obtained in this paper confirmed that this strategy was able to quickly find completely new lead molecules with basal activity as DPP-IV inhibitors. Moreover, we used molecular modeling to suggest how the most potent VS hit could be used as a lead molecule to find derivatives with significantly improved bioactivity.

2. Experimental

2.1. Set up of the starting databases for validating the virtual screening workflow and for lead discovery

The ability of our VS workflow to identify DPP-IV inhibitors was validated by applying it to an initial set of known 419 actives and 15,084 decoys (one active per 36 decoys [30]; see Figure 1). This set of known DPP-IV inhibitors was formed by molecules with a high activity value [*i.e.*, pX ≥ 7; where pX was calculated as the $-\log_{10}$ using the value for different activity measures (*e.g.*, IC_{50}, K_i, % of inhibition, etc.) with the goal of normalizing the bioactivity data from different experiments] and was obtained from Reaxys Medicinal Chemistry [31]. Before the VS, the 3D structures of the actives set were generated with OMEGA v2.5.1.4 [32] allowing just one conformation for each input molecule. Furthermore, the decoys were obtained with a modified version of DecoyFinder [33] that selected as decoys any molecule with a molecular weight (MW) within the range of the actives (resulting in a MW range of 215-586 Da for actives and of 300-440 Da for decoys).

Ligands for lead discovery purposes were downloaded from the purchasable subset of the ZINC12 database [34], which contains more than 16 million compounds. QikProp v4.5 [35] was then used to filter the ZINC molecules to discard those with bad ADME/Tox properties. Thus, only molecules that simultaneously fulfill the following drug-like properties were considered for the next step of the VS workflow: **(a)** MW at the 300-700 Da range; **(b)** only one violation for Lipinski's rule of five (predicted through the *RuleOfFive* parameter) [36]; **(c)** a maximum of 2 reactive/toxic functional groups (predicted through the *rtvFG* parameter); **(d)** high human oral absorption (predicted through the *HumanOralAbsorption* parameter, which must have a value of 3) [37]; and **(e)** number of property or descriptor values that fall outside the 95% range of similar values for known drugs at the 0-5 range.

One of the most important challenges of this VS workflow is that it should be able to find new lead molecules with basal DPP-IV bioactivity but with extremely low similarity to known active compounds. Consequently, the RDKit-Torsion fingerprint [38,39] was used to label all ZINC molecules that fulfill the previously described ADME/Tox filter according to their chemical structure. The same procedure was used to label 33 known inhibitors co-crystallized with DPP-IV with an IC_{50} better than 100 nM (see Table 1 and Figure S1). Then the

similarity between the RDKit-Torsion fingerprints of each co-crystallized inhibitor and each ZINC molecule was calculated using the Tanimoto coefficient [40]. For each comparison between one specific ZINC molecule and the different co-crystallized inhibitors, only the highest Tanimoto value was kept. Finally, ZINC molecules were sorted by decreasing the Tanimoto coefficient and only the bottom 1% of the sorted list (with coefficients in the range of 0.09474 to 0) was kept for the next step of the VS filter (see Figure 1).

2.2. Description and validation of the virtual screening

2.2.1. Ligand and protein setup

Before the docking filter, the 3D structure of all the remaining molecules (either actives or decoys or ZINC molecules for lead discovery) was prepared with LigPrep v3.5 [35] with the following settings: **(a)** the force field OPLS 2005 was used; **(b)** all possible ionization and tautomerization states at pH 7.0 ± 2.0 were generated with Epik; **(c)** the desalt option was activated; **(d)** chirality from input geometry was kept when generating stereoisomers; and **(e)** one low-energy ring conformation per ligand was generated.

The Protein Preparation Wizard (PPW) panel [35] was used to set up DPP-IV protein for use as a target in the following VS steps. Thus, chain A was prepared for two different PDB entries (*i.e.*, 1X70 and 3G0B [41,42]) in order to cover both possible positions of residue Tyr547 (for which the dihedral angle changes by 70° between the two orientations; see Figure 2) [43]. During the processing and refining steps of the PPW, all options were set to default with the exception of *remove original hydrogens*, *fill in missing side chains* and *cap termini* options, which were set to on.

2.2.2. Protein-ligand docking during the VS

During the VS workflow, protein-ligand docking studies were carried out using Glide v6.8 [35] with the following settings: **(a)** two different binding sites for DPP-IV were defined by using the previously curated coordinates of the two PDB files (*i.e.*, 1X70 and 3G0B) with the *Schrödinger's Grid Generation* panel (default options were used); **(b)** the standard precision mode (*i.e.*, SP) was used; **(c)** the maximum number of poses per ligand was increased to 32; and **(d)** the number of poses per ligand included in the post-docking minimization was increased to 32. The default values were used for the remaining docking parameters.

2.2.3. Structure-based pharmacophore screening

Docked poses were filtered through a couple of structure-based pharmacophores that were built to take into account the two different conformations that Tyr547 can adopt (see Figure 2). In this regard, docked poses obtained with 1X70 were filtered with the pharmacophore shown in Figure 2A whereas those obtained with 3G0B were filtered with the pharmacophore shown in Figure 2B. Both structure-based pharmacophores were designed by considering the most important interactions described for DPP-IV inhibition [44–47]. Thus, the two pharmacophores share three of the sites (*i.e.*, **P/D**, **H/R1** and **R2** that are associated, respectively, with interactions with the Glu205/Glu206 dyad, the S_1 pocket and the S_2 extensive subsite), whereas the location of the fourth site (*i.e.*, **R3**) depends on the conformation adopted by Tyr547 (see Figure 2). Associated tolerances are 2.3Å for **P/D**, 2.0Å for **H/R1**, 2.5Å for **R2** and 1.8Å for **R3**. Docked poses were filtered by the corresponding pharmacophore by using Phase v4.4 [35] and with the *score in place* option set

to on (*i.e.*, no re-orientation of the docked poses was allowed during the search). Thus, only docked poses simultaneously matching at least three pharmacophore sites (*i.e.*, **P/D**, **H/R1** and either **R2** or **R3**) were kept for the next VS filter.

2.2.4. Electrostatic and shape similarity screening

The software EON v2.2.0.5 [32] compares the poses for two different compounds by calculating the Tanimoto coefficients associated with either their electrostatic potentials, their shape or the combination of the Poisson-Boltzmann electrostatics and their shape. Thus, for the electrostatic potentials, the Tanimoto score is a value between -⅓ (*i.e.*, overlap of opposite charges between the two poses) and 1 (*i.e.*, identical electrostatic potential overlap). For the shape, the Tanimoto score is a value between 1 (*i.e.*, the same shape) and 0.

Fifteen complexes between DPP-IV and potent and selective non-covalent inhibitors (*i.e.*, 1X70, 2FJP, 2HHA, 2OLE, 2OPH, 2P8S, 2QOE, 2RGU, 3G0B, 3HAB, 3HAC, 3KWF, 3O95, 3WQH, 4PNZ) were superposed to 1X70 and 3G0B with the help of the PPW panel [35]. Then the experimental poses of their ligands were used during EON comparisons with the docked poses of the actives and decoys that passed the previous VS filter. This allows to find which of the Tanimoto scores provided by EON parameters (and considering too the combination of the coulombic part of the Poisson-Boltzmann electrostatics and their shape) and which threshold for these Tanimotos scores produces a better enrichment factor (see Figures 3, S2 and S3) and to determine the influence of the combination of the coulombic part of the Poisson-Boltzmann electrostatics and their shape. The highest value obtained for each Tanimoto score was kept from each comparison of one docked pose with the set of fifteen experimental poses.

2.3. Hit selection for further experimental assays on DPP-IV activity

The RDKit-Torsion fingerprints [39] of all the ZINC molecules that passed all the VS workflow filters were clustered on the basis of their Tanimoto similarities. Five structurally different compounds were then selected for *in vitro* assays of DPP-IV inhibitory activity on the basis of their commercial availability, cost and low chemical similarity to any molecule that has been experimentally shown to be bioactive as a human DPP-IV inhibitor (see Figure 4). These compounds were ZINC04299461, ZINC03823281, ZINC02751967, ZINC49076645 and ZINC71902582 and they were purchased from either Ambinter c/o Greenpharma (Orléans, France) or Epsilon Chimie (Brest-Guipavas, France).

2.4. In vitro assay of selected compounds' inhibition of DPP-IV

The DPP-IV enzyme purified from porcine kidney (product number 317640, Merck Millipore Corporation) was used to evaluate the effect of the selected compounds on DPP-IV activity. Stock solutions of DMSO diluted compounds were made and were subsequently diluted in a 50 mM Tris-Hcl buffer to a final concentration of 500 µM (1% DMSO concentration in the assay). The DPP-IV enzyme (diluted with 100 mM Tris HCl buffer pH 8.0 to 0.26 mU per well) and a test sample (10 µL) with a different concentration were pre-incubated for 10 min at 37°C using 96-well microplates to allow compound/enzyme interaction. Next, the enzymatic assay was initiated by the addition of 50 µL of the fluorimetric substrate H-Gly-Pro-AMC [product number I-1225, purchased from Bachem (Bubendorf, Switzerland)] at a final concentration of 0.01mM. Fluorescence was

measured in a Biotek FLx800 Fluorescence Microplate Reader at Ex:380nm/Em:460nm and 37°C for 30 min. Sitagliptin, a well-known DPP-IV inhibitor (that non-covalently binds to DPP-IV), was used as reference inhibitor and positive control. At least three independent assays were performed, each with two technical replicates. DPP-IV inhibition is expressed as a percentage, which is the difference of the activity in presence of test compounds versus the total activity of enzyme. Significant results showed $p<0.05$ with a Student's T test (SPSS software; SPSS, Chicago, USA).

2.5. Lead-optimization from ZINC02751967

Lead optimization was performed with CombiGlide v3.9 [35] by using the Virtual Combinatorial Screening workflow. The core-containing molecule was a derivative of the lowest-energy docked pose for ZINC02751967 with the ethoxycarbonyl moiety of the original molecule removed (compare Figures 5A and 6A) and the substituents were obtained from the Schrödinger CombiGlide Diverse Side-chain Collection v1.2 [35] (that has the most probable ionization and tautomeric states for 817 common functional groups in known drugs together with linkers of variable length). In two consecutive steps we established the points where the substituents had to be attached with the aim of improving the interactions at the: **(a)** S_1 pocket; and **(b)** S_2 extensive subsite of DPP-IV.

Thus, the following parameters were used for a single-position docking run: **(a)** the receptor grid for PDB code 3G0B was the same as the one previously used at protein-ligand docking step; **(b)** the *apply Glide core constraints* option was used within a maximum RMSD of 1.0Å; **(c)** the *Fully enumerated* option was selected; and **(d)** the *CombiGlide XP docking* mode was used. After this process, the resulting derivatives were filtered with the ADME filter set as *Druglike*. During this ADME filter, only two violations of the following criteria were allowed: **(a)** molecular weight less than 500 Da; **(b)** a maximum of 5 hydrogen bond donors; **(c)** a maximum of 10 hydrogen bond acceptors; **(d)** a predicted octanol/water partition coefficient (*i.e.*, logP) less than 5; **(e)** 10 or fewer rotatable bonds; and **(f)** a 150 $Å^2$ or less Van der Waals surface area. Next, the derivatives were filtered with the same pharmacophore, but a fourth compulsory site was required during the S_2 site enlargement (*i.e.*, **R2**). Finally, these poses were re-docked with the following settings: **(a)** the receptor grid for PDB code 3G0B was the same as the one previously used for the protein-ligand docking step; **(b)** the extra precision mode (*i.e.*, XP) was used; **(c)** the *Refine* option was used; and **(d)** the maximum number of poses per ligand was increased to 10.

3. Results and discussion

3.1. Description and validation of the virtual screening

A VS strategy was applied using sequential filters where the output molecules of one step were the input molecules for the next step and so on (see Figure 1). These steps mean that large libraries can be narrowed down to those molecules that are most likely to inhibit DPP-IV. Consequently, the following filters were applied in the present study: **(a)** ADME/tox filter (exclusively for VS purposes); **(b)** fingerprint similarity analyses with co-crystallized ligands (exclusively for VS purposes); **(c)** protein-ligand docking; **(d)** pharmacophore screening; and **(e)** shape/electrostatic analysis. It is important to note that the binding site of

DPP-IV is quite rigid except for the residue Tyr547, which can adopt two different orientations [43]. Consequently, two different crystallized structures for DPP-IV (each with one of the two conformations for Tyr547; see Figure 2) were used during the protein-ligand docking step and the pharmacophore filter to take into account this flexibility.

