Combined ligand- and receptor-based virtual screening methodology to identify structurally diverse PTP1B inhibitors

Aleix Gimeno[a], Andrea Ardid-Ruiz[b], María José Ojeda-Montes[a], Sarah Tomás-Hernández[a], Adrià Cereto-Massagué[a], Raúl Beltrán-Debón[a], Miquel Mulero[a], Cristina Valls[a], Gerard Aragonès[b], Manuel Suárez[b], Gerard Pujadas[a],[c],*, Santiago Garcia-Vallvé[a],[c]

[a]Research group in Cheminformatics & Nutrition, Departament de Bioquímica i Biotecnologia, Universitat Rovira i Virgili, Campus de Sescelades, 43007 Tarragona, Catalonia, Spain

[b]Nutrigenomics Research Group, Department of Biochemistry and Biotechnology, Universitat Rovira i Virgili (URV), Tarragona, Spain

[c]EURECAT, TECNIO, CEICS, Avinguda Universitat, 1, 43204 Reus, Catalonia, Spain

**Correspondence to:* Gerard Pujadas, Research group in Cheminformatics & Nutrition, phone: +34 977 55 95 65, fax: +34 977 55 82 32. Departament de Bioquímica i Biotecnologia, Facultat de Química, Universitat Rovira i Virgili, C/ Marcel·lí Domingo 1, Edifici N4, 43007 Tarragona, Catalonia, Spain. E-mail: gerard.pujadas@urv.cat

Abstract

Protein tyrosine phosphatase 1B (PTP1B) is a potential drug target for diabetes and obesity. However, designing PTP1B inhibitors that combine potency and bioavailability is a great challenge and new leads are needed to circumvent this problem. Virtual screening (VS) workflows can be used to find new PTP1B inhibitors with little chemical similarity to existing ones. Unfortunately, previous VS workflows for the identification of PTP1B inhibitors have several limitations, such as a small number of experimentally tested compounds and the low bioactivity of those compounds. We developed a VS workflow capable of identifying 15 structurally diverse PTP1B inhibitors from 20 compounds whose bioactivity was tested *in vitro*. Moreover, we have identified the two PTP1B inhibitors with the highest bioactivity reported by any VS (i.e. IC_{50} values of 1.4 and 2.1 μM), which can potentially be used as new lead compounds.

1. Introduction

Diabetes mellitus and obesity have become major public health problems in today's society, affecting more than 9%[1] and 13%[2] of the adult population worldwide, respectively. The combination of a sedentary lifestyle and an unbalanced diet promotes the development of obesity because it increases lipid storage and encourages the expansion of adipose tissue and the progressive loss of leptin sensitivity.[3] This triggers the secretion of cytokines, which ultimately results in a systemic inflammation driven by the immune response.[4,5] This inflammation can result in increased insulin resistance,[6,7] decreased insulin secretion by the pancreatic islets[8], and the increased permeability of the vascular endothelium,[3] favoring the development of type 2 diabetes mellitus (T2DM)[9,10] and cardiovascular disease (CVD).[11] Moreover, the higher waist circumference of obese patients is thought to be related to elevated blood pressure and risk of CVD.[12] Elevated waist circumference as a measure of central obesity, elevated blood pressure, atherogenic dyslipidemia (i.e. elevated TAG and lowered HDL cholesterol) and raised fasting glucose levels (as an outcome of increased insulin resistance and decreased insulin secretion) constitute a cluster of cardiovascular risk factors that define metabolic syndrome, which in turn confers a 5-fold increase of risk for T2DM.[13] All the cardiovascular risk factors involved in metabolic syndrome are part of a wide range of diseases that should not be treated separately, but recognized as components of a greater single disease.[14]

Protein tyrosine phosphatase 1B (PTP1B) is a phosphatase whose increased activity and expression are associated with resistance to the hormones insulin[15] and leptin.[16,17] PTP1B acts at several stages of the insulin and leptin signaling pathways.[18] In the insulin signaling pathway, PTP1B dephosphorylates the insulin receptor, thus diminishing insulin action. It also dephosphorylates IRS1/2, inhibiting insulin signal transduction.[19,20] As a result, both the Akt/PKB pathway and the Ras-MAPK pathway are affected. In the leptin signaling pathway, PTP1B dephosphorylates JAK2, altering the JAK-STAT pathway and therefore inhibiting the anorexigenic action of leptin in the hypothalamus, resulting in an increase in food intake and a decrease in energy expenditure.[21,22] Consequently, PTP1B inhibition has emerged as a promising strategy for the treatment of diabetes and obesity.[23,24]

Over the last two decades, different classes of PTP1B inhibitors have been identified through rational drug design.[25] However, achieving a compromise between the activity and pharmacokinetics of PTP1B inhibitors has proved a major challenge in medicinal chemistry due to the characteristics of the PTP1B binding site. The combination of the backbone amide protons of the P-loop (defined by residues 214 to 221 with HCSAGIGR sequence) and the basic nature of the Arg221 side chain create a highly positive-charged environment that has a preference for ligands with highly acidic groups (see Figure 1). VS workflows capable of further exploring the chemical space by identifying new chemical entities with diverse structures may prove useful in the pursuit of molecules that show potential for use as new lead compounds for the design of PTP1B inhibitors with a good balance between potency and bioavailability. To date, several VS workflows have been developed for the purpose of discovering novel PTP1B inhibitors.[26-33] Despite their success in identifying PTP1B inhibitors, they often present drawbacks such as few[31-33] or no compounds[26,27,29] tested *in vitro,* low percentages of active compounds[28,30] and relatively low activity values of the identified inhibitors[28,30,33] (see Table 1).

Therefore, the aim of this study was to design a new VS methodology and validate it by means of the *in vitro* testing of a structurally diverse set of compounds, which addresses the need to find new PTP1B inhibitors with increased potency that can be used as lead compounds in the search for new PTP1B inhibitors.

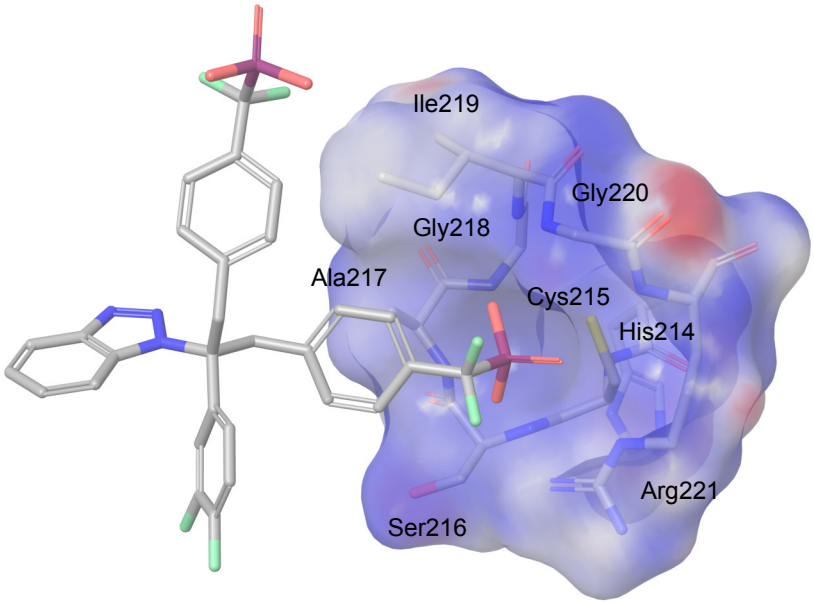


Figure 1. Representation of the crystal structure with PDB[34,35] code 1Q6M[36] showing the P-loop residues of PTP1B and the co-crystallized ligand. The ligand and the P-loop residues are shown in sticks and the molecular surface of the P-loop residues has been represented and colored according to their Poisson-Boltzmann electrostatic potentials from red (negative) to blue (positive), where potentials range from -50.0 to 50.0. This figure was obtained with Maestro[37] v10.7.

Table 1. Summary of the manuscripts that have been published containing VSs for PTP1B inhibitors. The first column contains the authors and the reference to the manuscript; the second column contains the number of compounds tested *in vitro*; the third column contains the number of PTP1B inhibitors identified by the *in vitro* assay; the fourth column contains the number of PTP1B inhibitors identified whose activity was found in the 1-10 μM range.

VS	Compounds tested *in vitro*	Compounds with PTP1B activity	Compounds with activity in the 1-10 μM range
Floriano *et al.*[25]	0	-	-
Bharatham *et al.*[26]	0	-	-
Park *et al.*[27]	225	62	0
Rao *et al.*[28]	0	-	-
Reddy *et al.*[29]	43	4	0
Ma *et al.*[30]	8	8	2
Balaramnavar *et al.*[31]	10	10	2
Chandra *et al.*[32]	5	5	0
Present study	20	15	2

2. Results and Discussion

We have developed a VS workflow capable of identifying PTP1B inhibitors from a given library of compounds. The workflow consists of a series of successive filters, namely a molecular weight (MW) filter, a random forest (RF) model based on fingerprints, protein-ligand docking and an electrostatic similarity analysis, in which the output molecules of each filter were the input of the next one (i.e. those molecules that do not accomplish one filter are eliminated for subsequent filters; see Figure 2). The performances of the RF model and the electrostatic similarity analysis were validated *in silico* with two different sets of actives and decoys. Finally, a library of 212,672 compounds obtained from the Specs[38] company was screened for the purpose of validating the VS protocol *in vitro* by finding new PTP1B inhibitors with previously undescribed scaffolds for such bioactivity.

2.1. MW filter

In the first VS step, the Specs library was filtered according to MW in order to discard molecules that were considered either too small or too large to fit on the PTP1B binding site. In order to determine the specific molecular weight range, all inhibitors of human PTP1B with an average pX bioactivity greater than or equal to 4 were obtained from Reaxys[39] (a total of 6,091 molecules) and the distribution of their MW was plotted (see Figure S1). Using the MW of these known actives as a reference, molecules with a MW lower than 200 Da and higher than 900 Da were removed from the library, leaving 65,223 remaining compounds and discarding about 70% of the initial compounds (see Figure 2). Although no validation was performed for this step, the rapid exclusion of compounds from the scope of the VS justifies the use of this filter at the beginning of the workflow.

Workflow steps	Number of compounds
Starting library	212,672
MW filter	65,223
Random Forests	30,838
Docking	11,429
Electrostatic similarity analysis	717
Activity assay	20

Figure 2. Diagram of the virtual screening indicating the different filters used and the number of compounds that overcome each of them.

2.2. RF model

Next, an RF model based on circular fingerprints (FPs) was developed in order to rapidly rule out the compounds least likely to be active. Circular FPs (also known as Morgan FPs because they are based on the Morgan algorithm) record the environment of each atom in the molecule up to a determined radius.[40] As hashed topological fingerprints, they have the advantage of being built around each molecule; therefore, any molecule can produce a meaningful FP. Despite the drawback that FP bits do not relate to the presence or absence of a particular substructure, since the RF model from these FPs was developed for the specific purpose of enriching the library in actives, circular FPs have been used to develop the model as they are one of the highest ranked FPs in terms of FP performance.[40]

As a supervised machine learning algorithm based on fingerprints, in this RF model, FP bits are related to bioactivity and the output probabilities are a function of the presence of important structural characteristics for the bioactivity of known actives. Therefore, the model should be capable of identifying molecules with similar structural characteristics to known actives, but with different overall structures, recognizing compounds that present crucial features for activity as compounds with a high probability of being active and vice versa. Because the RF model is not the last step of the VS workflow (see Figure 2), compounds presenting a probability of being active of less than 20% were excluded. For this threshold, in a cross-validation using actives and decoys, the RF model performed with average precision and sensitivity values of 94.1% and 98.7% (see Table 2 for performance details). Despite this relatively low threshold, about 50% of the compounds from the Specs library that passed the MW filter were discarded and 30,838 compounds remained (see Figure 2). Therefore, the development of this predictive model allowed us to quickly dismiss a large number of molecules with low probabilities of inhibiting PTP1B due to the low computational cost of FP calculation and machine learning, thus proving a good filtering option for application at early stages of the VS workflow.

Table 2. Statistical parameters of the RF model cross-validation. The values correspond to the means of the 5-fold cross-validation for each parameter.

	Sensitivity	Specificity	Precision	Fall-out	False Negative Rate	False Discovery Rate	Accuracy	F1 Score	Matthews Correlation Coefficient
5-fold cross-validation mean	0.9875	0.9379	0.9408	0.0621	0.0125	0.0592	0.9627	0.9636	0.9265

2.3. Protein-ligand docking

The compounds resulting from the RF filter were docked onto the binding site of PTP1B using the protein from the crystal structure with PDB[34,35] code 1Q6M.[36] It has previously been reported that the use of pharmacophoric constraints during docking (i.e. a polar interaction with Arg221 and at least one interaction with the backbone amides of the residues of the P-loop) contributes to an increase in VS enrichment when

searching for PTP1B inhibitors.[41] Thus, with this idea in mind, the ligand interactions with the P-loop residues were analyzed for the crystal structures containing the most potent PTP1B inhibitors (see Table 3). In this analysis, a series of hydrogen bond interactions were identified to be common among most of these inhibitors and, therefore, the protein atoms involved in these interactions were used to set hydrogen bond constraints during the protein-ligand docking. A total of 6 protein residues were used to define these hydrogen bond constraints (see the *Grid generation* section for more details). Since the next step of the VS is a 3D-based similarity analysis with actives, hydrogen bond constraints are also helpful in orienting each molecule at the binding site and dismissing the molecules that are unable to form hydrogen bond interactions with the P-loop. These molecules are not of interest to us because their binding mode would be different to that of the molecules used as a reference in the subsequent electrostatic similarity analysis.