The reliability of the VS workflow was evaluated using a starting database of 419 actives for DPP-IV and 15,084 decoys. Figure 1 shows how many actives and decoys remained after applying each VS step. The first filter applied to this set of actives and decoys —based on docking and pharmacophore screening— placed the ligands in the binding site of the target requiring the most important interactions for the DPP-IV inhibition (*i.e.*, salt bridges/hydrogen bonds with the N-terminal recognition region, hydrophobic interactions with the S_1 pocket, interaction in S_2 pocket and π-π stacking with Tyr547) [44–47]. Thus, the number of actives and decoys were reduced to 267 and 6,363, respectively, but without a significant enrichment factor (*i.e.*, 1.5). However, the subsequent shape/electrostatic-potential comparison with co-crystallized ligands became a highly discriminative filter. Thus, Figures 3A and S2 show the distribution of the Tanimoto scores for actives and decoys provided by the different parameters calculated by EON [32]. A threshold for each parameter was established in order to recover the largest number of actives and remove the highest number of decoys. Then, a first cutoff of 0.7 for the coulombic part of the Poisson-Boltzmann electrostatics allowed us to remove nearly 70% of decoys and with almost no impact on the number of actives (*i.e.*, 2,039 and 236 molecules remain after the cutoff, respectively; see Figure 3A). In order to further concentrate the sample of active molecules, the Tanimoto scores for the remaining docked pose of each molecule were plotted again (see Figures 3B and S3). A second cutoff of 1.5 considering the combination of the coulombic part of the Poisson-Boltzmann electrostatics and their shape fields resulted in 101 actives and 137 decoys (see Figure 3B). Therefore, this two-step shape/electrostatic-potential filter produces an enrichment factor of 10.5 relative to the previous pharmacophore screening because the number of decoys was strongly reduced (from 6,363 to only 137) in comparison with the number of actives (from 267 to 101). Overall, the enrichment factor of the complete VS workflow is 15.7 and, therefore, these results show that this VS protocol is able to discern those molecules that can inhibit DPP-IV from those that do not affect its activity.

Once the VS workflow had been validated, it was decided to use the purchasable subset of the ZINC database [34] as the source of molecules for finding new lead structures with no (or very low) similarity to known DPP-IV inhibitors. To this end, the set of 16,538,200 molecules in that subset were first submitted to a filter aimed at discarding molecules that were either potentially toxic or exhibited poor ADME properties (see Figure 1). This filter reduced the number of molecules to be screened to 9,362,907 (see Figure 1). Then, the Tanimoto coefficient was calculated from the fingerprints in order to find those structures that could contribute to new scaffolds and that were significantly different from co-crystallized inhibitors. After sorting these 9,362,907 molecules in descending order of their Tanimoto coefficients, only the bottom 1% of the resulting list (with values from 0 to 0.095 for that coefficient) was selected for the next filter (*i.e.*, 93,629 molecules). From the docking and pharmacophore screening, we obtained 24,034 compounds with at least one pose simultaneously filling the S_1 pocket and interacting with the Glu dyad (*i.e.*, **H/R1** and **P/D**, respectively) and either filling the S_2 pocket or interacting with Tyr547 (*i.e.*, **R2** and **R3**, respectively; see Figure 2). Finally, these poses were submitted to a shape/electrostatic comparison with known experimental

poses for actives and only 404 structures were identified as VS hits with potential DPP-IV inhibitory bioactivity (see Figure 1).

3.2. Structure-activity relationship of selected compounds regarding the inhibition of DPP-IV

As mentioned before, the main goal of this study was to describe a computational strategy able to find new leads with no (or very low) similarity to known actives for DPP-IV. Thus, in order to select which of the 404 VS hits could be considered as new lead molecules and experimentally to test their bioactivity as DPP-IV inhibitors, we: **(a)** ensured that there were sufficient structural differences between the known DPP-IV inhibitors and the VS hits; **(b)** ensured that there were sufficient structural differences between the VS hits themselves; **(c)** visually inspected the docking poses; and **(d)** took into account commercial availability and cost. As result of these steps, the compounds ZINC04299461, ZINC03823281, ZINC02751967, ZINC49076645 and ZINC71902582 were selected in order to experimentally test their effects on the DPP-IV activity. Interestingly, when these five compounds were submitted to the SwissTargetPrediction webserver [48], DPP-IV was identified as a likely biological target for three of them (*i.e.*, ZINC03823281, ZINC04299461 and ZINC71902582 with probability values of 0.11, 0.29 and 0.30; respectively) which reinforce the VS results.

Figure 4 is a dendrogram in which the five selected hits are clustered according to their fingerprint similarity, thus revealing their structural diversity (*i.e.*, the maximum Tanimoto score among them is 0.0968 between ZINC49076645 and ZINC71902582). Moreover, a set of 15,024 molecules which have experimental bioactivity values for human DPP-IV were downloaded from Reaxys Medicinal Chemistry [31] for a fingerprint similarity analysis. After calculating the RDKit-Torsion [39] fingerprint for all of them, the highest Tanimoto coefficient was kept for comparison with the five selected hits. As Figure 4 shows, the highest Tanimoto is for the compound ZINC03823281 (*i.e.*, Tanimoto value of 0.536) which shares part of its structure with XRN.24962630 while the remaining structures are sufficiently different from the associated molecule.

The *in vitro* bioactivity assay shows that two out of these five selected compounds inhibit DPP-IV at 500 μM (*i.e.*, ZINC02751967 and ZINC03823281 significantly inhibited at 25.4% and 7.6%, respectively). Docking of these two compounds in the DPP-IV binding site (PDB entry 3G0B [42]) shows how these molecules match the main interactions determined by the pharmacophore (see Figure 5). From one side, both compounds use a primary or a secondary charged amine to interact with the N-terminal recognition region formed by the residues Glu205, Glu206 and Tyr662 [49]. From the other side, the hydrophobic S_1 pocket (formed by the residues Tyr631, Val656, Trp659, Tyr662, Tyr666 and Val711) [29] is filled by different moieties (*i.e.*, ZINC02751967 places an ethylsulfanyl group whereas ZINC03823281 places a phenyl ring). Therefore, both ligands are able to match the main sites for DPP-IV inhibition, these being: **(a)** a salt bridge and/or hydrogen bond interactions with the N-terminal recognition region (*i.e.*, **P/D** site of the pharmacophore; see Figure 2); and **(b)** hydrophobic contacts with the S_1 pocket (*i.e.*, **H/R1** site of the pharmacophore; see Figure 2) [44–46]. Additionally, ZINC02751967 and ZINC03823281 place a 5-methylfuran-3-yl group and a phenyl ring at **R3**, respectively, which allows them to interact with Tyr547 through a π-π stacking (see Figure 5). Tyr547 has been reported to be essential for the catalytic activity of the enzyme. Moreover, due to Tyr547 flexibility in DPP-IV (which is not possible in DPP8/9) it has been suggested, to be also involved in inhibitor selectivity

[50,51]. Finally, in the case of ZINC02751967, the cyano group is able to place a negative electrostatic environment around Arg125 (results not shown), whereas ZINC03823281 can interact with this residue by cation-π interaction (see Figure 5).

3.3. Lead optimization from ZINC02751967

The *in vitro* experiments demonstrated that, of the two VS hits that are bioactive as a DPP-IV inhibitors, ZINC02751967 is the more potent (see Figure 4). Consequently, this molecule is a promising lead compound for designing potent and selective DPP-IV inhibitors with very low similarity to existing actives. The lowest-energy docked pose of ZINC02751967 was used as the starting point for lead optimization (the XP GScore was -4.378 kcal/mol; see Figure 5A) with the purpose of improving the binding affinity with DPP-IV according to the most important interactions described for this target [44–47]. Therefore, an optimization process has been developed (see Figure 6) in order: **(a)** to improve the occupation of the hydrophobic S_1 pocket (*i.e.*, the **H/R1** site of the pharmacophore); and **(b)** to reach the S_2 extensive subsite (i.e., the **R2** site of the pharmacophore).

Firstly, the ethoxycarbonyl group was removed from the original structure of ZINC02751967 (see Figure 6A) because this fragment is not able to interact with any residue (see Figure 5A) and because of its very low contribution to the XP GScore (from -4.378 for ZINC02751967 to -4.122 kcal/mol for ZINC02751967 without the ethoxycarbonyl moiety). Moreover, this new ZINC02751967 derivative (*i.e.*, ZINC02751967-dev) allowed us to attach bigger substituents in order to reach the pockets due to the reduction of its molecular weight (which is within the parameters of Lipinski's rule during the combinatorial screening). Next, the lead optimization process initially focused on introducing a moiety that could better fill the hydrophobic S_1 cavity than the initial ethylsulfanyl group (*i.e.*, $\mathbf{R^1}$ label; see Figure 6A). As a result, ZINC02751967-dev-283 was selected on the basis of its XP GScore (*i.e.*, -7.072 kcal/mol; see Figure 6B). At this point, it is important to note that it has been experimentally shown that a better occupancy of the S_1 pocket results in higher bioactivities for DPP-IV inhibitors [52,53]. Consequently, the substitution of the original ethylsulfanyl substituent by a positively charged pyridin-4-yl moiety is expected to help increase the bioactivity of ZINC02751967-dev-283 relative to ZINC02751967-dev. Moreover, the docked pose of this new compound, not only maintains the original intermolecular interactions with the DPP-IV binding site (*i.e.*, two hydrogen bonds with residues of the N-terminal recognition region and the π-π stacking interaction with Tyr547) but also shows additional π-π and cation-π interactions with Tyr666 (which reinforces the hypothesis that ZINC02751967-dev-283 is a better DPP-IV inhibitor than ZINC02751967-dev). At this point, it is interesting to remark that these results are coherent with those previously found by our group that show that improvements in the bioactivity of DPP-IV inhibitors can be obtained by replacing an alkyl substituent at the **H/R1** site by a group that can bind with the lipophilic atoms of the S_1 pocket either by means of the so-called hydrophobic enclosure (where the two sides of the substituent are enclosed –at a 180º angle– on the hydrophobic environment of the S_1 pocket) or by means of π-cation interactions with the different aromatic side chains in this pocket [54]. In that sense, the positively charged pyridin-4-yl substituent of ZINC02751967-dev-283 would be able to perform, simultaneously, both kind of interactions.

A second optimization step was performed in order to reach the S_2 extensive subsite. This pocket has been shown to enhance the activity and selectivity of DPP-IV by interacting with Ser209, Phe357 and Arg358

[45,53,55]. Consequently, another point of attachment was placed in ZINC02751967-dev-283 (*i.e.*, **R²** label; see Figure 6B). The top eight derivatives of this second optimization step had docked poses that were able to further increase the XP GScore (*i.e.*, in the -8.800 to -7.319 Kcal/mol range; see Table 2). Most of them are expected to increase the binding affinity and selectivity for DPP-IV by either interacting with Ser209 through a hydrogen bond or by making a π-π stacking interaction with Phe357 (see Figures 6C, 6D and 6E). The 2D structures for the eight best docked poses and their corresponding XP GScore values are shown in Table 2.

4. Conclusion

The design of the computational strategy used in this study has been demonstrated to be suitable for identifying new lead compounds in purchasable databases with very low (or no) similarity to known actives. Therefore, this VS workflow is a good alternative to other computational approaches such as bioisosteric replacement and fragment-based drug design because it reduces the cost and time of designing new potent actives; that is, by using this VS workflow, the synthetic effort focuses solely on improving a core structure with the desired basal bioactivity for the target of interest. Moreover, this computational strategy is significantly faster than protein-ligand docking or pharmacophore screening. For instance, benchmarking studies have estimated that the fastest docking mode available in Glide (*i.e.*, HVTS) needs around 1.5 seconds to dock a ligand in a binding site by using a 2.2 GHz Opteron processor (*i.e.*, around 60,000 compounds per day) [56]. Additionally, obtaining conformers (a previous step for screening ligands with a pharmacophore) needs about 2 seconds per compound with a similar processor (*i.e.*, around 43,000 compounds per day) by using OMEGA with default parameters [57]. In contrast, with similar computational resources, we can reduce an initial sample of 9,362,907 molecules to the 1% of interest in only 6 hours (thus focusing computational resources only on those molecules that are potential candidates for finding new leads).

5. Future perspective

It is well assumed that the number of putative drug-like molecules is many orders of magnitude higher than the amount in current libraries and that, with molecular size, this number grows exponentially [58]. For instance, it is estimated that more than 10^{60} different molecules can be built by combining up to 30 atoms of C, N, O and S [59]. Thus, considering the importance of diversifying molecular scaffolds for improving the chances of success in drug discovery, it becomes evident that there is a strong need for computational protocols that can efficiently explore such vast virtual molecular libraries looking for these new scaffolds [60]. In that sense, our manuscript aims to offer a computationally cheap alternative to more computational demanding strategies that also share this goal [61]. Moreover, in our opinion, this strategy can be valid to other targets of pharmacological interest with similar success. In that sense, we plan to apply it to find completely new leads for PTP1B, MMP-13 and ACE.

6. Executive summary

- An *in silico* strategy is designed to find leads with no similarity to known actives.
- This fast strategy to mine large databases is based on a fingerprint similarity analysis which is performed to select these new scaffolds.
- This computational protocol is applied to a target of pharmacological interest, as DPP-IV, involved in the type II diabetes treatment.
- ZINC02751967 which has a Tanimoto value of 0.123 in comparison to the known DPP-IV inhibitors, experimentally confirmed to be new lead as DPP-IV inhibitor.
- ZINC02751967 is used as a lead compound to computationally suggest how to improve potency and selectivity.
- This computational approach to find new scaffolds can be valid to other targets of pharmacological interest with similar success in order to reduce cost and time.

Supplementary data

Supplementary data accompany this paper.

Financial & competing interests disclosure

This study was supported by research grants 2016PFR-URV-B2-67 and 2017PFR-URV-B2-69 from our University. AG's contract is supported by grant 2015FI_B00655 from the Catalan Government. MP and MM are Serra Húnter research fellows. We thank OpenEye Scientific Software Inc. for generously providing us with an academic license to use their programs. This manuscript has been edited in accordance with the English language usage of our University. The authors have no other relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript apart from those disclosed.

Table 1. Ligands from these PDB codes of DPP-IV complexes were used as references to calculate Tanimoto similarity with ZINC molecules based on RDKit-Torsion fingerprint (see the ligand structures at Figure S1).

Table 2. Best ZINC02751967-dev-283 derivatives obtained in the second optimization step. Molecules are sorted according to XP GScore. The name for each derivative was built by adding the code of the attached fragment (according to the CombiGlide Diverse Side-chain Collection) to the lead name.

Figure 1. The VS workflow used in the present study. The data corresponds to the number of molecules that remains after each VS step. The actives and decoys columns correspond to those molecules used for validating the VS. The ZINC column refers to data obtained when looking for new leads for DPP-IV inhibition. Enrichment factors were calculated during the validation for each step of the VS protocol as the quotient between the fraction of actives in the sample that survived the VS step and the fraction of actives in the sample before the VS step.

Figure 2. Structure-based pharmacophores used in this paper based on the crystal protein-ligand complex for the most important interactions. The difference between the two pharmacophores is due to the two different conformations of the residue Tyr547 (colored in pink) shown in the context of **(A)** the 1X70 active site and **(B)** the 3G0B active site. The pharmacophores are formed by a positive/hydrogen-bond donor feature (*i.e.*, **P/D**), a hydrophobic/aromatic ring site (*i.e.*, **H/R1**) and two aromatic ring sites (*i.e.*, **R2** and **R3**). The associated tolerances are 2.3Å for **P/D**, 2.0Å for **H/R1**, 2.5Å for **R2** and 1.8Å for **R3**. Two sites (*i.e.*, **P/D** and **H/R1**) together with a third site of the two remaining (*i.e.*, **R2** and **R3**) are required during the pharmacophore-based searches.

Figure 3. Histograms showing the distribution of the highest Tanimoto values for actives (shown in red) and decoys (shown in gray) for **(A)** the Electrostatic Tanimoto of the coulombic part of the Poisson-Boltzmann field and **(B)** the Electrostatic Tanimoto of the coulombic part of the Poisson-Boltzmann and Shape Tanimoto fields. Two consecutive cutoffs (red line) were applied to the set of actives and decoys by using these EON parameters in order to increase the enrichment factor of the VS validation.

Figure 4. Dendrogram of the five hits selected for experimental testing as a result of the VS workflow (framed in red). The dendrogram shows the distances of the Tanimoto coefficient which represents the fingerprint similarity of the hits. Each hit is attached to a chemical structure downloaded from the Reaxys database which has experimental bioactivity values for human DPP-IV (framed in blue). This molecule is the most similar in terms of fingerprint similarity to the VS hit. Compounds ZINC02751967 and ZINC03823281 are the only ones which show significant *in vitro* DPP-IV inhibition.

Figure 5. The best docked poses (with the corresponding XP GScore) for the compounds ZINC02751967 and ZINC03823281. Blue and orange dashed lines show π-π stacking and cation-π intermolecular interactions, respectively, whereas the red ones show either salt bridges (between the positively charged amine from ZINC03823281 and Glu206) or hydrogen bonds. Both panels are oriented the same way for easy comparison.

Figure 6. Lead optimization of a derivative of ZINC02751967 used with the aim of obtaining new molecules with improved potency and selectivity for DPP-IV. First, the ethoxycarbonyl group was removed from the initial ZINC02751967 (Figure 6A) because of its low contribution to the protein-ligand interaction (see Figure 5A). Then a substituent was attached to the ethylsulfanyl group of this ZINC02751967 derivative (*i.e.*, R^1 label) in order to improve the occupancy of S_1 pocket. The resulting derivative (Figure 6B) was selected for further optimization. Next, another point for attaching the substituents (*i.e.*, R^2 label) was placed in these new derivatives in order to reach the S_2 extensive subsite. The docked poses of some of the most potent derivatives after this second optimization step are shown (Figure 6C-6E). The name for each derivative was built by adding the code of the attached fragment (according to the CombiGlide Diverse Side-chain Collection) to the lead name (see also in the Table 2 the 2D structure and XP GScore for the best eight derivatives obtained during the second optimization step). Blue and orange dashed lines show π-π stacking and cation-π intermolecular interactions, respectively, whereas the red ones show the hydrogen bonds.

References

1. Jain S, Jacob M, Walker L, Tekwani B. Screening North American plant extracts *in vitro* against *Trypanosoma brucei* for discovery of new antitrypanosomal drug leads. *BMC Complement. Altern. Med.* 16, 131 (2016).
2. Ghorab MM, Alsaid MS. Novel 4-aminoquinazoline derivatives as new leads for anticancer drug discovery. *Acta Pharm.* 65(3), 299–309 (2015).
3. Brown DG, Lister T, May-Dracka TL. New natural products as new leads for antibacterial drug discovery. *Bioorg. Med. Chem. Lett.* 24(2), 413–8 (2014).
4. Pohlit AM, Lima RB, Frausin G, *et al.* Amazonian plant natural products: perspectives for discovery of new antimalarial drug leads. *Molecules.* 18(8), 9219–40 (2013).
5. Martins da Silva SJ, Brown SG, Sutton K, *et al.* Drug discovery for male subfertility using high-throughput screening: A new approach to an unsolved problem. *Hum. Reprod.* 32(5), 974–984 (2017).
6. Alonso-Padilla J, Rodríguez A. High throughput screening for anti-*Trypanosoma cruzi* drug discovery. *PLoS Negl. Trop. Dis.* 8(12), e3259 (2014).
7. Kessel S, Cribbes S, Déry O, *et al.* High-throughput 3D tumor spheroid screening method for cancer drug discovery using celigo image cytometry. *J. Lab. Autom.* 1–12 (2016).
8. Annang F, Pérez-Moreno G, García-Hernández R, *et al.* High-throughput screening platform for natural product-based drug discovery against 3 neglected tropical diseases: Human African trypanosomiasis, leishmaniasis, and Chagas disease. *J. Biomol. Screen.* 20(1), 82–91 (2015).
9. Chen S, Feng Z, Wang Y, *et al.* Discovery of novel ligands for TNF-α and TNF receptor-1 through structure-based virtual screening and biological assay. *J. Chem. Inf. Model.* 57(5), 1101–1111 (2017).
10. Froes TQ, Melo MCC, Souza GEP, Castilho MS, Soares DM. Virtual screening and biological evaluation of novel antipyretics compounds. *Chem. Biol. Drug Des.* 38(1), 42–49 (2017).
11. Cui W, Lv W, Qu Y, *et al.* Discovery of 2-((3-cyanopyridin-2-yl)thio)acetamides as human lactate dehydrogenase A inhibitors to reduce the growth of MG-63 osteosarcoma cells: Virtual screening and biological validation. *Bioorg. Med. Chem. Lett.* 26(16), 3984–7 (2016).
12. Talari FS, Bagherzadeh K, Golestanian S, Jarstfer M, Amanlou M. Potent human telomerase inhibitors: Molecular dynamic simulations, multiple pharmacophore-based virtual screening, and biochemical assays. *J. Chem. Inf. Model.* 55(12), 2596–610 (2015).
13. Mendes V, Blundell TL. Targeting tuberculosis using structure-guided fragment-based drug design. *Drug Discov. Today.* 22(3), 546–554 (2017).
14. Benmansour F, Trist I, Coutard B, *et al.* Discovery of novel dengue virus NS5 methyltransferase non-nucleoside inhibitors by fragment-based drug design. *Eur. J. Med. Chem.* 125, 865–880 (2017).
15. Doak BC, Norton RS, Scanlon MJ. The ways and means of fragment-based drug design. *Pharmacol. Ther.* 167, 28–37 (2016).
16. Tuyishime M, Lawrence R, Cocklin S. Core chemotype diversification in the HIV-1 entry inhibitor class using field-based bioisosteric replacement. *Bioorg. Med. Chem. Lett.* 26(1), 228–34 (2016).
17. Chandna N, Kumar S, Kaushik P, *et al.* Synthesis of novel celecoxib analogues by bioisosteric replacement of sulfonamide as potent anti-inflammatory agents and cyclooxygenase inhibitors. *Bioorg. Med. Chem.* 21(15), 4581–90 (2013).
18. Jiang Z, Wang Y, Wang W, *et al.* Discovery of highly potent triazole antifungal derivatives by heterocycle-benzene bioisosteric replacement. *Eur. J. Med. Chem.* 64, 16–22 (2013).

19. Sterling T, Irwin JJ. ZINC 15 – Ligand discovery for everyone. *J. Chem. Inf. Model.* 55(11), 2324–37 (2015).

20. Ojeda-Montes MJ, Gimeno A, Tomas-Hernández S, *et al.* Activity and selectivity cliffs for DPP-IV inhibitors: Lessons we can learn from SAR studies and their application to virtual screening. *Med. Res. Rev.* 38(6), 1874–1915 (2018).

**** Recent review on DPP-IV focusing on the binding site features to suggest how virtual screening protocols might be improved to favor the early identification of potent and selective DPP-IV inhibitors in molecular databases.**

21. Zeng S, Xie H, Zeng L, *et al.* Discovery of potent dipeptidyl peptidase IV inhibitors through pharmacophore hybridization and hit-to-lead optimization. *Bioorg. Med. Chem.* 21(7), 1749–55 (2013).

22. Juillerat-Jeanneret L. Dipeptidyl peptidase IV and its inhibitors: Therapeutics for type 2 diabetes and what else? *J. Med. Chem.* 57(6), 2197–212 (2014).

23. Mentlein R. Dipeptidyl-peptidase IV (CD26) – role in the inactivation of regulatory peptides. *Regul. Pept.* 85(1), 9–24 (1999).

24. Thoma R, Löffler B, Stihle M, Huber W, Ruf A, Hennig M. Structural basis of proline-specific exopeptidase activity as observed in human dipeptidyl peptidase-IV. *Structure.* 11(8), 947–59 (2003).

25. Nabeno M, Akahoshi F, Kishida H, *et al.* A comparative study of the binding modes of recently launched dipeptidyl peptidase IV inhibitors in the active site. *Biochem. Biophys. Res. Commun.* 434(2), 191–6 (2013).

26. Power O, Nongonierma A, Jakeman P, Fitzgerald R. Food protein hydrolysates as a source of dipeptidyl peptidase IV inhibitory peptides for the management of type 2 diabetes. *Proc. Nutr. Soc.* 73(1), 34–46 (2014).