Table 3. Summary of the interactions between crystallized PTP1B inhibitors with an IC_{50} or K_i lower than or equal to 100 nM and their target. Data was obtained from the LigPlot[42] and PoseView[43] diagrams in the PDBsum[44] and PDB[34,35] databases. Non-bonded contacts are indicated with check marks and hydrogen bonds are indicated with the PDB[34,35] label of the protein atom that is involved in the interaction. The crystallized ligands of these protein-ligand complexes were used as a reference to perform electrostatic comparisons. The ligands of all 14 protein-ligand complexes were used to perform 14 separate validations (see Figure S2), but only the ligands of the 11 underlined protein-ligand complexes were selected as queries for the electrostatic similarity analysis.

	1Q6J	1Q6M	1Q6N	1Q6P	1Q6S	1Q6T	2FJN	2QBP	2QBQ	2VEU	2VEW	2VEY	2ZMM	2ZN7
His214	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Cys215	SG	✓	SG	SG	SG	SG	SG	✓	✓	✓, SG	✓	✓	✓	✓
Ser216	N	N	N	N	N	N	N	✓	✓	N	N	N	✓	✓
Ala217	✓, N	✓, N	✓, N	✓, N	✓, N	✓, N	✓, N	✓	✓	✓, N	✓, N	✓, N	✓	✓
Gly218	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Ile219	✓, N	✓, N	✓, N	✓, N	✓, N	✓, N	✓, N	-	-	✓, N	N	✓, N	✓	✓
Gly220	N	N	N	N	N	N	N	-	✓	N	N	N	✓	✓
Arg221	N, NE, NH2	N, NE, NH2	N, NE, NH2	N, NE, NH2	N, NE, NH1	N, NE, NH2	N, NE, NH2	N, NE	N, NE	N, NE, NH2	N, NE	N, NE, NH2	N, NE	N, NE

Although this docking step allows us to discard compounds that would not fit in the binding site as well as those that would not be capable of establishing hydrogen bonds with the P-loop residues, it is not considered a filter *per se,* included to increase enrichment in actives, but instead it is considered a means of generating hypotheses on how each compound may bind to the binding site. Therefore, no validation was performed for this step of the VS. After applying this filter to the remaining compounds from the Specs library, 114,823 docked poses corresponding to 11,429 compounds survived this step (see Figure 2).

2.4. Electrostatic similarity analysis

Once the docked poses were obtained, an electrostatic similarity analysis was performed in order to compare the electrostatic potential of the 114,823 docked poses for the remaining 11,429 molecules to those of the experimental poses for potent PTP1B inhibitors (see Figure 2). Thus, 11 crystallized complexes

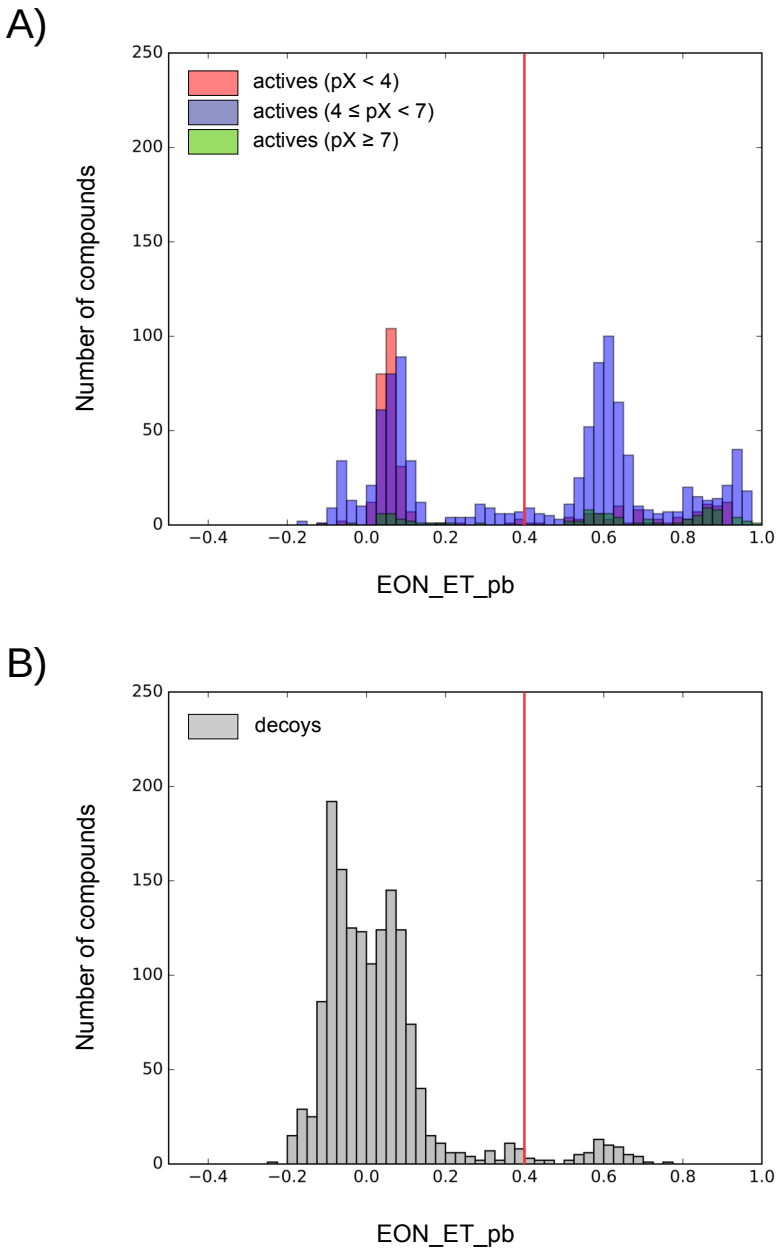


Figure 3. Histogram representation of the highest electrostatic Tanimoto (i.e. EON_ET_pb) values obtained in the comparison of the validation set to all queries. Panel A shows the histogram for the actives in the validation set. Actives with a pX lower than 4 are in red, actives with a pX from 4 to 7 are in blue and actives with a pX higher than 7 are in green. Panel B shows the histogram for the decoys in the validation set. The EON_ET_pb cutoff of 0.4 is represented as a red line.

between PTP1B and PTP1B inhibitors with IC_{50} or K_i activity values between 1 and 100 nM were superposed to the crystal structure with PDB[34,35] code 1Q6M[36] (the same PDB[34,35] structure that was used during the protein-ligand docking step of the VS) and used as a reference for the electrostatic comparison (see Table 3). Comparing the superposed experimental poses directly with the 114,823 docked poses of the 11,429 compounds allows us to overcome the limitations of docking scoring functions by selecting the most reliable docking pose for each compound based on its similarity to the known binding mode of an active compound.

To validate the use of these co-crystallized PTP1B inhibitors as reference compounds for the electrostatic potential comparison *in silico*, a validation set of 1,424 actives and 1,500 decoys was prepared and docked to 1Q6M[36] in the same conditions as the library compounds and compared with the set of the 11 experimental poses (see Table 3) in order to obtain the electrostatic Tanimoto (i.e. EON_ET_pb) coefficients of the docked poses for actives and decoys from the validation set for each comparison. Once the EON_ET_pb values were computed for each molecule, only the docked pose that presented the highest EON_ET_pb value regardless of the reference compound used for the comparison was kept. After discarding all other docking poses, histograms of the electrostatic Tanimoto coefficients for the actives and the decoys in the validation set were plotted, separating the actives into 3 groups depending on whether their bioactivity had a pX value lower than 4, between 4 and 7 or higher than 7 (see Figure 3). As Figure 3 shows, applying an electrostatic Tanimoto (i.e. EON_ET_pb) cutoff of 0.4 to the validation set allowed us to discard 1,436 decoys (96% of the initial decoys) and retain 739 actives (52% of the initial actives). Therefore, this cutoff was applied to our compound library after performing the same electrostatic potential comparison in order to enrich it with active molecules, leaving 717 remaining compounds (see Figure 2).

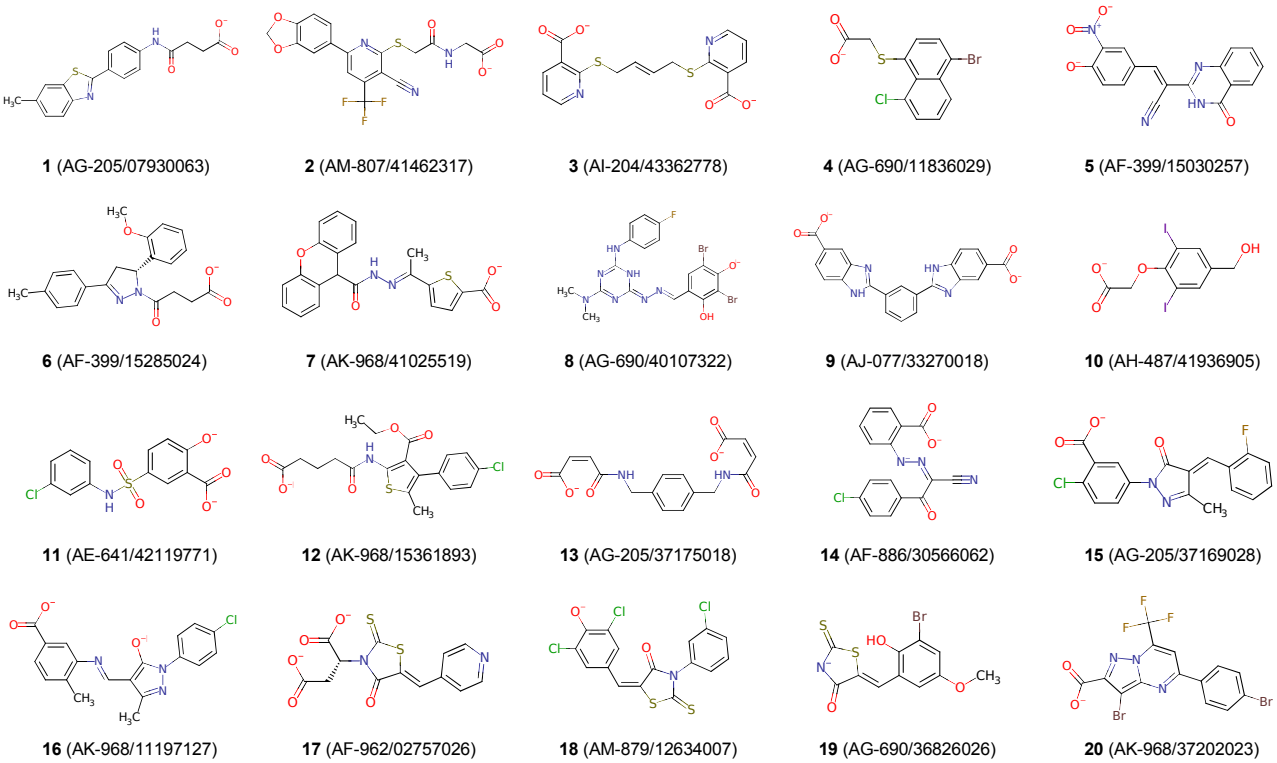


Figure 4. 2D structures of the 20 hit compounds. Each compound is identified with its Specs ID number. MarvinSketch[45] was used to draw the structures. The protonation state of each compound corresponds to the protonation state of the docked pose selected for that compound.

2.5. Clustering and hit selection

After having computed their EON_ET_pb similarities with experimental poses and applied the cutoff, the resulting 717 hit molecules were clustered using the HDBSCAN[46] algorithm together with the 1,424 active compounds used in the validation of the electrostatic similarity filter. Thus, hit compounds that were grouped in the same cluster as known active compounds were excluded. Clusters that contained only hit molecules were identified and the molecule with the highest EON_ET_pb value for one of the 11 experimental poses was selected from each cluster. The selected molecules were sorted according to electrostatic similarity and were visually inspected to examine the diversity of their structures. Finally, the top 20 molecules were selected for activity tests (see Figure 4). Figure 5 shows their structural diversity in a fingerprint-based dendogram. This allowed us to identify the molecules that had similar electrostatic potential to active compounds, but with a different structure. These molecules may present similar chemical properties to those of active compounds and may thus be able to establish similar interactions with the protein environment, despite their low structural similarity to known PTP1B inhibitors (see Figure S3).

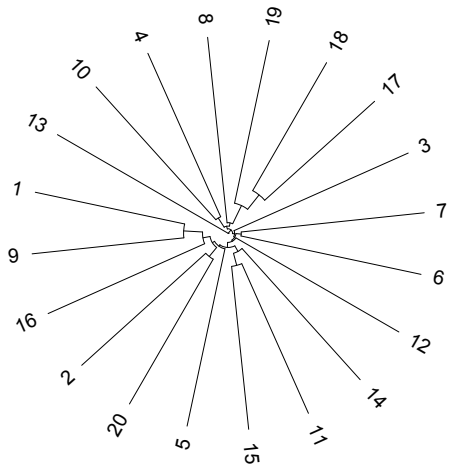


Figure 5. *Dendogram based on fingerprints showing the structural diversity of the 20 hit compounds. The FP used to obtain the distance matrix was the OpenEyePath FP. iTOL[47] was used to draw the dendogram.*

2.6. Activity assays and comparison with other VS looking for PTP1B-inhibitors

The 20 hit compounds were purchased from the Specs[38] company. Figure S4 shows the LC-UV/Vis and Q-TOF spectra and each compound. This data was obtained at the Center for Omic Sciences (COS, Reus, Catalonia) and shows that all compounds have a purity above 95.0% (see Table S1) except for compounds **13** (87.96%) and **17** (93.56%). The PTP1B inhibitory activity of these 20 hit compounds was analyzed *in vitro* at concentrations of 100 μM and 10 μM (see Table 4). Fifteen of the 20 hit compounds (**2**, **3**, **4**, **5**, **6**, **7**, **9**, **10**, **12**, **13**, **14**, **15**, **18**, **19** and **20**) showed PTP1B inhibition at a concentration of 100 μM. Although compounds **2**, **3**, **4**, **6**, **9**, **10**, **12**, **13**, **14**, **15**, **18**, **19** and **20** displayed reduced or no inhibition at the concentration of 10 μM, interestingly, compounds **5** (with a purity of 97.77%) and **7** (with a purity of 99.64%) displayed complete inhibition of the enzyme at both concentrations and were selected for IC_{50} determination (see Table 4). Compound **8** was not selected because it inhibited the enzyme at the concentration of 10 μM, but not at the

concentration of 100 μM. The IC_{50} values of compound **5** and compound **7** were 1.4 and 2.1 μM, respectively (see Table 4).