27. Chien CH, Huang LH, Chou CY, *et al.* One site mutation disrupts dimer formation in human DPP-IV proteins. *J. Biol. Chem.* 279(50), 52338–45 (2004).

28. Zettl H, Schubert-Zsilavecz M, Steinhilber D. Medicinal chemistry of incretin mimetics and DPP-4 inhibitors. *ChemMedChem.* 5(2), 179–85 (2010).

29. Kuhn B, Hennig M, Mattei P. Molecular recognition of ligands in dipeptidyl peptidase IV. *Curr. Top. Med. Chem.* 7(6), 609–19 (2007).

30. Mysinger MM, Carchia M, Irwin JJ, Shoichet BK. Directory of useful decoys, enhanced (DUD-E): Better ligands and decoys for better benchmarking. *J. Med. Chem.* 55(14), 6582–94 (2012).

31. Reaxys Medicinal Chemistry. www.reaxys.com.

32. OpenEye Scientific Software, Santa Fe, NM. www.eyesopen.com.

33. Cereto-Massagué A, Guasch L, Valls C, Mulero M, Pujadas G, Garcia-Vallvé S. DecoyFinder: An easy-to-use python GUI application for building target-specific decoy sets. *Bioinformatics*. 28(12), 1661–1662 (2012).

34. Irwin JJ, Sterling T, Mysinger MM, Bolstad ES, Coleman RG. ZINC: A free tool to discover chemistry for biology. *J. Chem. Inf. Model.* 52(7), 1757–68 (2012).

35. Schrödinger Release 2015-4: Protein Preparation Wizard; Epik v3.3; Prime; LigPrep v3.6; QikProp v4.6; Glide v6.9; CombiGlide v3.9; Phase v4.5, Schrödinger, LLC, New York, NY (2015).

36. Lipinski CA, Lombardo F, Dominy BW, Feeney PJ. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Deliv. Rev.* 46(1–3), 3–26 (2001).

37. Kelder J, Grootenhuis PD, Bayada DM, Delbressine LP, Ploemen JP. Polar molecular surface as a dominating determinant for oral absorption and brain penetration of drugs. *Pharm. Res.* 16(10), 1514–9 (1999).

38. Nilakantan R, Bauman N, Dixon JS, Venkataraghavan R. Topological torsion: A new molecular descriptor for SAR applications. Comparison with other descriptors. *J. Chem. Inf. Model.* 27(2), 82–85 (1987).

39. RDKit, Open-Source Cheminformatics. www.rdkit.org.

40. Maggiora G, Vogt M, Stumpfe D, Bajorath J. Molecular similarity in medicinal chemistry. *J. Med. Chem.* 57(8), 3186–204 (2014).

41. Kim D, Wang L, Beconi M, *et al.* (2*R*)-4-oxo-4-[3-(trifluoromethyl)-5,6-dihydro[1,2,4]triazolo[4,3-a]pyrazin-7(8*H*)-yl]-1-(2,4,5-trifluorophenyl)butan-2-amine: A potent, orally active dipeptidyl peptidase IV inhibitor for the treatment of type 2 diabetes. *J. Med. Chem.* 48(1), 141–51 (2005).

42. Zhang Z, Wallace MB, Feng J, *et al.* Design and synthesis of pyrimidinone and pyrimidinedione inhibitors of dipeptidyl peptidase IV. *J. Med. Chem.* 54(2), 510–24 (2011).

43. Sheehan SM, Mest HJ, Watson BM, *et al.* Discovery of non-covalent dipeptidyl peptidase IV inhibitors which induce a conformational change in the active site. *Bioorg. Med. Chem. Lett.* 17(6), 1765–8 (2007).

44. Nojima H, Kanou K, Terashi G, *et al.* Comprehensive analysis of the Co-structures of dipeptidyl peptidase IV and its inhibitor. *BMC Struct. Biol.* 16(1), 11 (2016).

*** A comprehensive analysis of DPP-IV X-ray complexes with different inhibitors to clarify whether DPP-IV alters its binding site structure according to the inhibitor and whether this enzyme has a common rule for inhibitor binding**

45. Smelcerovic A, Miljkovic F, Kolarevic A, *et al.* An overview of recent dipeptidyl peptidase-IV inhibitors: Linking their structure and physico-chemical properties with SAR, pharmacokinetics and toxicity. *Curr. Top. Med. Chem.* 15(23), 2342–72 (2015).

*** Recent review on natural and synthetic DPP-IV inhibitors, focusing on the association between their chemical structure and mechanism of action.**

46. Liu Y, Hu Y. Novel DPP-4 inhibitors against diabetes. *Future Med. Chem.* 6(7), 793–808 (2014).

47. Patel B, Ghate M. Computational studies on structurally diverse dipeptidyl peptidase IV inhibitors: An approach for new antidiabetic drug development. *Med. Chem. Res.* 22(9), 4505–4521 (2013).

48. Gfeller D, Grosdidier A, Wirth M, Daina A, Michielin O, Zoete V. SwissTargetPrediction: A web server for target prediction of bioactive small molecules. *Nucleic Acids Res.* 42, W32-8 (2014).

49. Ojeda MJ, Cereto-Massagué A, Valls C, Pujadas G. DPP-IV, an important target for antidiabetic functional food design. In: *FoodInformatics, Applications of chemical information to food chemistry*. Mayorga M,Medina-Franco K, Luis J (Eds). Springer, Switzerland, 177–212 (2014).

50. Scapin G. Structural chemistry and molecular modeling in the design of DPP4 inhibitors. In: *Multifaceted roles of crystallography in modern drug discovery*. Scapin G, Patel D, Arnold E (Eds.). Springer Netherlands 53–67 (2015).

51. Bjelke JR, Christensen J, Branner S, *et al.* Tyrosine 547 constitutes an essential part of the catalytic mechanism of dipeptidyl peptidase IV. *J. Biol. Chem.* 279(33), 34691–7 (2004).

52. Liu Y, Hu Y, Liu T. Recent advances in non-peptidomimetic dipeptidyl peptidase 4 inhibitors: Medicinal chemistry and preclinical aspects. *Curr. Med. Chem.* 19(23), 3982–99 (2012).

53. Patel BD, Ghate MD. Recent approaches to medicinal chemistry and therapeutic potential of dipeptidyl peptidase-4 (DPP-4) inhibitors. *Eur. J. Med. Chem.* 74, 574–605 (2014).

54. Guasch L, Ojeda MJ, González-Abuín N, *et al.* Identification of novel human dipeptidyl peptidase-IV inhibitors of natural origin (part I): Virtual screening and activity assays. *PLoS One*. 7(9), e44971 (2012).

55. Rummey C, Metz G. Homology models of dipeptidyl peptidases 8 and 9 with a focus on loop predictions near the active site. *Proteins*. 66(1), 160–71 (2007).

56. Schrödinger – Knowledge Base. www.schrodinger.com/kb/1012.

57. Hawkins PCD, Skillman AG, Warren GL, Ellingson BA, Stahl MT. Conformer generation with OMEGA: Algorithm and validation using high quality structures from the Protein Databank and Cambridge Structural Database. *J. Chem. Inf. Model.* 50(4), 572–84 (2010).

58. Fink T, Bruggesser H, Reymond JL. Virtual exploration of the small-molecule chemical universe below 160 Daltons. *Angew. Chem. Int. Ed. Engl.* 44(10), 1504–8 (2005).

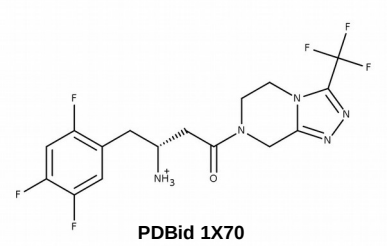
59. Bohacek RS, McMartin C, Guida WC. The art and practice of structure-based drug design: A molecular modeling perspective. *Med. Res. Rev.* 16(1), 3–50 (1996).

60. Reymond JL. The chemical space project. *Acc. Chem. Res.* 48(3), 722–30 (2015).

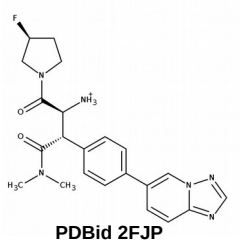
61. Lyu J, Wang S, Balius TE, *et al.* Ultra-large library docking for discovering new chemotypes. *Nature.* 566(7743), 224–229 (2019).

Supplementary Material

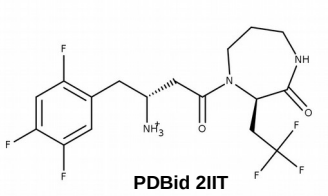
Mining large databases to find new leads with low similarity to known actives: application to find new DPP-IV inhibitors