To our knowledge, these compounds represent the highest inhibitory potency for PTP1B obtained by any VS method. Moreover, although previous VS protocols[31,32] have also obtained PTP1B inhibitors with activity values in the low μM range, they were similar in structure. For example, Ma *et al.*[31] reported two inhibitors (labeled **6** and **4** by the authors) with IC_{50} values of 4.2 and 4.5 μM, respectively, but the compounds were structurally similar to one another and to previously identified PTP1B inhibitors (see Figure S5). Moreover, Balaramnavar *et al.*[32] reported two other inhibitors (labeled **115** and **116** by the authors) with IC_{50} values of 7.5 and 8.3 μM, respectively, that share the same core structure and a different aryl sulfonyl substituent (see Figure S6). In contrast, the PTP1B inhibitors obtained in this work are more potent than those found by Ma *et al.*[31] or Balaramnavar *et al.*,[32] and they are structurally diverse and different from previously described PTP1B inhibitors (see Figure 5). Remarkably, 15 of the 20 compounds tested show bioactivity as PTP1B inhibitors. Therefore, this method offers unprecedented performance when using VS for exploring the chemical space in pursuit of new scaffolds with PTP1B inhibitory activity.

Table 4. Results of the activity assays performed on the 20 hit compounds showing the percentages of inhibition at the substrate concentrations of 100 μM and 10 μM and the IC_{50} values of compounds **5** and **7** in μM. ND refers to "not determined". [a] Negative inhibition values were considered as 0% inhibition. [b] Inhibition values higher than 100% were considered as 100% inhibition. [c] Suramin was used as a positive control. The experimental data used to calculate the IC_{50} values in this table can be found in Figure S7.

Compound	PTP1B inhibition (%)		PTP1B IC_{50} (μM)
	100 μM	10 μM	
1	0[a]	0[a]	ND
2	12	0	ND
3	27	0[a]	ND
4	44	0[a]	ND
5	100[b]	100[b]	1.4
6	5	0[a]	ND
7	100[b]	100[b]	2.1
8	0[a]	77	ND
9	29	4	ND
10	72	11	ND
11	0[a]	0[a]	ND
12	37	0[a]	ND
13	27	3	ND
14	89	0[a]	ND
15	26	0[a]	ND
16	0[a]	0[a]	ND
17	0[a]	0[a]	ND
18	16	0[a]	ND
19	57	42	ND
20	40	0[a]	ND
Suramin[c]	90	93	ND

2.7. Performance of our VS workflow on compounds identified by other VSs

We have analyzed how our VS workflow acts on the hit compounds identified by previous PTP1B VSs (see Table 1). At this point, it is worth remarking that although the bioactivity of a total of 291 compounds was assayed in these studies, no data is available on the 2D structure for: **(a)** the false positive compounds; and **(b)** 53 out of the 62 actives identified by Park *et al*.[28] Thus, we can only describe the performance for 36 PTP1B inhibitors from Table 1:

- Park *et al*.[28]: out of the 9 compounds with reported PTP1B inhibitory activity available from this study, 5 of them were identified as hits by our VS methodology (*i.e.*, compounds **1**, **3**, **4**, **8** and **9**). Compounds **5** and **6** from this study were discarded in the protein-ligand docking step of our VS workflow because no docked poses were able to fulfill the hydrogen bond constraints specified in the protein-ligand docking step. This could be explained by the fact that hydrogen bond acceptors in both compounds are surrounded by bulky groups which may prevent them from interacting with the hydrogen bond donors of the P-loop. Compounds **2** and **7**, on the other hand, fulfilled the docking constraints, but their docked poses showed low electrostatic similarity with the co-crystallized ligands used as queries and, therefore, they were discarded.

- Reddy *et al.*[30]: out of the 4 compounds with reported PTP1B inhibitory activity in this study, 3 of them were identified as hits by our VS methodology (*i.e.*, compounds **AU-247**, **AU-2525** and **AU-2439**). Compound **AU-008** was the only one that the VS was not able to predict as active and it was discarded by the RF model, possibly due to its low structural similarity with known PTP1B inhibitors.

- Ma *et al.*[31]: out of the 8 compounds with reported PTP1B inhibitory activity in this study, only 1 of them was identified as a hit by our VS methodology (*i.e.*, compound **5**). Compounds **1** and **2** were discarded by the RF model and compounds **3**, **4**, **6**, **7** and **8** fulfilled the docking constraints but displayed an electrostatic similarity value lower than the defined cutoff value. As these five compounds present high structural similarity, it is no surprise that all of them are discarded for the same reason. Compound **5** (the only hit from Ma *et al.* that succeed in our VS workflow) is a special case in the sense that, despite its structural similarity with compounds **3**, **4**, **6**, **7** and **8**, it is the only compound in that study that has a -COOH group which is not only able to fulfill the protein-ligand docking constraints but also provides a negative environment close to the P-loop residues that is similar to the one provided by the co-crystallized compounds used as queries during the electrostatic comparison.

- Balaramnavar *et al.*[32]: out of the 10 compounds with reported PTP1B inhibitory activity in this study, only 1 of them was identified as a hit by our VS methodology (*i.e.*, compound **117**). Compounds **116** and **117** reported by this work were the only ones able to fulfill our protein-ligand docking constraints due to the presence of a $-NO_2$ group and a -COOH group, respectively at their variable substructure. Interestingly, the protein-ligand docking results obtained by Balaramnavar *et al.*[32] suggest that all

ten compounds used the hydroxyl group present in the substructure shared by all of them to act as a hydrogen bond acceptor of the protein donors located at the P-loop. In contrast, no pose for any of these 10 compounds in our protein-ligand protocol was able to use this hydroxyl group to make hydrogen bonds with the P-loop residues. Moreover, using our protocol with the same PDB structure that Balaramnavar *et al.*[32] used (*i.e.*, 1AAX[48]) also gave no docked poses with the hydroxyl group accepting protons from the P-loop and only docked poses for compounds **116** and **117** were again obtained. Therefore, such strong differences between protein-ligand docking results could be the result of differences in how the programs used to perform this step consider the steric hindrance between protein binding site and ligand (which would allow certain poses for one program that are forbidden for another one). Finally, only compound **117** surpassed the electrostatic similarity filter. This is surely the consequence of using query ligands at this step with negatively charged groups to interact with the P-loop (as compound **117** does with its -COOH substituent in the docked pose we obtained).

- Chandra *et al.*[33]: out of the 5 compounds with reported PTP1B inhibitory activity in this study, 3 of them were identified as hits by our VS methodology (*i.e.*, compounds **SB01794SC**, **JFD02789** and **JFD03705**). Compounds **BTB12807** and **JFD02644** were discarded in the last step of the workflow as they did not show enough electrostatic similarity with the reference compounds.

Overall, our VS workflow is able to identify 13 out of the 36 compounds identified by other VS methodologies as PTP1B inhibitors. From the 23 discarded molecules, 3 and 10 of them are eliminated by the RF and the electrostatic similarity filters, respectively. Remarkably, the protein-ligand docking step (that was not specifically designed as a filter and whose aim was to orient each molecule at the binding site and dismiss those that were unable to form hydrogen bond interactions with the P-loop) also discarded 10 compounds. However, as mentioned above, this could be related to differences in how the programs consider the steric hindrance between the protein binding site and the corresponding ligand, as 8 of these compounds shared the same substructure that is expected to hydrogen bond to the P-loop .[32]

3. Conclusions

Due to the lack of PTP1B inhibitors presenting both a good activity and pharmacokinetic profile, new candidate compounds presenting PTP1B inhibitory activity through novel structures need to be identified for use as lead compounds. Despite previous efforts, over the last two decades very few PTP1B inhibitors resulting from VS have been reported. Our VS methodology demonstrates the potential of VS workflows because it was able to: **a)** find 15 compounds capable of inhibiting PTP1B at 100 μM (75% of the compounds experimentally tested); **b)** find 2 PTP1B inhibitors with an IC_{50} value in the 1-10 μM range; **c)** find PTP1B inhibitors with structural diversity; and **d)** find PTP1B inhibitors whose structure is not similar to that of known actives. Although further pharmacokinetic studies are required, this VS workflow has discovered several novel compounds capable of inhibiting PTP1B and has shown its potential for identifying more potent

and structurally diverse molecules which could be used as lead compounds for the treatment of diabetes and obesity.

Nevertheless, although a good proportion of the tested compounds have been found to be active in this study, their activity values are not as high as those described in several structure-activity relationship studies.[36,49] Keeping in mind that: **(a)** the percentages of PTP1B inhibition of the 20 hits show poor or no correlation with either the docking score or the electrostatic Tanimoto (see Figure S8); and **(b)** the required protein-ligand interactions were the establishment of at least one hydrogen bond with one of the residues of the P-loop and the implicit presence of a negative charge in that region, these results suggest that it would be necessary to consider further interactions within the binding site to achieve higher inhibitory potency during a VS workflow. For instance, Padney *et al.*[50] developed and validated different 3D-QSAR models from a set of peptidomimetic competitive inhibitors and concluded that: **(a)** hydrogen bonding with the carboxylate group of Asp48 is an important interaction for PTP1B inhibitor activity; and **(b)** the inclusion of steric bulks in the region around Phe182 are detrimental factors for PTP1B inhibitory activity. In another study, Gupta *et al.*[51] remarked the importance of hydrogen bonding with the side chain oxygen of Gln262 by comparing the binding pose of compound **5b** obtained in that study to the binding poses of the corresponding acetamide derivatives and their respective bioactivities. Moreover, the same study shows the key role played by Asp181, Ser216 and Arg221 in the anchorage of compound **5b** to the active site. From our docking studies, we observed that most of our compounds did not perform further hydrogen bonding interactions apart from those required with the residues of the P-loop (see Figure S9). Interestingly, the compounds with more activity identified in our study are the only ones capable of performing an additional π-π interaction with Tyr46 in the active site (*i.e.*, compounds **5** and **7**; see Figure S9) which, together with a second π-π interaction with Phe182, has been reported to be present in other complexes between PTP1B and potent inhibitors.[51,52] Moreover, the requirement of an interaction with secondary binding site residues like Arg24 and Arg254 would contribute to the selectivity of the VS hits over the highly homologous T-cell protein tyrosine phosphatase (TCPTP).[48,51,53] Thus, this indicates that there is still room for improvement in our VS protocol as seeking interactions with other residues in the PTP1B binding site may result in the identification of PTP1B inhibitors with higher potency and selectivity.

4. Experimental section

4.1. RF model

In order to prepare the molecules for the application of the RF model, ChemAxon's Standardizer[54] was used to generate canonical representations of each molecule and Morgan fingerprints of radius 2 were calculated with RDKit.[55] The RF model was built and validated with a set of 4,693 actives and 4,676 decoys using the Python package Scikit-learn[56] v0.17. The actives were obtained from ChEMBL[57] and Reaxys[39] and correspond to inhibitors of human PTP1B from those databases with bioactivity in the 1 to 15 range for pX, and whose activity was determined by measuring IC_{50} or K_i. MW-based decoys were obtained from the ZINC[58] database using Decoyfinder.[59] Some parameters of the model were adjusted by individually

evaluating their effect upon the performance of the model (see Figure S10). We chose to use 100 trees and to split the training and the test set into 80% and 20%, respectively. The model was validated via 5-fold cross-validation. The RF model performs well, with average precision and sensitivity values of 94.1% and 98.7% (see Table 2 for performance details). The output classification probabilities were calibrated using Platt scaling.[60] Thanks to probability calibration, the RF model provides the predicted probability of a compound being active. It is worth noting that the probability threshold can be increased or decreased depending on the library size to obtain the desired amount of input compounds for the following step of the virtual screening.

4.2. Selection of the crystal structure used for docking

Crystal structures for PTP1B/inhibitor complexes were obtained by VHELIBS[61] from PDB_REDO[62] under the Uniprot[63] accession number for human PTP1B (i.e. P18031) and the fitting of the coordinates of the inhibitors and the binding site relative to their corresponding electron density map was analyzed. Protein-ligand interaction schemes from the PDBsum[44] website were used to confirm the non-peptide and reversible character of the PTP1B inhibitor present in each complex. Furthermore, complexes with at least one mutation in their amino acid sequence were discarded. Thus, from the initial 72 structures corresponding to PTP1B-inhibitor complexes in the PDB_REDO,[62] 11 were rejected by VHELIBS[61] because either the inhibitor or the receptor binding site coordinates did not correctly fill their corresponding electron density map. Of the remaining 61 structures, 13 were discarded because the ligand was a peptide or covalently bound to PTP1B. From the remaining 48 structures, the one with PDB[34,35] code 1Q6M[36] (corresponding to the complex with an inhibitor that shows one of the lowest IC_{50}; i.e. 13 nM) was selected as the target for protein-ligand docking.