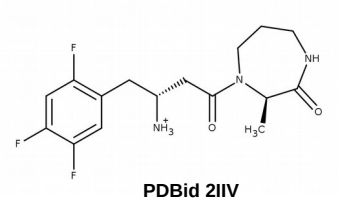
PDBid 1X70
IC_{50} = 18 nM



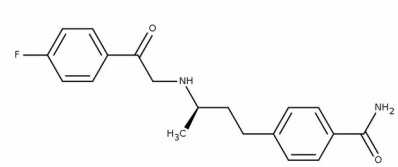
PDBid 2FJP
IC_{50} = 4.3 nM



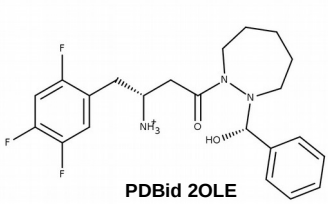
PDBid 2IIT
IC_{50} = 2.6 nM



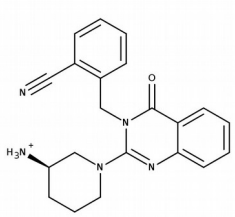
PDBid 2IIV
IC_{50} = 6.6 nM



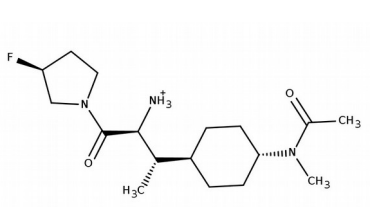
PDBid 2OGZ
IC_{50} = 84 nM



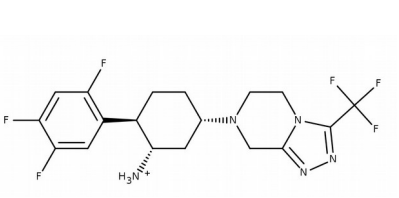
PDBid 2OLE
IC_{50} = 70 nM



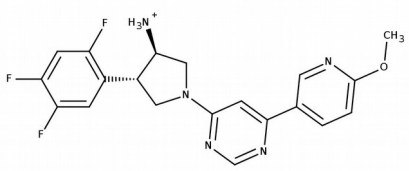
PDBid 2ONC
IC_{50} = 13 nM



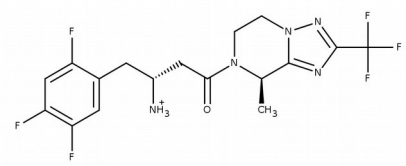
PDBid 2OPH
IC_{50} = 16 nM



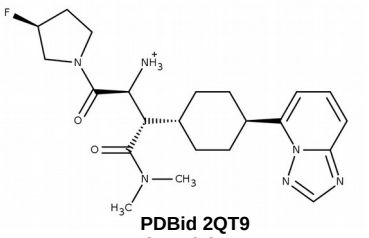
PDBid 2P8S
IC_{50} = 21 nM



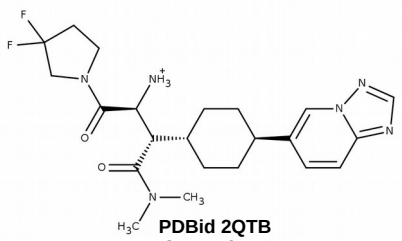
PDBid 2QJR
IC_{50} = 6.4 nM



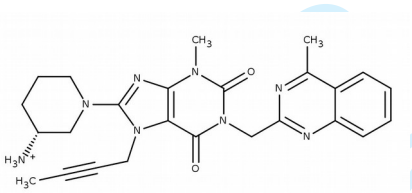
PDBid 2QOE
IC_{50} = 25 nM



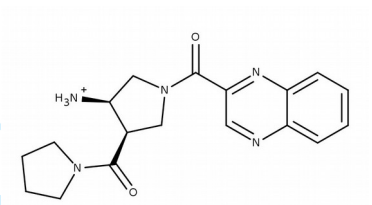
PDBid 2QT9
IC_{50} = 2.3 nM



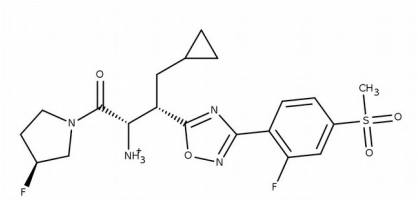
PDBid 2QTB
IC_{50} = 4.8 nM



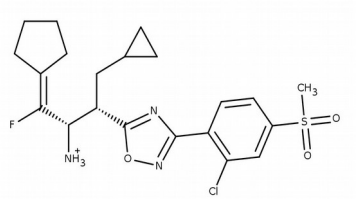
PDBid 2RGU
IC_{50} = 1 nM



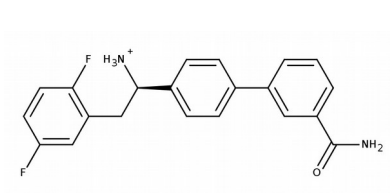
PDBid 2RIP
IC_{50} = 65 nM



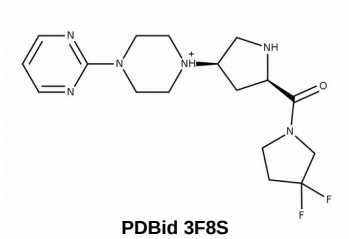
PDBid 3C43
IC_{50} = 19 nM



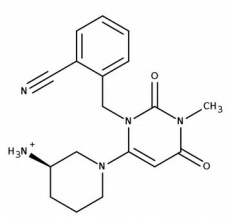
PDBid 3C45
IC_{50} = 0.21 nM



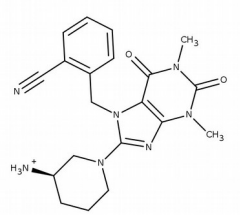
PDBid 3D4L
IC_{50} = 32 nM



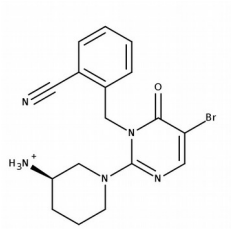
PDBid 3F8S
IC_{50} = 13 nM



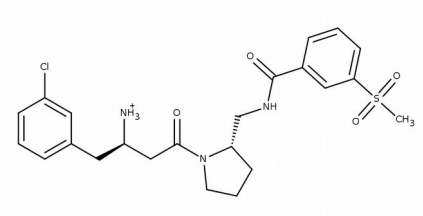
PDBid 3G0B
IC_{50} = 7 nM



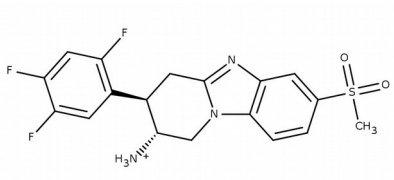
PDBid 3G0D
IC_{50} = 5 nM



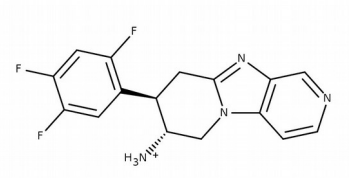
PDBid 3G0G
IC_{50} = 5 nM



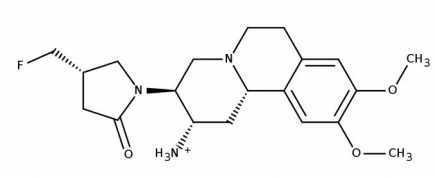
PDBid 3H0C
IC_{50} = 0.38 nM



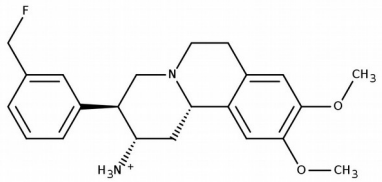
PDBid 3HAB
IC_{50} = 4.2 nM



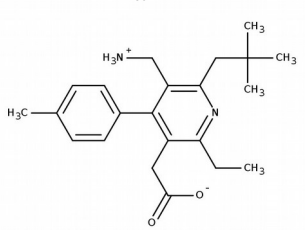
PDBid 3HAC
IC_{50} = 6.7 nM



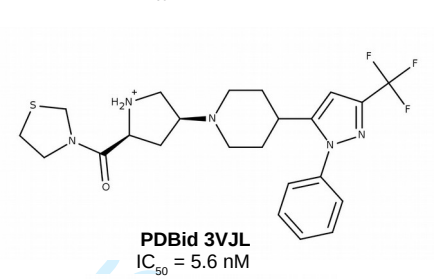
PDBid 3KWF
IC_{50} = 6.8 nM



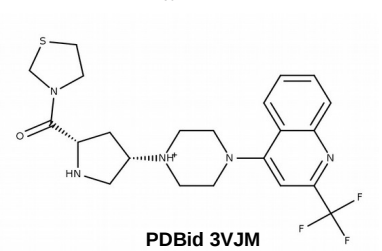
PDBid 3KWJ
IC_{50} = 0.5 nM



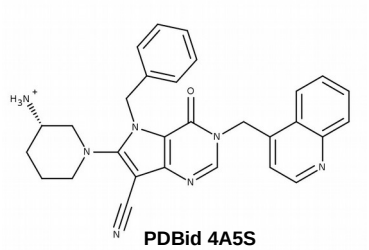
PDBid 3O95
IC_{50} = 5.3 nM



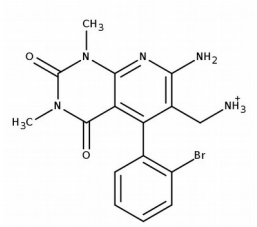
PDBid 3VJL
IC_{50} = 5.6 nM



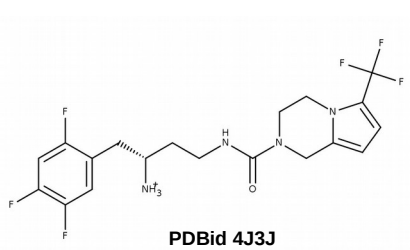
PDBid 3VJM
IC_{50} = 0.37 nM



PDBid 4A5S
IC_{50} = 17 nM



PDBid 4G1F
IC_{50} = 17 nM



PDBid 4J3J
IC_{50} = 31 nM

Figure S1. 2D structure of the inhibitors in PDB complexes with DPP-IV that were used as references to calculate Tanimoto similarity with ZINC molecules based on RDKit-Torsion fingerprint.

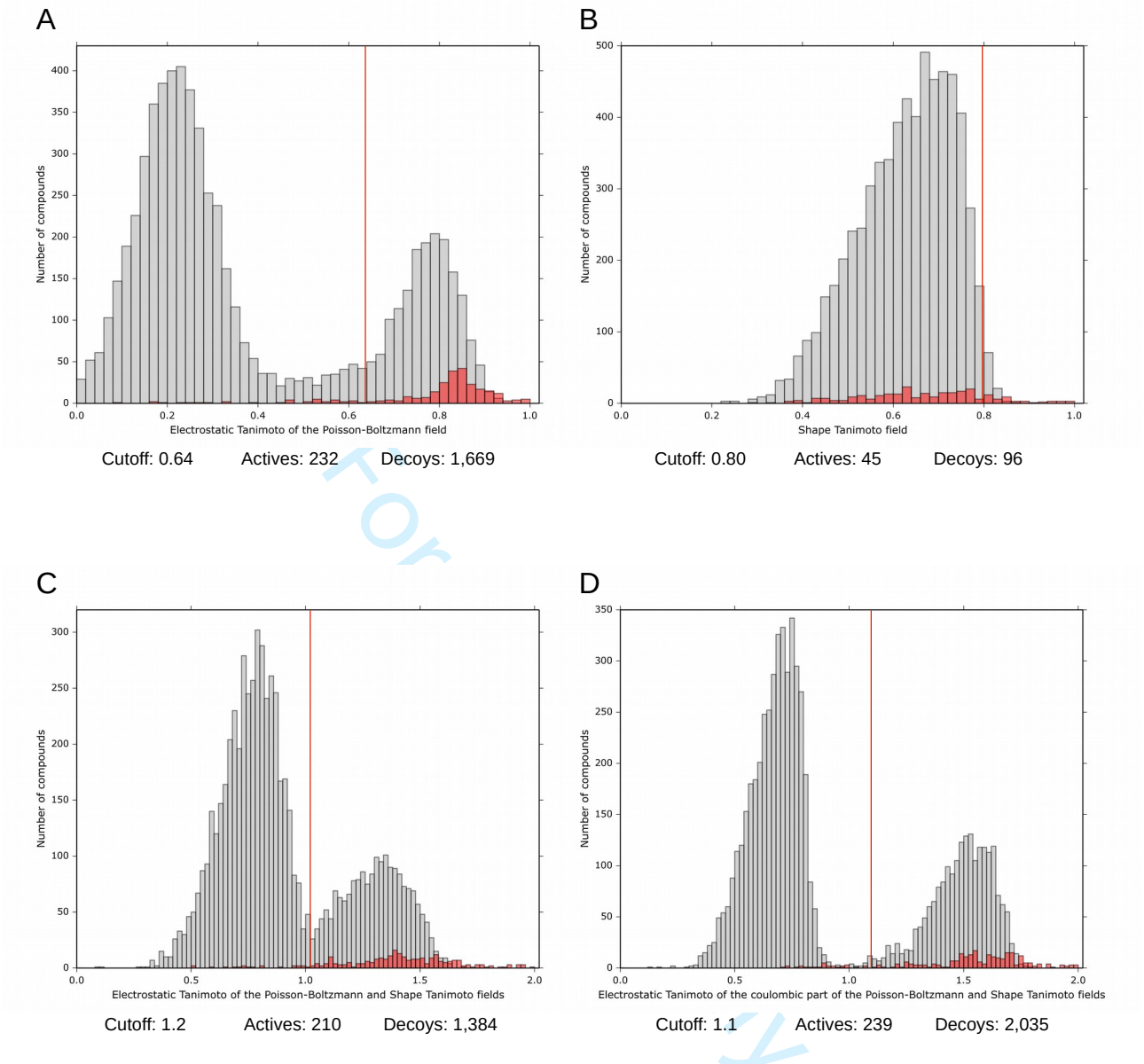


Figure S2. Histograms showing the distribution of the highest Tanimoto values for actives (shown in red) and decoys (shown in gray) for **(A)** the Electrostatic Tanimoto of the Poisson-Boltzmann field; **(B)** the Shape Tanimoto field; **(C)** the Electrostatic Tanimoto of the Poisson-Boltzmann and Shape Tanimoto fields and **(D)** the Electrostatic Tanimoto of the coulombic part of the Poisson-Boltzmann and Shape Tanimoto fields. Different cutoffs (red line) were applied to the set of actives and decoys by using these EON parameters in order to increase the enrichment factor of the VS validation.