4.3. Ligand setup for docking

Before docking, ligand molecules were prepared with Maestro's v10.7 LigPrep[37,64] with default parameter values except for the following options: **a)** respect chiralities from input geometry when generating stereoisomers; **b)** use Epik[65] v3.7 for ionization and tautomerization; **c)** use 7.0 as effective pH; and **d)** use 2.0 as pH tolerance for generated structures.

4.4. Protein preparation

The A chain of the crystal structure with PDB[34,35] code 1Q6M[36] was prepared by using Maestro's v10.7 Protein Preparation Wizard[37,66] through the following procedure: **a)** remove original hydrogens; **b)** cap termini; **c)** generate ionization and tautomeric states of the ligand with Epik[65] (here the original state of the ligand was selected since it showed the lowest state penalty and the highest hydrogen bond count; i.e. state penalty: 0.17 Kcal/mol, Q: -4 and hydrogen bond count: 10); **d)** assign hydrogen bonds at pH 7 with

PROPKA; **e)** use force field OPLS_2005 to minimize the structure at 0.30 Å; and **f)** remove all water molecules from the structure.

4.5. Grid generation

The grid for protein-ligand docking was generated with Maestro[37] v10.7 by using default parameter values except for the following settings: **a)** the grid center coordinates were (0.25, 72.57, 70.0); **b)** aromatic hydrogens were included as hydrogen bond donors; **c)** halogens were included as acceptors; **d)** the inner box size was (10, 10, 10); **e)** the outer box size was (30, 30, 30); and **f)** hydrogen bond constraints were defined on the backbone nitrogens of the residues Ser216, Ala217, Ile219, Gly220 and Arg221, as well as the side-chain nitrogens of the residue Arg221 and the thiol group of the residue Cys215.

4.6. Docking

Protein-ligand docking was performed with Glide[67] v7.2 by using default parameter values except for the following settings: **a)** SP precision; **b)** enhance planarity of conjugated π groups; **c)** include aromatic H as donors; **d)** include halogens as acceptors; **e)** write out at most 10 poses per ligand; **f)** include 50 poses per ligand in post-docking minimization; and **g)** require accomplishment of one hydrogen bond constraint.

4.7. Electrostatic similarity analysis

The software EON[68] compares the poses for two different compounds by calculating Tanimoto coefficients associated either to their electrostatic potentials (i.e. Poisson-Boltzmann electrostatics and the coulombic part of the Poisson-Boltzmann electrostatics), to their shape, or to the combination of Poisson-Boltzmann electrostatics and their shape. The Poisson-Boltzmann electrostatics metric was used here, obtaining the electrostatic Tanimoto value (i.e. EON_ET_pb), which was in the -⅓ to 1 range (where a value of 1 corresponds to identical electrostatic potential overlap whereas negative values correspond to the overlap of opposite charges between the two poses). The validation set used in all electrostatic similarity comparisons was composed of 1,424 actives and 1,500 decoys. The actives were obtained from Reaxys[39] and corresponded to inhibitors of human PTP1B with a bioactivity in the 1 to 15 range for pX and whose activity was determined by measuring IC_{50} or K_i. MW-based decoys were obtained from the ZINC[58] database using Decoyfinder.[59] Prior to the analysis, the docking of the validation set was performed following the procedure described above.

In order to determine which crystallized ligands to use as references for the electrostatic similarity analysis, the PTP1B crystal structures containing ligands with IC_{50} or K_i activity values between 1 and 100 nM (a total of 14 crystal structures; see Table 3) were obtained and superposed to the crystal structure with PDB[34,35] code 1Q6M.[36] The validation set was then used to perform 14 separate validations (i.e. one for each of the

14 crystallized ligands in the superposed crystal structures; see Figure S2) using the crystallized ligands as queries. In each validation, only the docked pose that presented the highest electrostatic Tanimoto with the corresponding query was kept for each library compound. After discarding the rest of the docking poses, histograms of the electrostatic Tanimoto coefficients for the actives and the decoys in the validation set were plotted, separating the actives into 3 groups depending on whether their pX value was lower than 4, between 4 and 7 or higher than 7 (see Figure S2). For 3 of the 14 initial crystallized actives (ligands in crystal structures with the PDB[34,35] codes 2VEU,[69] 2VEW,[69] 2VEY;[69] see Figure S2) the electrostatic Tanimoto value distribution of the actives was very similar to that of the decoys, so no electrostatic Tanimoto value could be used as a cutoff to differentiate between the two groups, and for this reason these queries were discarded. The remaining 11 crystallized actives (ligands in crystal structures with the PDB[34,35] codes 1Q6J, [36] 1Q6M,[36] 1Q6N,[36] 1Q6P,[36] 1Q6S,[36] 1Q6T,[36] 2FJN,[49] 2QBP,[70] 2QBQ,[70] 2ZMM,[71] 2ZN7;[71] see underlined PDB codes in Table 3 and Figure S2) were selected as queries for the *in silico* validation of the electrostatic similarity analysis.

4.8. Clustering

The HDBSCAN[46] Clustering Library from Python was used for clustering. First, a distance matrix was developed using scikit-learn based on the “rogerstanimoto” metric and then HDBSCAN[46] clustering was performed with the following settings: **a)** minimum cluster size: 2; **b)** metric: “precomputed”.

4.9. Bioactivity assays

The colorimetric PTP1B Drug Discovery Kit (BML-AK822) was purchased from Enzo Life Sciences. The compounds were fully dissolved in DMSO at a concentration of 25 mM. For the inhibitory activity assay, each compound was dissolved in 1X assay buffer to obtain the final concentrations of 100 μM and 10 μM in a volume of 45 μL per well with the respective DMSO concentrations of 0.4% and 0.04%. Controls with DMSO were prepared at these concentrations to ensure that DMSO had no effect on PTP1B activity (data not shown). After warming for 10 min to the assay temperature of 37 ºC, 5 μL of human recombinant PTP1B stock solution (0.6 ng/μM) were added. The reaction was initiated with 50 μL of the substrate IR5 (insulin receptor β residues 1142-1153 with pY at position 1146) dissolved in 1X assay buffer to a concentration of 60 μM. After an incubation period of 2 h at 37 ºC the reaction was terminated by adding of 25 μL of Biomol Red Reagent. PTP1B dephosphorylates the IR5 substrate resulting in the release of orthophosphate, which was quantified after incubating for 30 min at 620 nm using the microtiter plate reader EON Microplate (BioTek, Vermont, USA). The maximum absorbance values were detected 10 min after exposure to Biomol Red reagent. In this initial bioactivity assay, only two measures per compound were performed (one at 10 μM and another one at 100 μM concentration; see Table 4). For the determination of IC_{50} values of the two most active compounds, they were diluted in 1X assay buffer to obtain the following final concentrations: 1000, 300, 100, 30, 10, 3, 1, 0.3, 0.1, 0.03, 0.01, 0.003, 0.001, and 0.0003 μM. Suramin was used as the positive

control. The percentages of inhibition at each concentration were determined using triplicates and IC_{50} values were calculated using a four parameter logistic regression (see Figure S7).

Acknowledgments

This study was supported by research grants 2014PFR-URV-B2-67 and 2015PFR-URV-B2-67 from our University. AG's contract is supported by grant 2015FI_B00655 from the Government of Catalonia. MM is a Serra Hunter research fellow. We thank OpenEye Scientific Software and ChemAxon Ltd. for kindly providing us with a software bursary to use their programs. This manuscript has been edited by the English language service of our university.

Conflict of interest

The authors declare no conflict of interest.

Keywords

Virtual screening; Protein tyrosine phosphatase 1B; Inhibitors; Obesity; Type 2 diabetes mellitus

References

[1] International Diabetes Federation, https://www.idf.org/

[2] World Health Organization, http://www.who.int/

[3] L. Yao, O. Herlea-Pana, J. Heuser-Baker, Y. Chen, J. Barlic-Dicen, *J. Immunol. Res.*, **2014**, *2014*, 181450.

[4] E.K. Anderson, D.A. Gutierrez, A.H. Hasty, *Curr. Opin. Lipidol.*, **2010**, *21*, 172–7.

[5] B.K. Surmi, A.H. Hasty, *Vascul. Pharmacol.*, **2010**, *52*, 27–36.

[6] N. Dali-Youcef, M. Mecili, R. Ricci, E. Andrès, *Ann. Med.*, **2013**, *45*, 242–253.

[7] U.J. Jung, M.-S. Choi, *Int. J. Mol. Sci.*, **2014**, *15*, 6184–223.

[8] Y.S. Lee, H. Morinaga, J.J. Kim, W. Lagakos, S. Taylor, M. Keshwani, G. Perkins, H. Dong, A.G. Kayali, I.R. Sweet, J. Olefsky, *Cell*, **2013**, *153*, 413–25.

[9] N. Esser, S. Legrand-Poels, J. Piette, A.J. Scheen, N. Paquot, *Diabetes Res. Clin. Pract.*, **2014**, *105*, 141–50.

[10] V.R. Richardson, K.A. Smith, A.M. Carter, *Immunobiology*, **2013**, *218*, 1497–504.

[11] L.F. Van Gaal, I.L. Mertens, C.E. De Block, *Nature*, **2006**, *444*, 875–80.

[12] M.Y. Donath, É. Dalmas, N.S. Sauter, M. Böni-Schnetzler, *Cell Metab.*, **2013**, *17*, 860–72.

[13] K.G.M.M. Alberti, R.H. Eckel, S.M. Grundy, P.Z. Zimmet, J.I. Cleeman, K.A. Donato, J.-C. Fruchart, W.P.T. James, C.M. Loria, S.C. Smith, International Diabetes Federation Task Force on Epidemiology and Prevention, and B.I. Hational Heart, Lung, American Heart Association, World Heart Federation, International Atherosclerosis Society, International Association for the Study of Obesity, *Circulation*, **2009**, *120*, 1640–5.

[14] S. Guo, *J. Endocrinol.*, **2014**, *220*, T1–T23.

[15] R. Di Paola, L. Frittitta, G. Miscio, M. Bozzali, R. Baratta, M. Centra, D. Spampinato, M.G. Santagati, T. Ercolino, C. Cisternino, T. Soccio, S. Mastroianno, V. Tassi, P. Almgren, A. Pizzuti, R. Vigneri, V. Trischitta, *Am. J. Hum. Genet.*, **2002**, *70*, 806–12.

[16] N.T. Lam, S.D. Covey, J.T. Lewis, S. Oosman, T. Webber, E.C. Hsu, A.T. Cheung, T.J. Kieffer, *J. Mol. Endocrinol.*, **2006**, *36*, 163–74.

[17] C.D. Morrison, C.L. White, Z. Wang, S.-Y. Lee, D.S. Lawrence, W.T. Cefalu, Z.-Y. Zhang, T.W. Gettys, *Endocrinology*, **2007**, *148*, 433–40.

[18] S. Koren, I.G. Fantus, *Best Pract. Res. Clin. Endocrinol. Metab.*, 2007, *21*, 621–40.

[19] M.F. Cicirelli, N.K. Tonks, C.D. Diltz, J.E. Weiel, E.H. Fischer, E.G. Krebs, *Proc. Natl. Acad. Sci. U. S. A.*, 1990, *87*, 5514–8.

[20] B.J. Goldstein, A. Bittner-Kowalczyk, M.F. White, M. Harbeck, *J. Biol. Chem.*, **2000**, *275*, 4283–9.

[21] A. Cheng, N. Uetani, P.D. Simoncic, V.P. Chaubey, A. Lee-Loy, C.J. McGlade, B.P. Kennedy, M.L. Tremblay, *Dev. Cell*, **2002**, *2*, 497–503.

[22] J.M. Zabolotny, K.K. Bence-Hanulec, A. Stricker-Krongrad, F. Haj, Y. Wang, Y. Minokoshi, Y.-B. Kim, J.K. Elmquist, L.A. Tartaglia, B.B. Kahn, B.G. Neel, *Dev. Cell*, **2002**, *2*, 489–95.

[23] P.-A. Thiebaut, M. Besnier, E. Gomez, V. Richard, *J. Mol. Cell. Cardiol.*, **2016**, *101*, 50–57.

[24] A.K. Saxena, G. Pandey, S. Gupta, A.B. Singh, A.K. Srivastava, *Bioorganic Med. Chem. Lett.*, **2009**, *19*, 2320–2323.

[25] A.P. Combs, *J. Med. Chem.*, **2010**, *53*, 2333–2344.

[26] W.B. Floriano, N. Vaidehi, G. Zamanakos, W.A. Goddard, *J. Med. Chem.*, **2004**, *47*, 56–71.

[27] K. Bharatham, N. Bharatham, K.W. Lee, *Arch. Pharm. Res.*, **2007**, *30*, 533–542.

[28] H. Park, B.R. Bhattarai, S.W. Ham, H. Cho, *Eur. J. Med. Chem.*, **2009**, *44*, 3280–3284.

[29] P.S. Rao, C. Muvva, K. Geethanjali, S.B. Bastipati, R. Kalashikam, *Bioinformation*, **2012**, *8*, 834–7.

[30] M.V.V.V.S. Reddy, C. Ghadiyaram, S.K. Panigrahi, S. Hosahalli, L.N. Mangamoori, *Mini Rev. Med. Chem.*, **2013**, *13*, 1602–6.

[31] Y. Ma, Y.-Y. Jin, Y.-L. Wang, R.-L. Wang, X.-H. Lu, D.-X. Kong, W.-R. Xu, *Chem. Biol. Drug Des.*, **2014**, *83*, 697–709.

[32] V.M. Balaramnavar, R. Srivastava, N. Rahuja, S. Gupta, A.K. Rawat, S. Varshney, H. Chandasana, Y.S. Chhonker, P.K. Doharey, S. Kumar, S. Gautam, S.P. Srivastava, R.S. Bhatta, J.K. Saxena, A.N. Gaikwad, A.K. Srivastava, A.K. Saxena, *Eur. J. Med. Chem.*, **2014**, *87*, 578–94.

[33] S. Chandra, J. Pandey, A.K. Tamrakar, M.I. Siddiqi, *J. Mol. Graph. Model.*, **2017**, *71*, 242–256.

[34] RCSB PDB, http://www.rcsb.org

[35] H.M. Berman, *Nucleic Acids Res.*, **2000**, *28*, 235–242.

[36] G. Scapin, S.B. Patel, J.W. Becker, Q. Wang, C. Desponts, D. Waddleton, K. Skorey, W. Cromlish, C. Bayly, M. Therien, J.Y. Gauthier, C.S. Li, C.K. Lau, C. Ramachandran, B.P. Kennedy, E. Asante-Appiah, *Biochemistry*, **2003**, *42*, 11451–9.

[37] Schrödinger Release 2016-3: Maestro, Schrödinger, LLC, New York, NY, **2016**.

[38] Specs, http://www.specs.net

[39] Reaxys, https://www.reaxys.com/

[40] A. Cereto-Massagué, M.J. Ojeda, C. Valls, M. Mulero, S. Garcia-Vallvé, G. Pujadas, *Methods*, **2014**, *71*, 6–11.

[41] M.A. Ghattas, N. Atatreh, E. V. Bichenkova, R.A. Bryce, *J. Mol. Graph. Model.*, **2014**, *52*, 114–23.

[42] A.C. Wallace, R.A. Laskowski, J.M. Thornton, *Protein Eng.*, **1995**, *8*, 127–34.

[43] K. Stierand, P.C. Maass, M. Rarey, *Bioinformatics*, **2006**, *22*, 1710–6.

[44] PDBsum, www.ebi.ac.uk/pdbsum/

[45] Marvin 16.10.10.0, **2016**, ChemAxon, http://www.chemaxon.com

[46] L. McInnes, J. Healy, S. Astels, *J. Open Source Softw.*, **2017**, *2*, 1–3.

[47] I. Letunic, P. Bork, *Nucleic Acids Res.*, **2016**, *44*, W242-5.

[48] Y.A. Puius, Y. Zhao, M. Sullivan, D.S. Lawrence, S.C. Almo, Z.Y. Zhang, *Proc. Natl. Acad. Sci. U. S. A.*, **1997**, *94*, 13420–5.

[49] E. Asante-Appiah, S. Patel, C. Desponts, J.M. Taylor, C. Lau, C. Dufresne, M. Therien, R. Friesen, J.W. Becker, Y. Leblanc, B.P. Kennedy, G. Scapin, *J. Biol. Chem.*, **2006**, *281*, 8010–5.

[50] G. Pandey, A.K. Saxena, *J. Chem. Inf. Model.*, **2006**, *46*, 2579–90.

[51] S. Gupta, K. Varshney, R. Srivastava, N. Rahuja, A.K. Rawat, A.K. Srivastava, A.K. Saxena, *Medchemcomm*, **2013**, 1382–1387.

[52] S.R. Klopfenstein, A.G. Evdokimov, A.-O.O. Colson, N.T. Fairweather, J.J. Neuman, M.B. Maier, J.L. Gray, G.S. Gerwe, G.E. Stake, B.W. Howard, J.A. Farmer, M.E. Pokross, T.R. Downs, B. Kasibhatla, K.G. Peters, *Bioorganic Med. Chem. Lett.*, **2006**, *16*, 1574–8.

[53] D. Patel, M. Jain, S.R. Shah, R. Bahekar, P. Jadav, A. Joharapurkar, N. Dhanesha, M. Shaikh, K.V.V.M. Sairam, P. Kapadnis, *Bioorganic Med. Chem. Lett.*, **2012**, *22*, 1111–1117.

[54] Standardizer 16.10.10.0, **2016**, ChemAxon, http://www.chemaxon.com

[55] RDKit: Open-source cheminformatics, http://www.rdkit.org

[56] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, G. Louppe, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, É. Duchesnay, **2012**, *12*, 2825–2830.

[57] A.P. Bento, A. Gaulton, A. Hersey, L.J. Bellis, J. Chambers, M. Davies, F.A. Krüger, Y. Light, L. Mak, S. McGlinchey, M. Nowotka, G. Papadatos, R. Santos, J.P. Overington, *Nucleic Acids Res.*, **2014**, *42*, D1083-90.

[58] J.J. Irwin, T. Sterling, M.M. Mysinger, E.S. Bolstad, R.G. Coleman, *J. Chem. Inf. Model.*, **2012**, *52*, 1757–68.

[59] A. Cereto-Massagué, L. Guasch, C. Valls, M. Mulero, G. Pujadas, S. Garcia-Vallvé, *Bioinformatics*, **2012**, *28*, 1661–1662.

[60] J. Platt, *Adv. large margin Classif.*, **1999**, *10*, 61–74.

[61] A. Cereto-Massagué, M.J. Ojeda, R.P. Joosten, C. Valls, M. Mulero, M.J. Salvado, A. Arola-Arnal, L. Arola, S. Garcia-Vallvé, G. Pujadas, *J. Cheminform.*, **2013**, *5*, 36.

[62] PDB_REDO, http://www.cmbi.ru.nl/pdb_redo/

[63] A. Bateman, M.J. Martin, C. O'Donovan, M. Magrane, E. Alpi, R. Antunes, B. Bely, M. Bingley, C. Bonilla, R. Britto, B. Bursteinas, H. Bye-AJee, A. Cowley, A. Da Silva, M. De Giorgi, T. Dogan, F. Fazzini, L.G. Castro, L. Figueira, P. Garmiri, G. Georghiou, D. Gonzalez, E. Hatton-Ellis, W. Li, W. Liu, R. Lopez, J. Luo, Y. Lussi, A. MacDougall, A. Nightingale, B. Palka, K. Pichler, D. Poggioli, S. Pundir, L. Pureza, G. Qi, S. Rosanoff, R. Saidi, T. Sawford, A. Shypitsyna, E. Speretta, E. Turner, N. Tyagi, V. Volynkin, T. Wardell, K. Warner, X. Watkins, R. Zaru, H. Zellner, I. Xenarios, L. Bougueleret, A. Bridge, S. Poux, N. Redaschi, L. Aimo, G. ArgoudPuy, A. Auchincloss, K. Axelsen, P. Bansal, D. Baratin, M.C. Blatter, B. Boeckmann, J. Bolleman, E. Boutet, L. Breuza, C. Casal-Casas, E. De Castro, E. Coudert, B. Cuche, M. Doche, D. Dornevil, S. Duvaud, A. Estreicher, L. Famiglietti, M. Feuermann, E. Gasteiger, S. Gehant, V. Gerritsen, A. Gos, N. Gruaz-Gumowski, U. Hinz, C. Hulo, F. Jungo, G. Keller, V. Lara, P. Lemercier, D. Lieberherr, T. Lombardot, X. Martin, P. Masson, A. Morgat, T. Neto, N. Nouspikel, S. Paesano, I. Pedruzzi, S. Pilbout, M. Pozzato, M. Pruess, C. Rivoire, B. Roechert, M. Schneider, C. Sigrist, K. Sonesson, S. Staehli, A. Stutz, S. Sundaram, M. Tognolli, L. Verbregue, A.L. Veuthey, C.H. Wu, C.N. Arighi, L. Arminski, C. Chen, Y. Chen, J.S. Garavelli, H. Huang, K. Laiho, P. McGarvey, D.A. Natale, K. Ross, C.R. Vinayaka, Q. Wang, Y. Wang, L.S. Yeh, J. Zhang, *Nucleic Acids Res.*, 2017, *45*, D158–D169.

[64] Schrödinger Release 2016-3: LigPrep, Schrödinger, LLC, New York, NY, **2016**.

[65] Schrödinger Release 2016-3: Epik, Schrödinger, LLC, New York, NY, **2016**.

[66] Schrödinger Release 2016-3: Schrödinger Suite 2016-3 Protein Preparation Wizard; Epik, Schrödinger, LLC, New York, NY, 2016; Impact, Schrödinger, LLC, New York, NY, 2016; Prime, Schrödinger, LLC, New York, NY, **2016**.

[67] Schrödinger Release 2016-3: Glide, Schrödinger, LLC, New York, NY, **2016**.

[68] EON 2.2.0.5: OpenEye Scientific Software, Santa Fe, NM., http://www.eyesopen.com

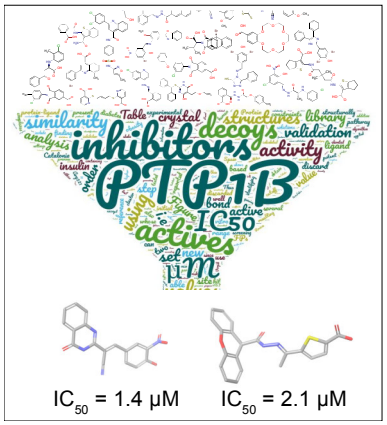
[69] B. Douty, B. Wayland, P.J. Ala, M.J. Bower, J. Pruitt, L. Bostrom, M. Wei, R. Klabe, L. Gonneville, R. Wynn, T.C. Burn, P.C.C. Liu, A.P. Combs, E.W. Yue, *Bioorganic Med. Chem. Lett.*, **2008**, *18*, 66–71.

[70] D.P. Wilson, Z.-K.K. Wan, W.-X.X. Xu, S.J. Kirincich, B.C. Follows, D. Joseph-McCarthy, K. Foreman, A. Moretto, J. Wu, M. Zhu, E. Binnun, Y.-L.L. Zhang, M. Tam, D. V. Erbe, J. Tobin, X. Xu, L. Leung, A. Shilling, S.Y. Tam, T.S. Mansour, J. Lee, *J. Med. Chem.*, **2007**, *50*, 4681–98.

[71] Z.-K. Wan, J. Lee, R. Hotchandani, A. Moretto, E. Binnun, D.P. Wilson, S.J. Kirincich, B.C. Follows, M. Ipek, W. Xu, D. Joseph-McCarthy, Y.-L. Zhang, M. Tam, D. V Erbe, J.F. Tobin, W. Li, S.Y. Tam, T.S. Mansour, J. Wu, *ChemMedChem*, **2008**, *3*, 1525–9.

Table of contents

Herein we have developed a virtual screening workflow consisting of a combination of ligand- and receptor-based methods which has been able to identify 15 structurally diverse PTP1B inhibitors, two of them with IC_{50} values in the 1 - 10 μM range.



Supplementary material for

Combined ligand- and receptor-based virtual screening methodology to identify structurally diverse PTP1B inhibitors

Aleix Gimeno[a], Andrea Ardid-Ruiz[b], María José Ojeda-Montes[a], Sarah Tomás-Hernández[a], Adrià Cereto-Massagué[a], Raúl Beltrán-Debón[a], Miquel Mulero[a], Cristina Valls[a], Gerard Aragonès[b], Manuel Suárez[b], Gerard Pujadas[a],[c],*, Santiago Garcia-Vallvé[a],[c]

[a]Research group in Cheminformatics & Nutrition, Departament de Bioquímica i Biotecnologia, Universitat Rovira i Virgili, Campus de Sescelades, 43007 Tarragona, Catalonia, Spain

[b]Nutrigenomics Research Group, Department of Biochemistry and Biotechnology, Universitat Rovira i Virgili (URV), Tarragona, Spain

[c]EURECAT, TECNIO, CEICS, Avinguda Universitat, 1, 43204 Reus, Catalonia, Spain

**Correspondence to:* Gerard Pujadas, Research group in Cheminformatics & Nutrition, phone: +34 977 55 95 65, fax: +34 977 55 82 32. Departament de Bioquímica i Biotecnologia, Facultat de Química, Universitat Rovira i Virgili, C/ Marcel·lí Domingo 1, Edifici N4, 43007 Tarragona, Catalonia, Spain. E-mail: gerard.pujadas@urv.cat

Table S1. Purity and exact mass of the 20 compounds whose bioactivities have been tested.