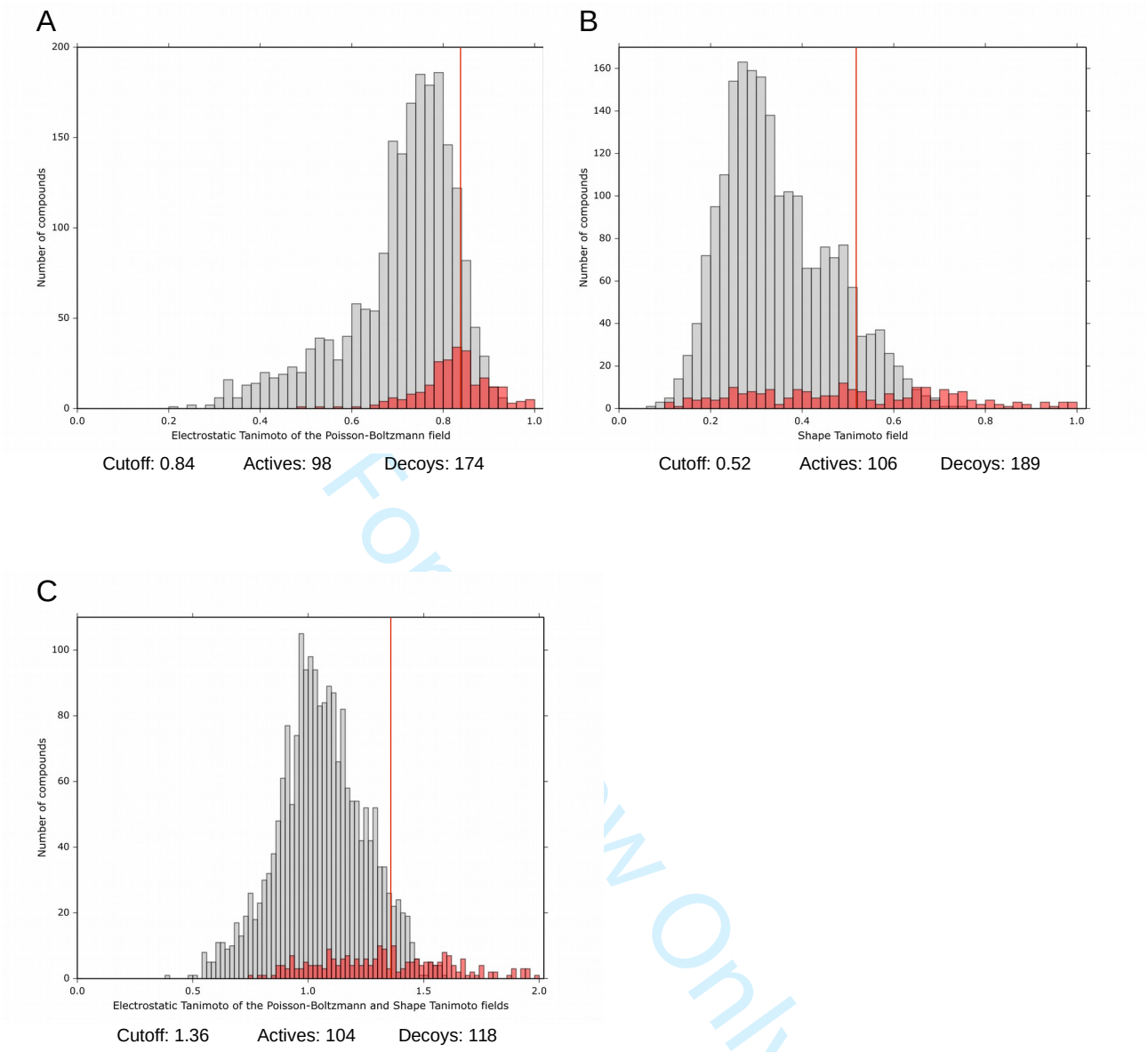


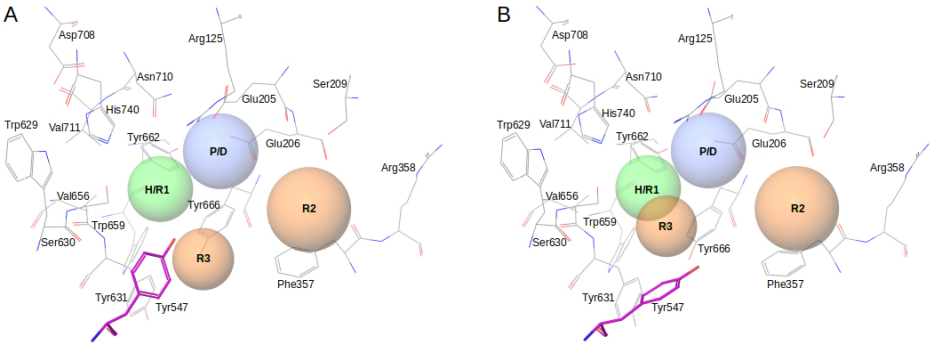
Figure S3. Histograms showing the distribution of the highest Tanimoto values for the pose of actives (shown in red) and decoys (shown in gray) that are above the first cutoff of 0.7 for the Electrostatic Tanimoto of the coulombic part of the Poisson-Boltzmann (see Figure 3A). The panels represent **(A)** the Electrostatic Tanimoto of the Poisson-Boltzmann field; **(B)** the Shape Tanimoto field and **(C)** the Electrostatic Tanimoto of the Poisson-Boltzmann and Shape Tanimoto fields. Different cutoffs (red line) were applied for the set of actives and decoys by using these EON parameters in order to increase the enrichment factor of the VS validation.

Figure 1

Workflow steps	Validation Actives	Decoys	EF	Virtual Screening ZINC Database
Starting Database	419	15,084		16,538,200
ADME QikProp v4.5	---	---		9,362,907
Fingerprint RDKit Torsion (Bottom 1%)	---	---		93,629
SP Docking Glide v6.8 **Pharmacophore** Phase v4.4	267	6,363	1.5	24,034
Similarity & Electrostatic analysis EON v2.2.0.5	101	137	10.5	404
***In vitro* assay**				5

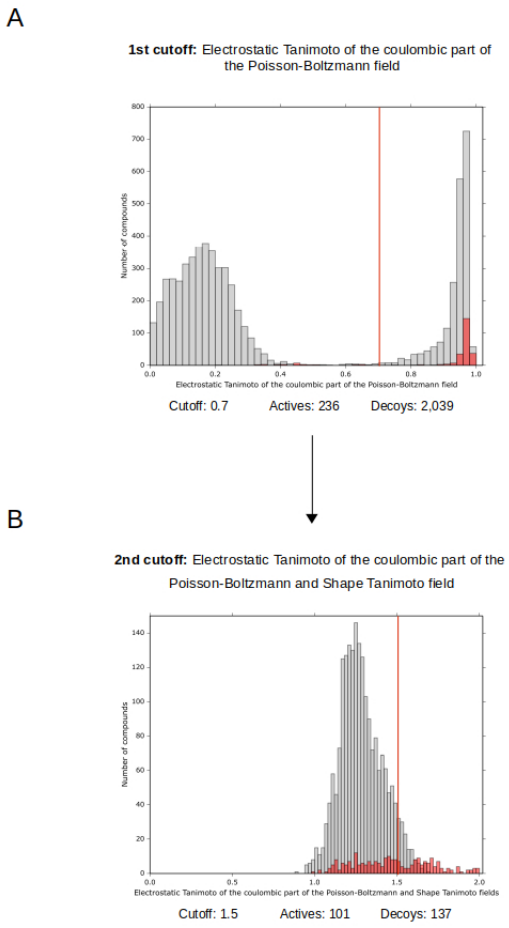
The VS workflow used in the present study. The data corresponds to the number of molecules that remains after each VS step. The actives and decoys columns correspond to those molecules used for validating the VS. The ZINC column refers to data obtained when looking for new leads for DPP-IV inhibition. Enrichment factors were calculated during the validation for each step of the VS protocol as the quotient between the fraction of actives in the sample that survived the VS step and the fraction of actives in the sample before the VS step.

Figure 2



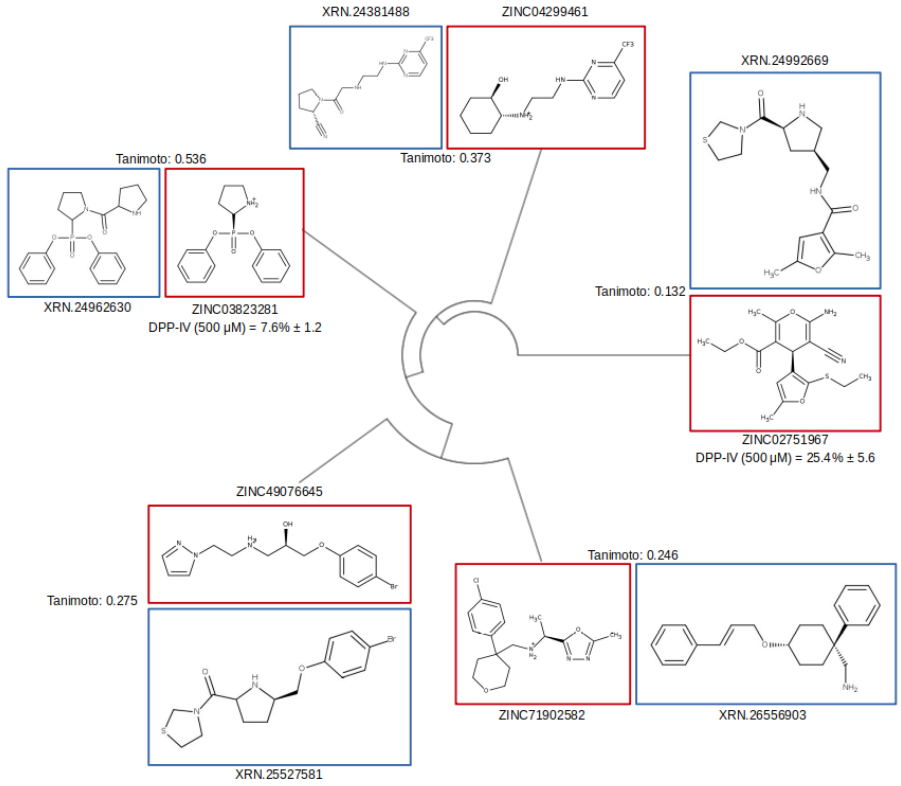
Structure-based pharmacophores used in this paper based on the crystal protein-ligand complex for the most important interactions. The difference between the two pharmacophores is due to the two different conformations of the residue Tyr547 (colored in pink) shown in the context of (A) the 1X70 active site and (B) the 3G0B active site. The pharmacophores are formed by a positive/hydrogen-bond donor feature (i.e., P/D), a hydrophobic/aromatic ring site (i.e., H/R1) and two aromatic ring sites (i.e., R2 and R3). The associated tolerances are 2.3Å for P/D, 2.0Å for H/R1, 2.5Å for R2 and 1.8Å for R3. Two sites (i.e., P/D and H/R1) together with a third site of the two remaining (i.e., R2 and R3) are required during the pharmacophore-based searches.

Figure 3



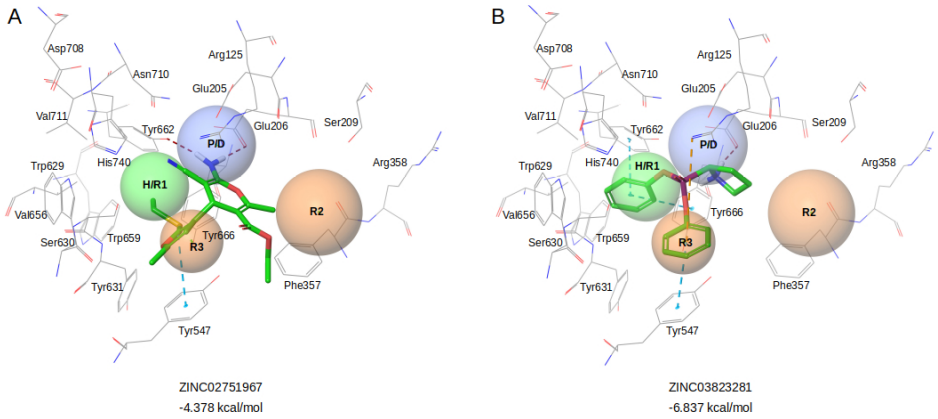
Histograms showing the distribution of the highest Tanimoto values for actives (shown in red) and decoys (shown in gray) for (A) the Electrostatic Tanimoto of the coulombic part of the Poisson-Boltzmann field and (B) the Electrostatic Tanimoto of the coulombic part of the Poisson-Boltzmann and Shape Tanimoto fields. Two consecutive cutoffs (red line) were applied to the set of actives and decoys by using these EON parameters in order to increase the enrichment factor of the VS validation.