Compound	Specs ID	Purity (%)	Formula	Pseudo molecular ion adduct	Theoretical m/z	Measured m/z
1	AG-205/07930063	98.98	C18 H16 N2 O3 S	[M+H]+	341.0954	341.0999
2	AM-807/41462317	98.92	C18 H12 F3 N3 O5 S	[M+H]+	440.0523	440.052
3	AI-204/43362778	97.92	C16 H14 N2 O4 S2	[M+H]+	363.0468	363.0466
4	AG-690/11836029	98.43	C12 H8 Br Cl O2 S	[M]·+	329.9111	329.9102
5	AF-399/15030257	97.77	C17 H10 N4 O4	[M+H]+	335.0775	335.0766
6	AF-399/15285024	98.72	C21 H22 N2 O4	[M+H]+	367.1652	367.1663
7	AK-968/41025519	99.64	C21 H16 N2 O4 S	[M+H]+	393.0904	393.095
8	AG-690/40107322	95.29	C18 H16 Br2 F N7 O2	[M+H]+	539.9789	539.9781
9	AJ-077/33270018	97.56	C22 H14 N4 O4	[M+H]+	399.1088	399.1092
10	AH-487/41936905	96.38	C9 H8 I2 O4	[M+NH4]+	451.8850	451.884
11	AE-641/42119771	99.65	C13 H10 Cl N O5 S	[M+H]+	328.0041	328.003
12	AK-968/15361893	99.29	C19 H20 Cl N O5 S	[M+H]+	410.0823	410.0812
13	AG-205/37175018	87.96	C16 H16 N2 O6	[M+H]+	333.1081	333.1074
14	AF-886/30566062	99.27	C16 H10 Cl N3 O3	[M+H]+	328.0483	328.0475
15	AG-205/37169028	96.02	C18 H12 Cl F N2 O3	[M+H]+	359.0593	359.0613
16	AK-968/11197127	99.33	C19 H16 Cl N3 O3	[M+H]+	370.0953	370.0994
17	AF-962/02757026	93.56	C13 H10 N2 O5 S2	[M+H]+	339.0104	339.0147
18	AM-879/12634007	96.84	C16 H8 Cl3 N O2 S2	[M+H]+	415.9135	415.9108
19	AG-690/36826026	96.37	C11 H8 Br N O3 S2	[M+H]+	345.9202	345.9189
20	AK-968/37202023	98.16	C14 H6 Br2 F3 N3 O2	[M+H]+	463.8852	463.8832

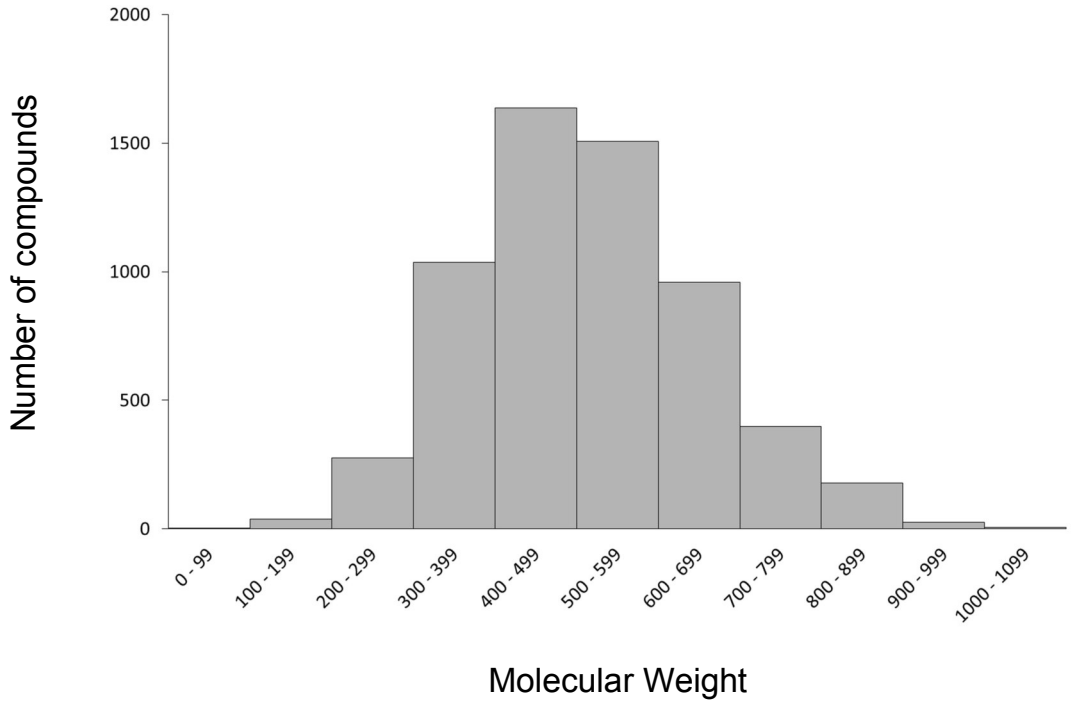


Figure S1. Histogram representation of the molecules used as references to establish a MW range for the MW filter.

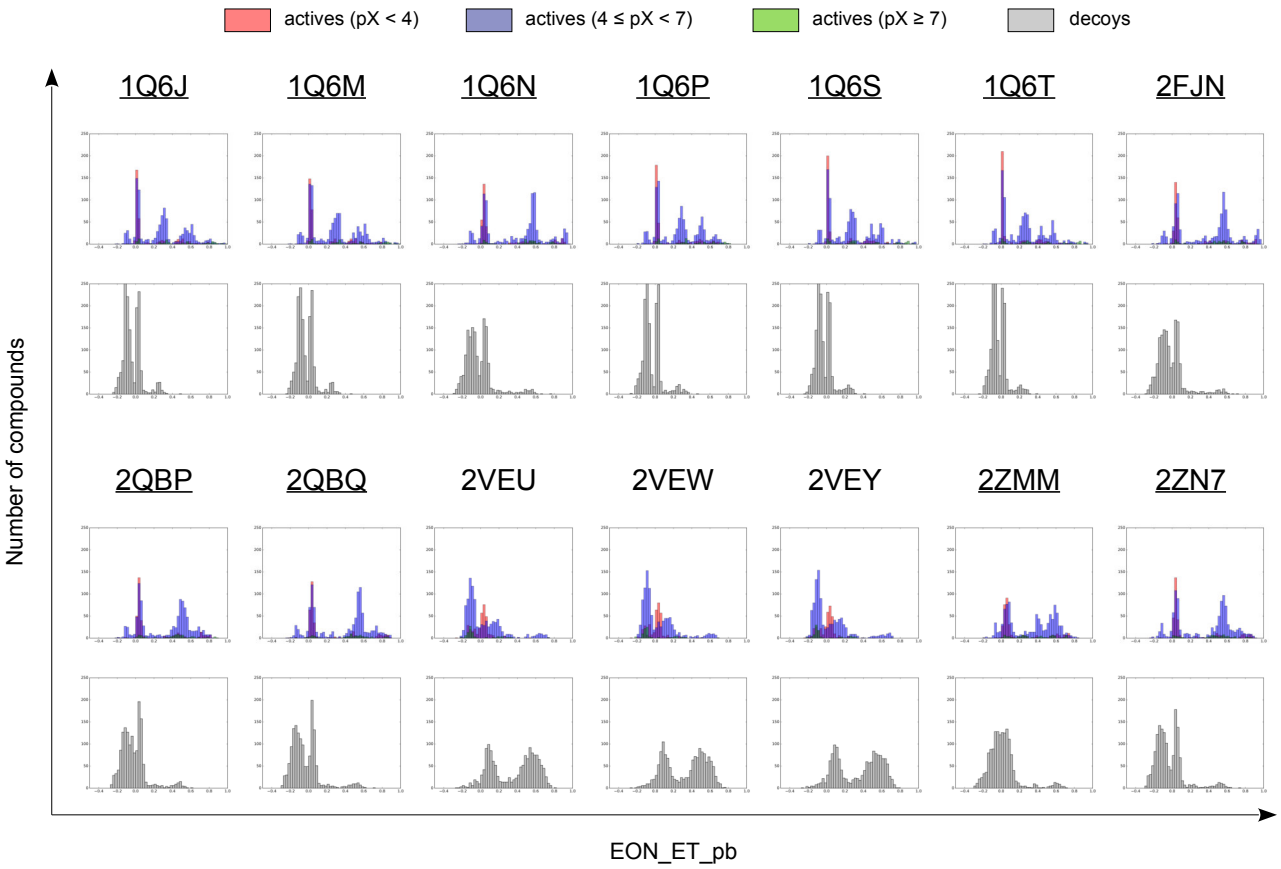
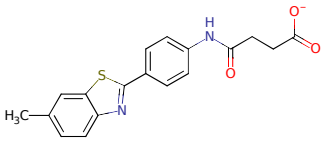
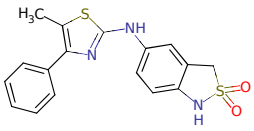


Figure S2. Histogram representation of the highest electrostatic Tanimoto (i.e. EON_ET_pb) values obtained in the comparison of the validation set to each query. For each query, the PDB[34,35] code of the crystal structure from which it was obtained and two histograms are shown: one corresponding to the actives and one corresponding to the decoys. In the actives histogram, actives with pX lower than 4 are in red, actives with pX from 4 to 7 are in blue and actives with pX higher than 7 are in green. In the decoys histogram, decoys are in gray. The 11 queries selected for the *in silico* validation of the electrostatic similarity analysis are underlined.

A)

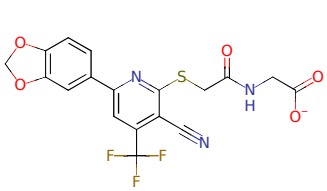


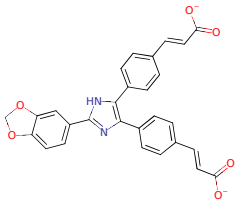


1 (AG-205/07930063) Reaxys RN: 25950177

Tanimoto: 0.39

B)

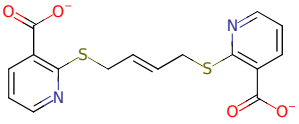


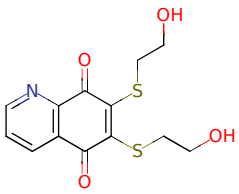


2 (AM-807/41462317) Reaxys RN: 24985723

Tanimoto: 0.34

C)

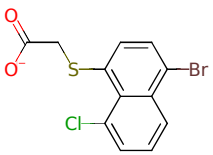


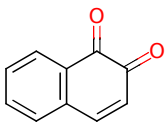


3 (AI-204/43362778) Reaxys RN: 24734948

Tanimoto: 0.31

D)

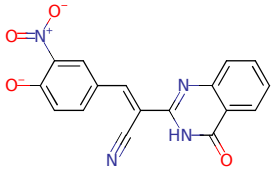


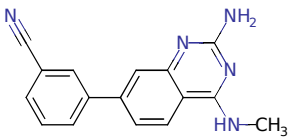


4 (AG-690/11836029) Reaxys RN: 606546

Tanimoto: 0.31

E)

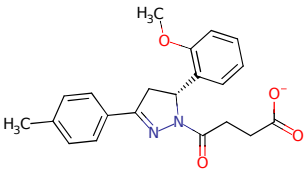


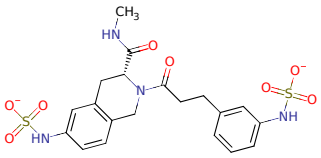


5 (AF-399/15030257) Reaxys RN: 12659520

Tanimoto: 0.46

F)

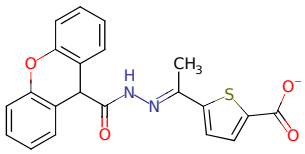


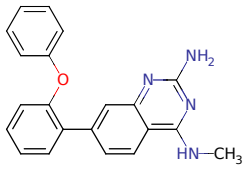


6 (AF-399/15285024) Reaxys RN: 25330632

Tanimoto: 0.25

G)

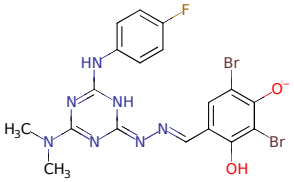


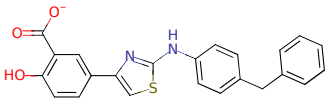


7 (AK-968/41025519) Reaxys RN: 12601206

Tanimoto: 0.27

H)

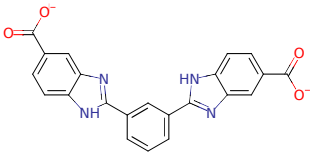


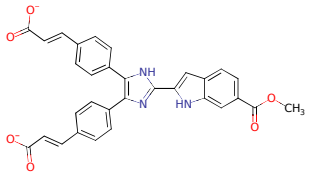


8 (AG-690/40107322) Reaxys RN: 12509120

Tanimoto: 0.30

I)

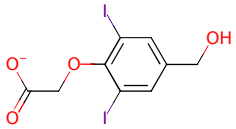


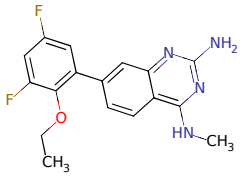


9 (AJ-077/33270018) Reaxys RN: 24985689

Tanimoto: 0.55

J)

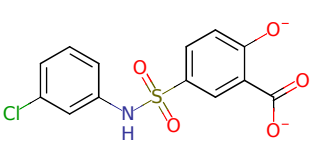


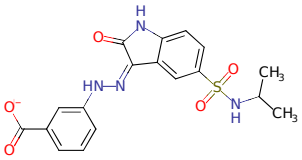


10 (AH-487/41936905) Reaxys RN: 12601276

Tanimoto: 0.23

K)

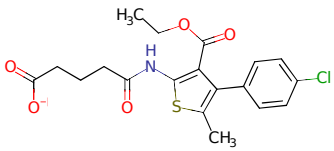


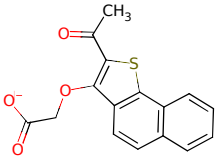


11 (AE-641/42119771) Reaxys RN: 13005572

Tanimoto: 0.43

L)

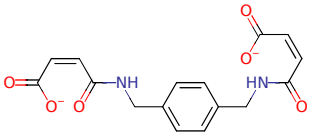


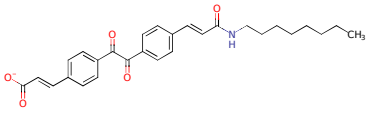


12 (AK-968/15361893) Reaxys RN: 28048260

Tanimoto: 0.39

M)

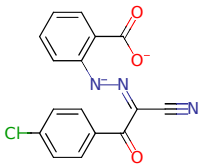


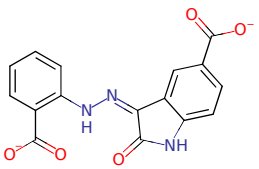


13 (AG-205/37175018) Reaxys RN: 24985673

Tanimoto: 0.39

N)

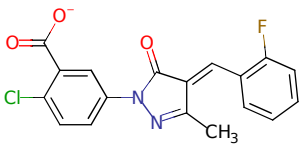


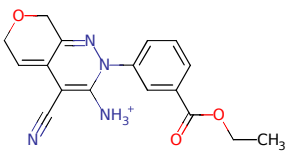


14 (AF-886/30566062) Reaxys RN: 13005538

Tanimoto: 0.48

O)

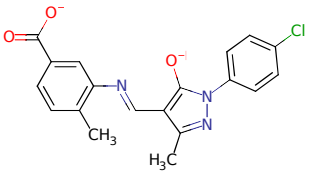


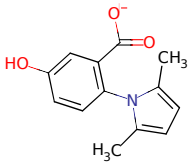


15 (AG-205/37169028) Reaxys RN: 13942891

Tanimoto: 0.34

P)

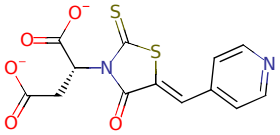


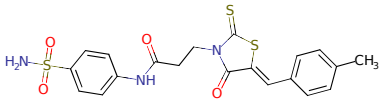


16 (AK-968/11197127) Reaxys RN: 15419640

Tanimoto: 0.34

Q)

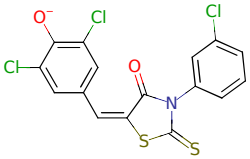


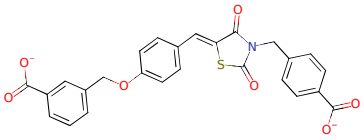


17 (AF-962/02757026) Reaxys RN: 25432438

Tanimoto: 0.31

R)

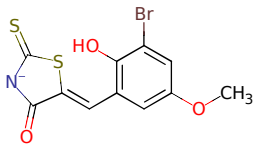


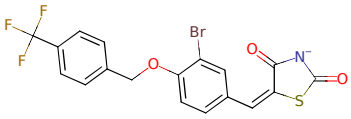


18 (AM-879/12634007) Reaxys RN: 26119874

Tanimoto: 0.34

S)

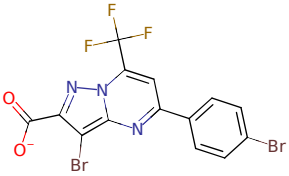


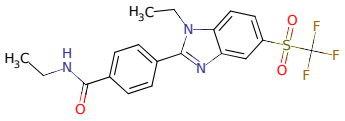


19 (AG-690/36826026) Reaxys RN: 25232528

Tanimoto: 0.31

T)



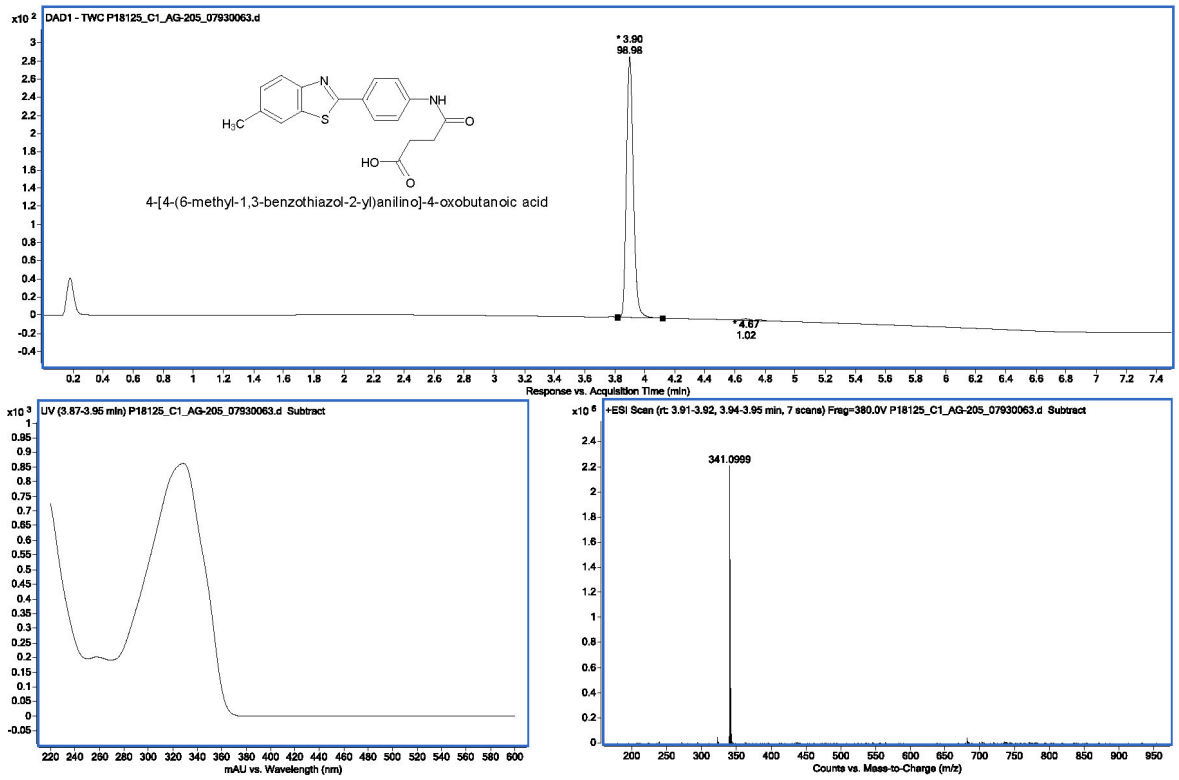


20 (AK-968/37202023) Reaxys RN: 15705274

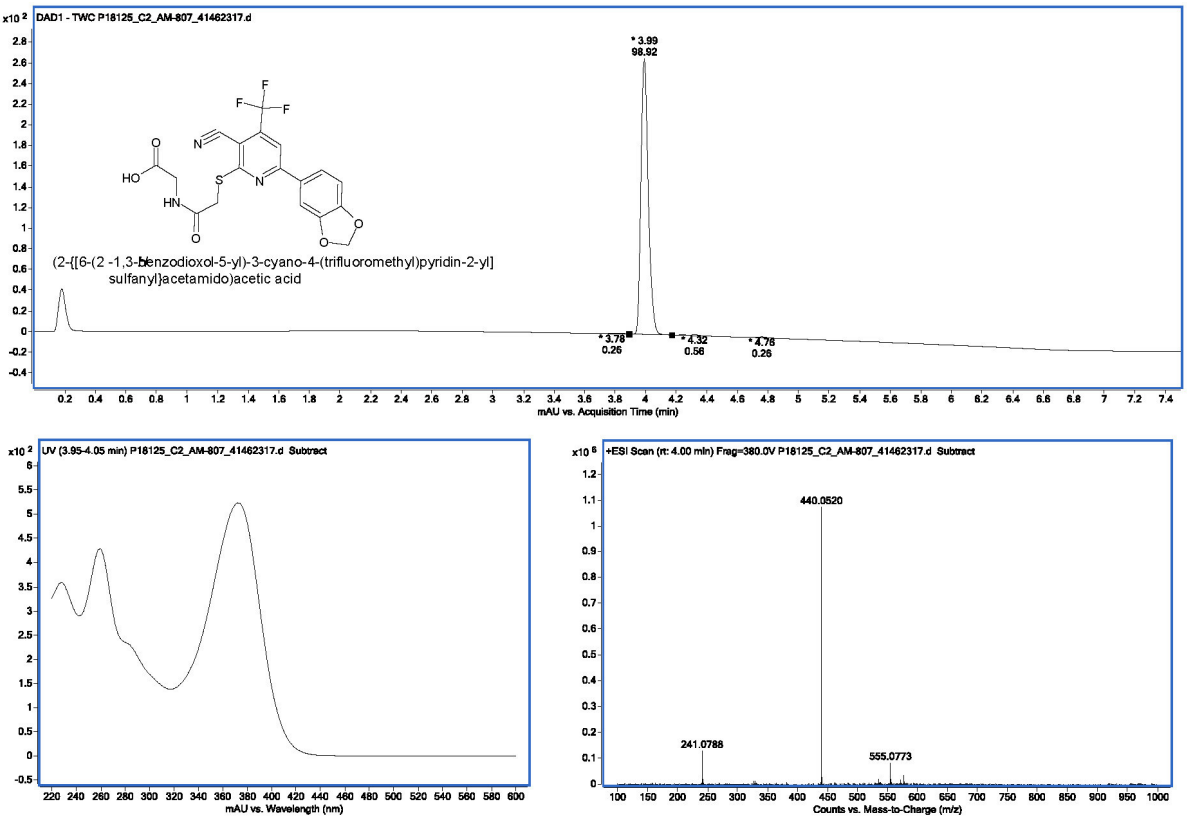
Tanimoto: 0.34

Figure S3. In each panel, a hit compound is represented in 2D, together with the most similar active compound that was used during the clustering. In each case, the most similar active compound is labeled with its Reaxys[39] Registry Number. The Tanimoto similarity value resulting from the comparison of the OpenEyePath FP of both compounds is shown below. MarvinSketch[45] was used to draw the structures. The protonation state of each compound corresponds to the protonation state of the docked pose selected for that compound.

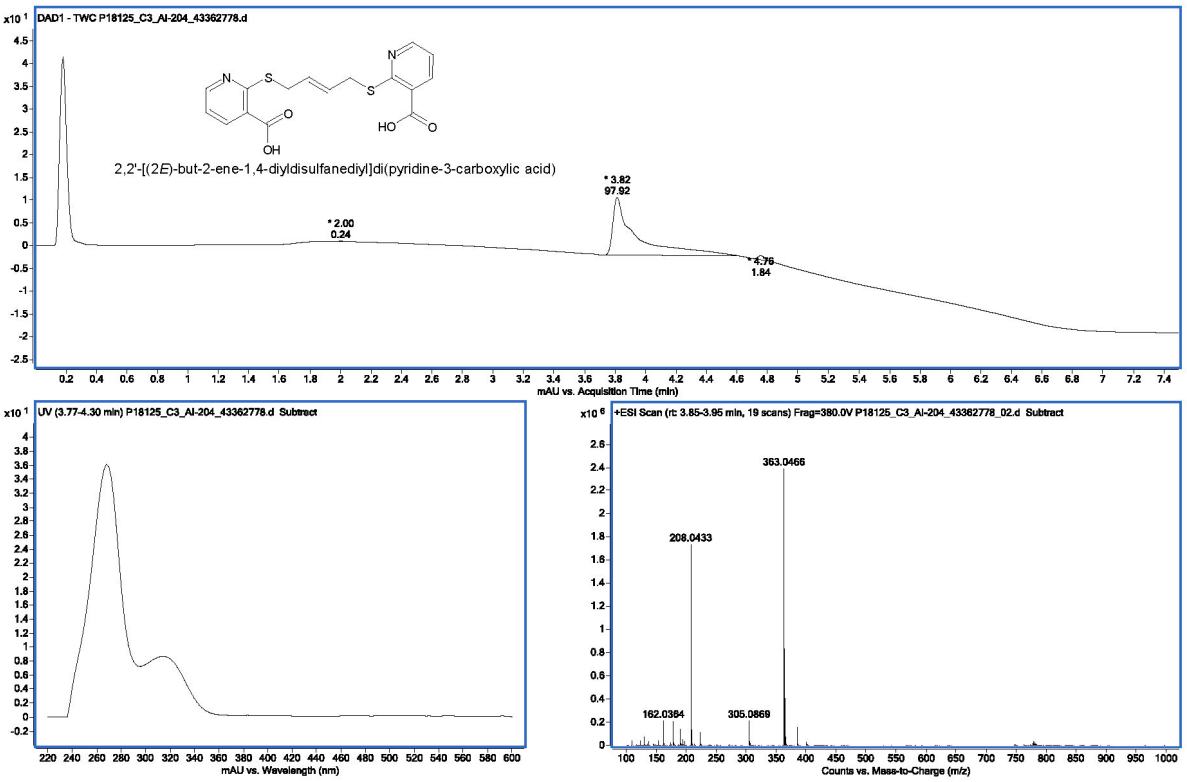
A)


DAD1 - TWC P18125_C1_AG-205_07930063.d
* 3.90
98.98
* 4.67
1.02
4-[4-(6-methyl-1,3-benzothiazol-2-yl)anilino]-4-oxobutanoic acid
Response vs. Acquisition Time (min)
UV (3.87-3.95 min) P18125_C1_AG-205_07930063.d Subtract
mAU vs. Wavelength (nm)
+ESI Scan (rt: 3.91-3.92, 3.94-3.95 min, 7 scans) Frag=380.0V P18125_C1_AG-205_07930063.d Subtract
341.0999
Counts vs. Mass-to-Charge (m/z)