Figure 4



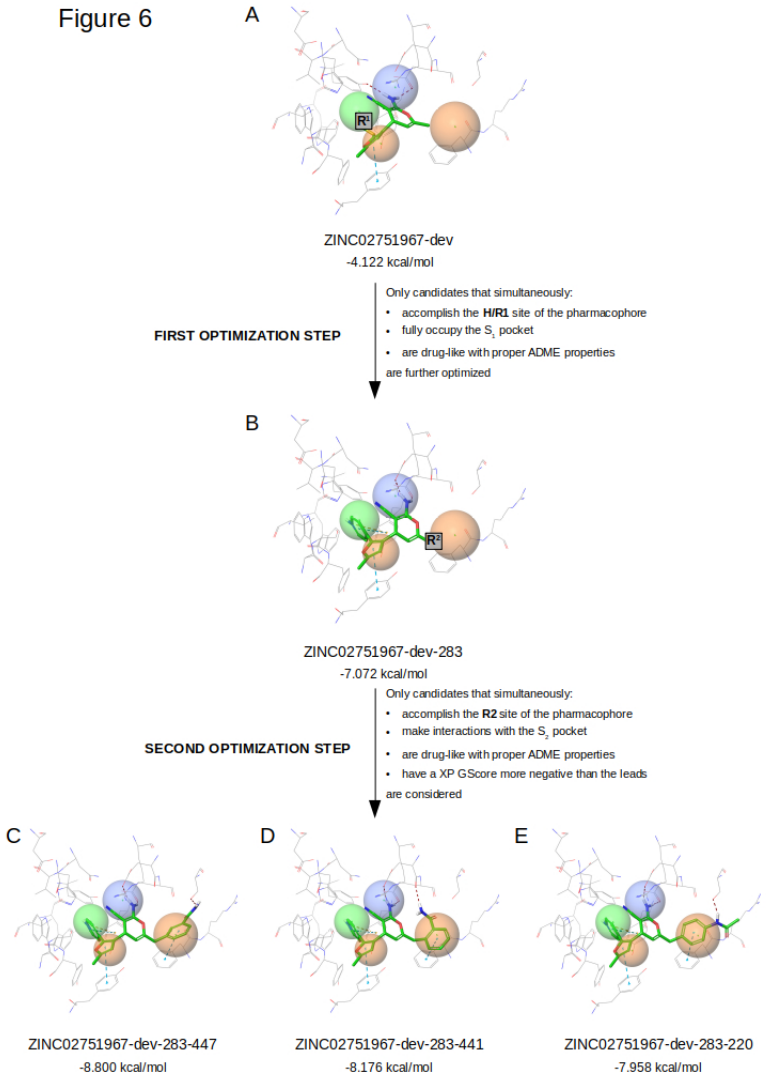
Dendrogram of the five hits selected for experimental testing as a result of the VS workflow (framed in red). The dendrogram shows the distances of the Tanimoto coefficient which represents the fingerprint similarity of the hits. Each hit is attached to a chemical structure downloaded from the Reaxys database which has experimental bioactivity values for human DPP-IV (framed in blue). This molecule is the most similar in terms of fingerprint similarity to the VS hit. Compounds ZINC02751967 and ZINC03823281 are the only ones which show significant in vitro DPP-IV inhibition.

Figure 5



The best docked poses (with the corresponding XP GScore) for the compounds ZINC02751967 and ZINC03823281. Blue and orange dashed lines show π-π stacking and cation-π intermolecular interactions, respectively, whereas the red ones show either salt bridges (between the positively charged amine from ZINC03823281 and Glu206) or hydrogen bonds. Both panels are oriented the same way for easy comparison.

Figure 6

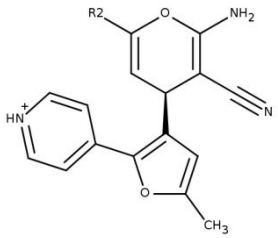


Lead optimization of a derivative of ZINC02751967 used with the aim of obtaining new molecules with improved potency and selectivity for DPP-IV. First, the ethoxycarbonyl group was removed from the initial ZINC02751967 (Figure 6A) because of its low contribution to the protein-ligand interaction (see Figure 5A). Then a substituent was attached to the ethylsulfanyl group of this ZINC02751967 derivative (i.e., R1 label) in order to improve the occupancy of S1 pocket. The resulting derivative (Figure 6B) was selected for further optimization. Next, another point for attaching the substituents (i.e., R2 label) was placed in these new derivatives in order to reach the S2 extensive subsite. The docked poses of some of the most potent derivatives after this second optimization step are shown (Figure 6C-6E). The name for each derivative was built by adding the code of the attached fragment (according to the CombiGlide Diverse Side-chain Collection) to the lead name (see also in the Table 2 the 2D structure and XP GScore for the best eight derivatives obtained during the second optimization step). Blue and orange dashed lines show π-π stacking and cation-π intermolecular interactions, respectively, whereas the red ones show the hydrogen bonds.

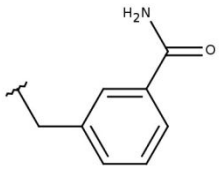
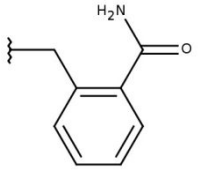
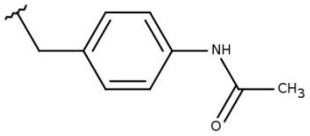
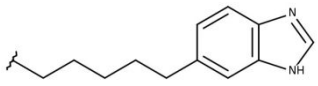
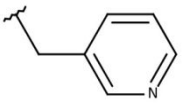
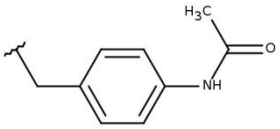
Table 1

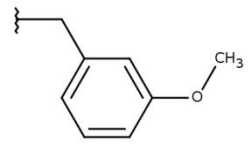
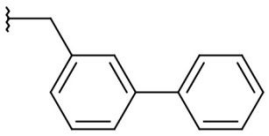
PDB codes										
1X70	2FJP	2IIT	2IIV	2OGZ	2OLE	2ONC	2OPH	2P8S	2QJR	2QOE
2QT9	2QTB	2RGU	2RIP	3C43	3C45	3D4L	3F8S	3G0B	3G0D	3G0G
3H0C	3HAB	3HAC	3KWF	3KWJ	3O95	3VJL	3VJM	4A5S	4G1F	4J3J

Table 2.



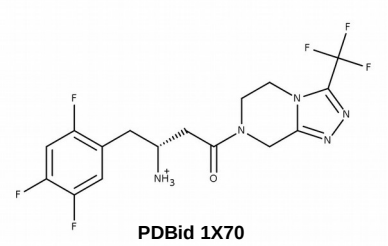
ZINC02751967-dev-283

ZINC02751967-dev	R2 substituent	XP GScore (Kcal/mol)
283-447		-8.800
283-441		-8.176
283-220		-7.958
283-312		-7.706
283-278		-7.651
283-500		-7.578

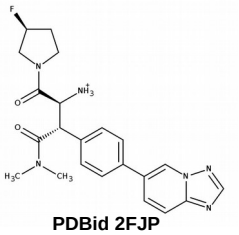
283-386		-7.418
283-236		-7.319

Supplementary Material

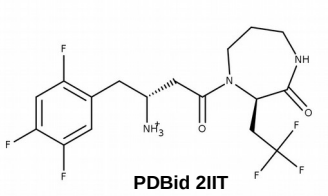
Mining large databases to find new leads with low similarity to known actives: application to find new DPP-IV inhibitors



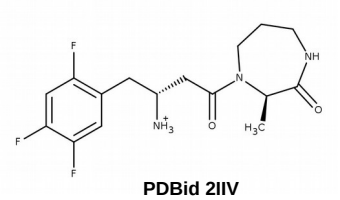
PDBid 1X70
IC_{50} = 18 nM



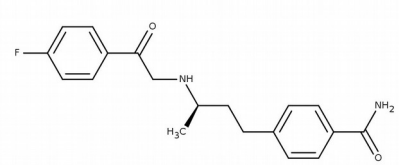
PDBid 2FJP
IC_{50} = 4.3 nM



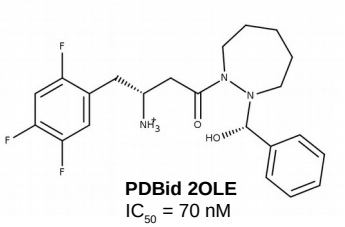
PDBid 2IIT
IC_{50} = 2.6 nM



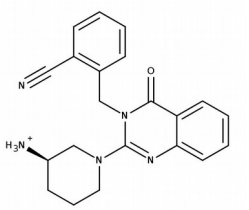
PDBid 2IIV
IC_{50} = 6.6 nM



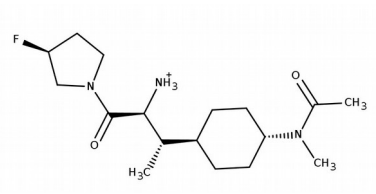
PDBid 2OGZ
IC_{50} = 84 nM



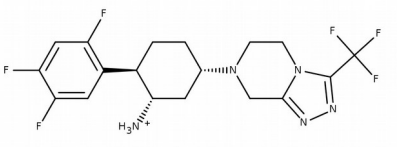
PDBid 2OLE
IC_{50} = 70 nM



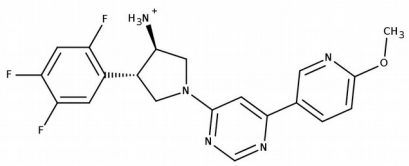
PDBid 2ONC
IC_{50} = 13 nM



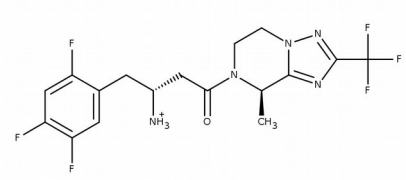
PDBid 2OPH
IC_{50} = 16 nM



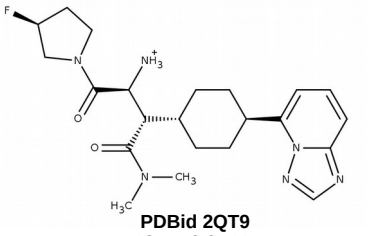
PDBid 2P8S
IC_{50} = 21 nM



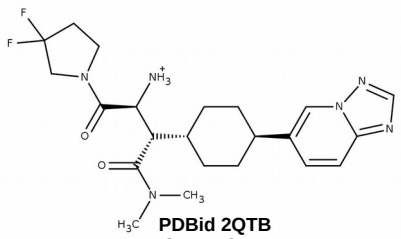
PDBid 2QJR
IC_{50} = 6.4 nM



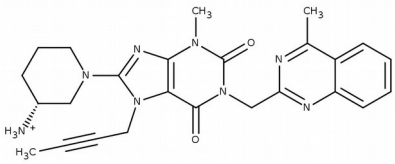
PDBid 2QOE
IC_{50} = 25 nM



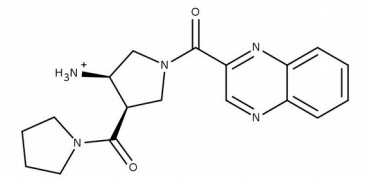
PDBid 2QT9
IC_{50} = 2.3 nM



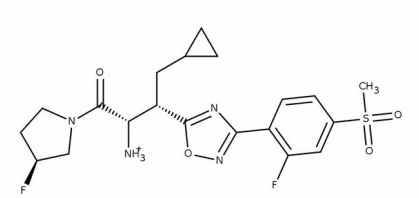
PDBid 2QTB
IC_{50} = 4.8 nM



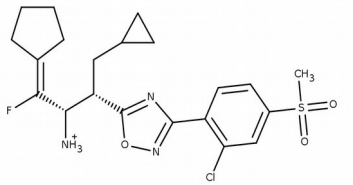
PDBid 2RGU
IC_{50} = 1 nM



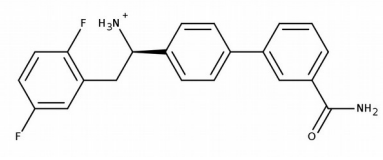
PDBid 2RIP
IC_{50} = 65 nM



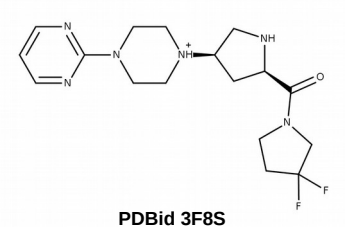
PDBid 3C43
IC_{50} = 19 nM



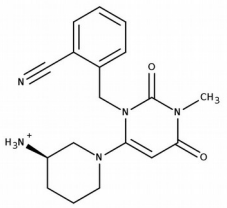
PDBid 3C45
IC_{50} = 0.21 nM



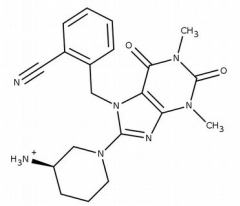
PDBid 3D4L
IC_{50} = 32 nM



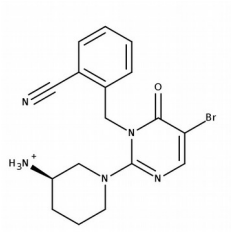
PDBid 3F8S
IC_{50} = 13 nM



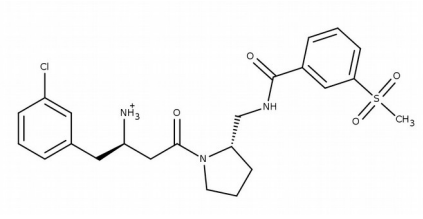
PDBid 3G0B
IC_{50} = 7 nM



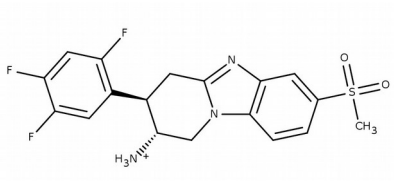
PDBid 3G0D
IC_{50} = 5 nM



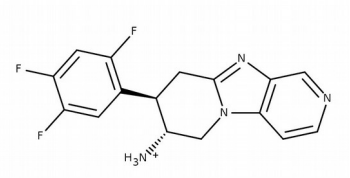
PDBid 3G0G
IC_{50} = 5 nM



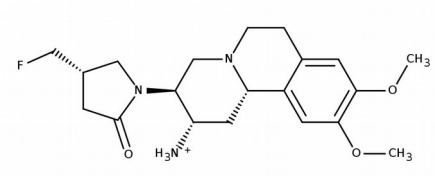
PDBid 3H0C
IC_{50} = 0.38 nM



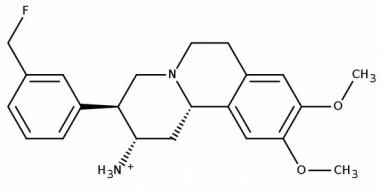
PDBid 3HAB
IC_{50} = 4.2 nM



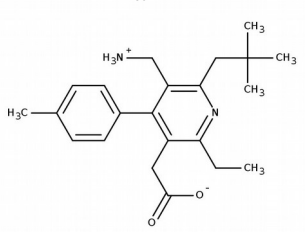
PDBid 3HAC
IC_{50} = 6.7 nM



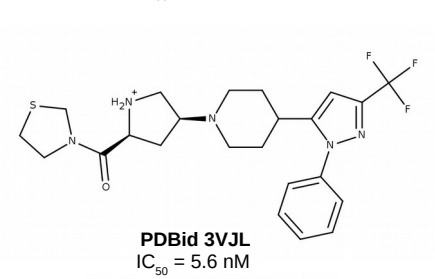
PDBid 3KWF
IC_{50} = 6.8 nM



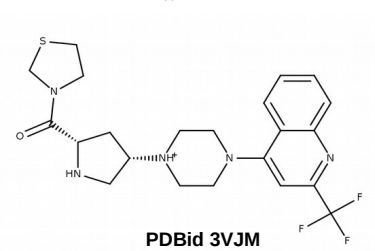
PDBid 3KWJ
IC_{50} = 0.5 nM



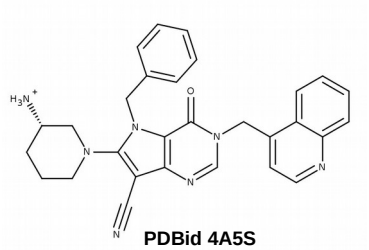
PDBid 3O95
IC_{50} = 5.3 nM



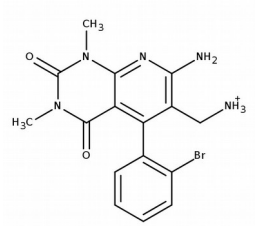
PDBid 3VJL
IC_{50} = 5.6 nM



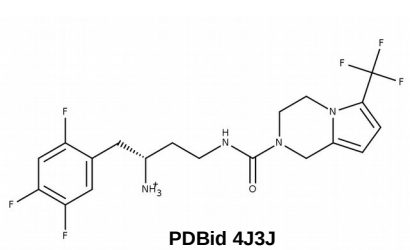
PDBid 3VJM
IC_{50} = 0.37 nM



PDBid 4A5S
IC_{50} = 17 nM



PDBid 4G1F
IC_{50} = 17 nM



PDBid 4J3J
IC_{50} = 31 nM

Figure S1. 2D structure of the inhibitors in PDB complexes with DPP-IV that were used as references to calculate Tanimoto similarity with ZINC molecules based on RDKit-Torsion fingerprint.

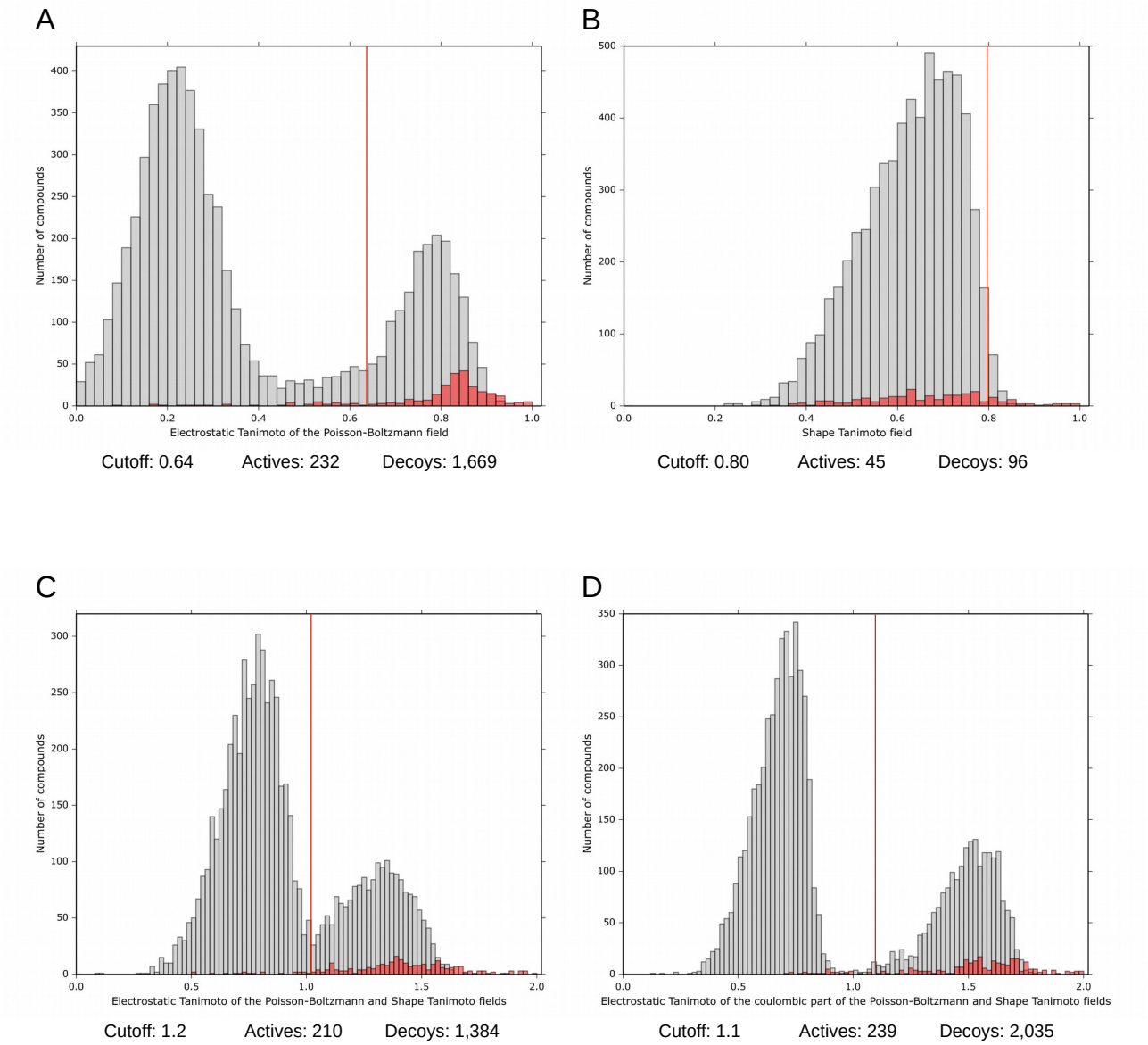
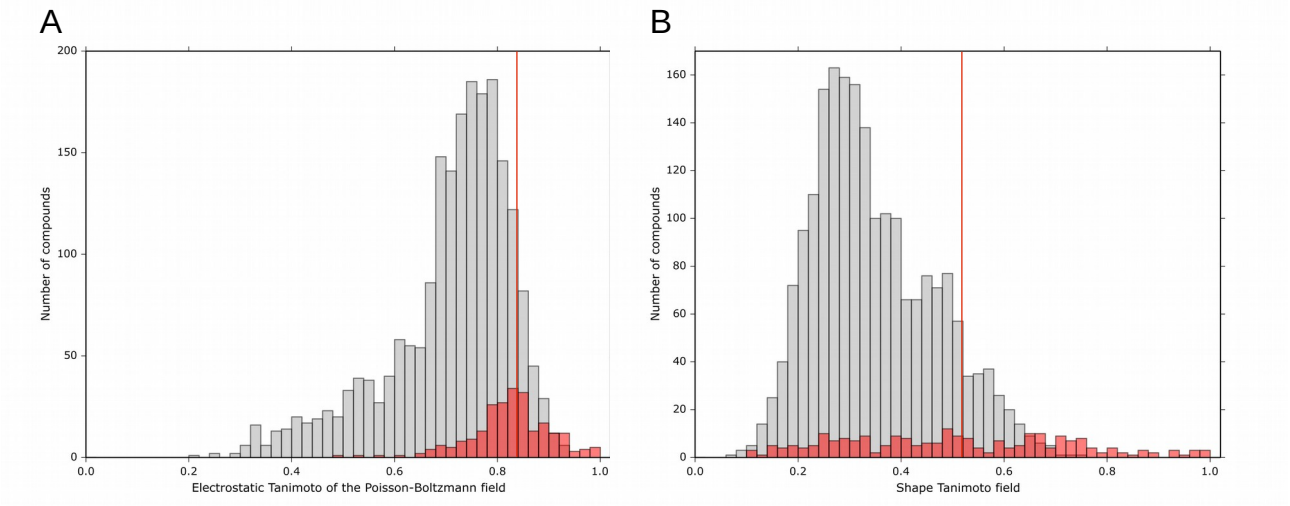


Figure S2. Histograms showing the distribution of the highest Tanimoto values for actives (shown in red) and decoys (shown in gray) for **(A)** the Electrostatic Tanimoto of the Poisson-Boltzmann field; **(B)** the Shape Tanimoto field; **(C)** the Electrostatic Tanimoto of the Poisson-Boltzmann and Shape Tanimoto fields and **(D)** the Electrostatic Tanimoto of the coulombic part of the Poisson-Boltzmann and Shape Tanimoto fields. Different cutoffs (red line) were applied to the set of actives and decoys by using these EON parameters in order to increase the enrichment factor of the VS validation.



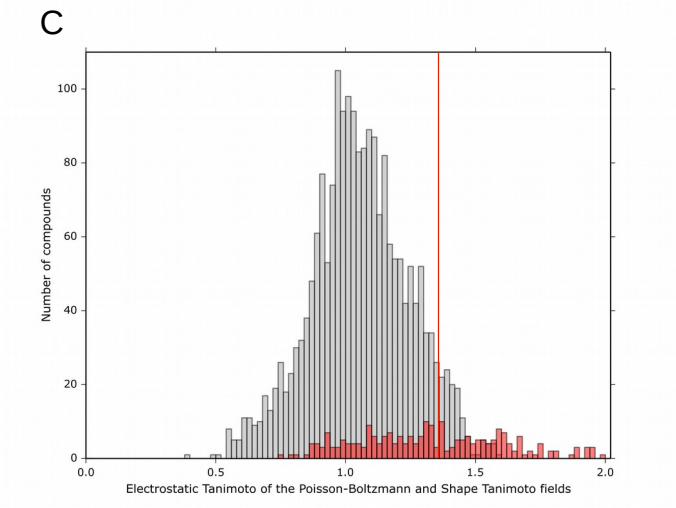


Figure S3. Histograms showing the distribution of the highest Tanimoto values for the pose of actives (shown in red) and decoys (shown in gray) that are above the first cutoff of 0.7 for the Electrostatic Tanimoto of the coulombic part of the Poisson-Boltzmann (see Figure 3A). The panels represent **(A)** the Electrostatic Tanimoto of the Poisson-Boltzmann field; **(B)** the Shape Tanimoto field and **(C)** the Electrostatic Tanimoto of the Poisson-Boltzmann and Shape Tanimoto fields. Different cutoffs (red line) were applied for the set of actives and decoys by using these EON parameters in order to increase the enrichment factor of the VS validation.