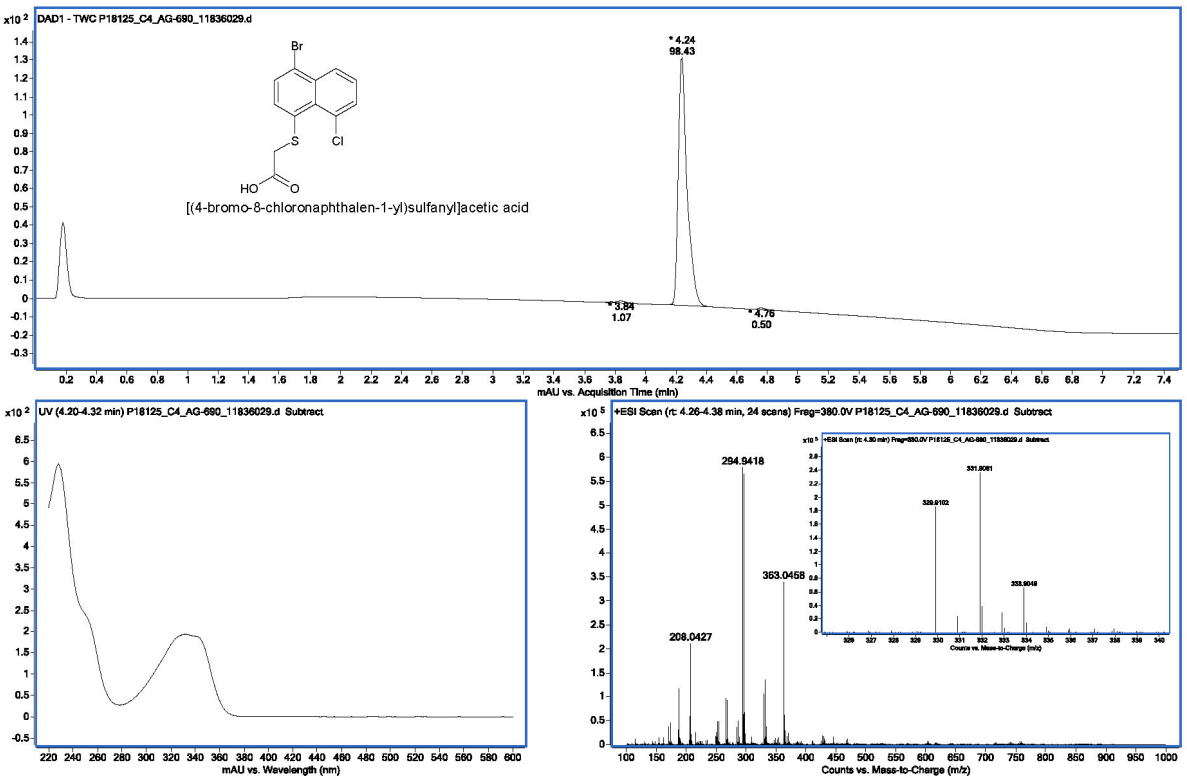
B)


DAD1 - TWC P18125_C2_AM-807_41462317.d
* 3.99
98.92
* 3.78
0.26
* 4.32
0.56
* 4.76
0.26
(2-{[6-(2 -1,3-benzodioxol-5-yl)-3-cyano-4-(trifluoromethyl)pyridin-2-yl]
sulfanyl}acetamido)acetic acid
mAU vs. Acquisition Time (min)
UV (3.95-4.05 min) P18125_C2_AM-807_41462317.d Subtract
mAU vs. Wavelength (nm)
+ESI Scan (rt: 4.00 min) Frag=380.0V P18125_C2_AM-807_41462317.d Subtract
440.0520
241.0788
555.0773
Counts vs. Mass-to-Charge (m/z)

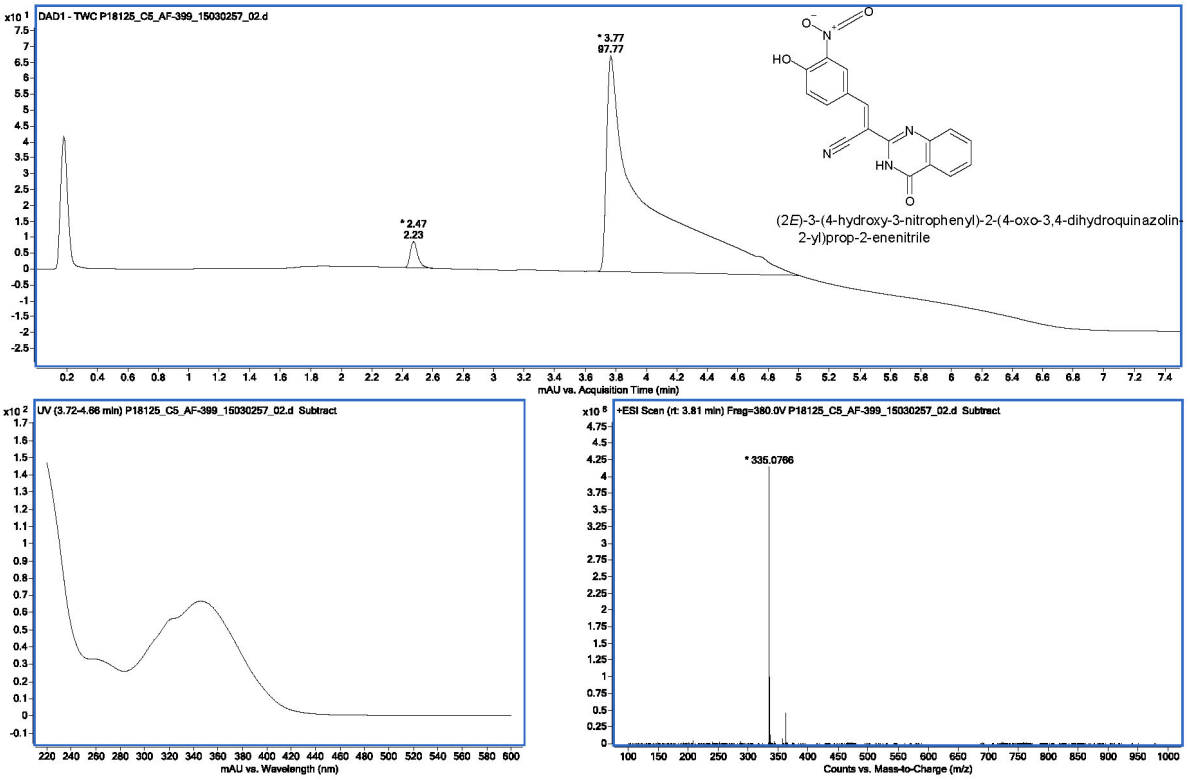
C)


DAD1 - TWC P18125_C3_AI-204_43362778.d
2,2'-[(2E)-but-2-ene-1,4-diyldisulfanediyl]di(pyridine-3-carboxylic acid)
* 3.82
97.92
* 2.00
0.24
* 4.76
1.84
mAU vs. Acquisition Time (min)
UV (3.77-4.30 min) P18125_C3_AI-204_43362778.d Subtract
mAU vs. Wavelength (nm)
+ESI Scan (rt: 3.85-3.95 min, 19 scans) Frag=380.0V P18125_C3_AI-204_43362778_02.d Subtract
363.0466
208.0433
162.0364
305.0869
Counts vs. Mass-to-Charge (m/z)

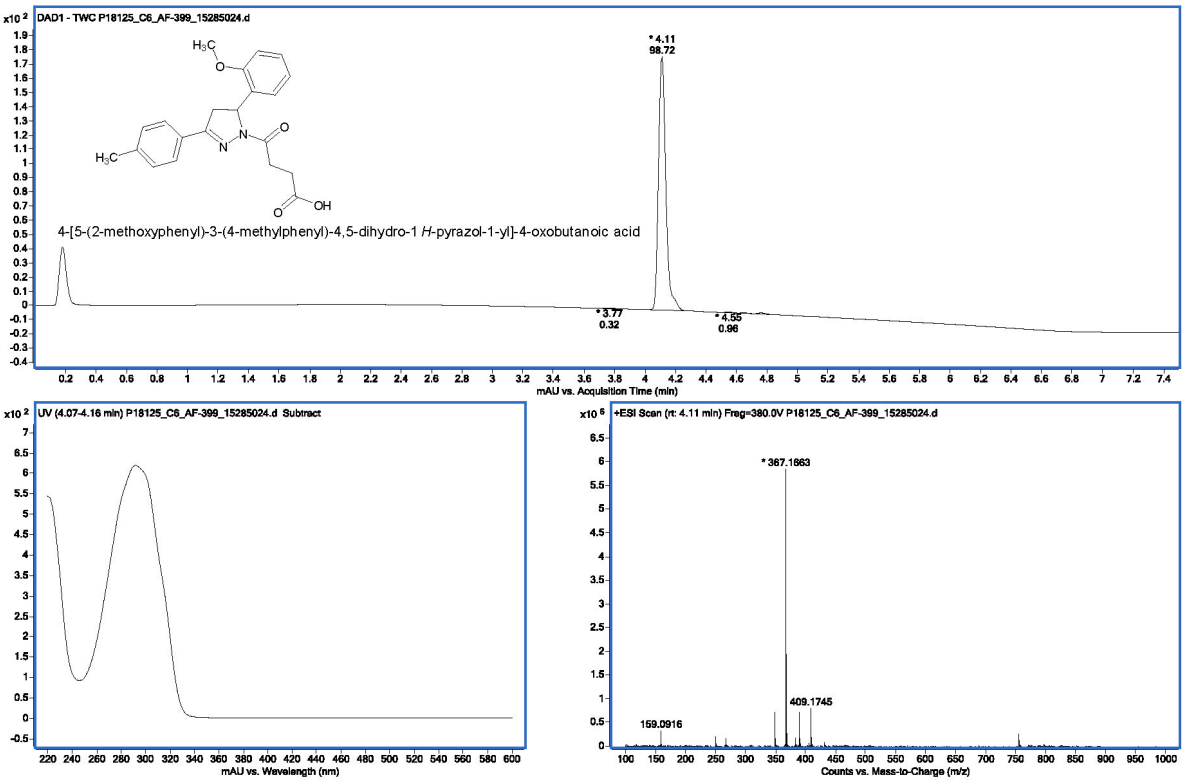
D)


DAD1 - TWC P18125_C4_AG-690_11836029.d
[(4-bromo-8-chloronaphthalen-1-yl)sulfanyl]acetic acid
* 4.24
98.43
* 3.84
1.07
* 4.76
0.50
mAU vs. Acquisition Time (min)
UV (4.20-4.32 min) P18125_C4_AG-690_11836029.d Subtract
mAU vs. Wavelength (nm)
+ESI Scan (rt: 4.26-4.38 min, 24 scans) Frag=380.0V P18125_C4_AG-690_11836029.d Subtract
294.9418
353.0458
208.0427
Counts vs. Mass-to-Charge (m/z)

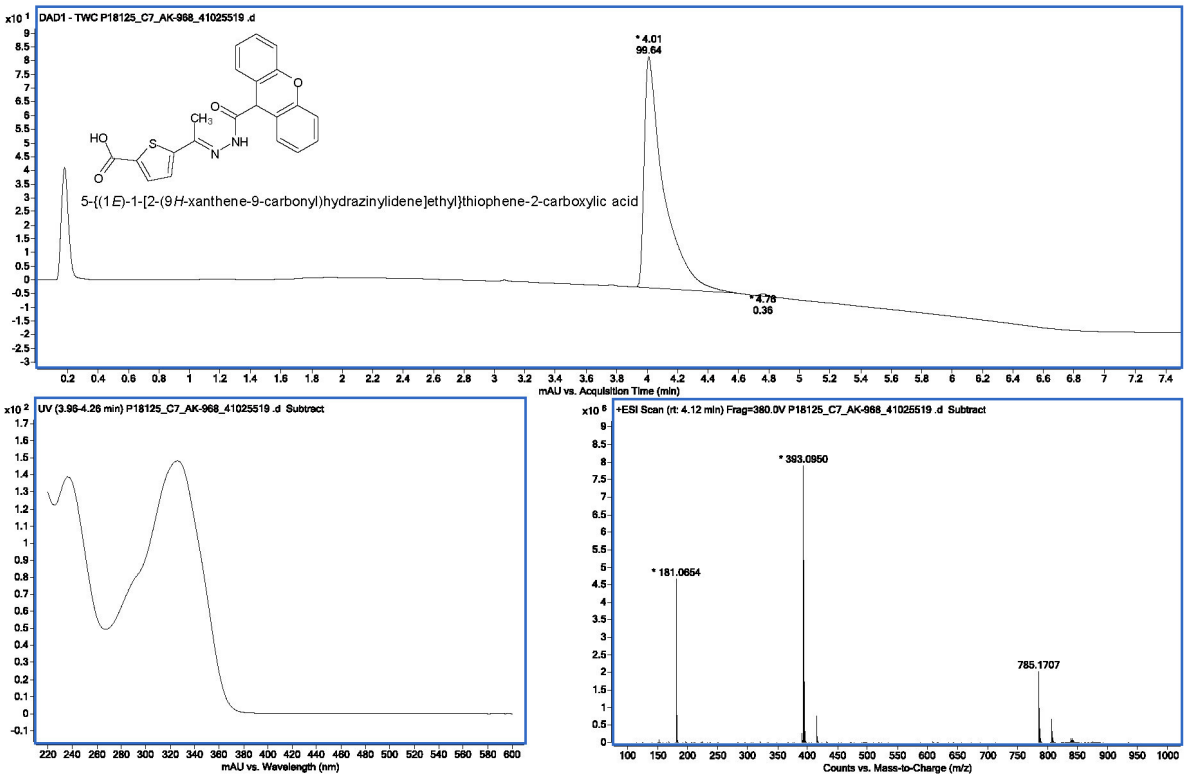
E)


DAD1 - TWC P18125_C5_AF-399_15030257_02.d
* 3.77
97.77
* 2.47
2.23
(2E)-3-(4-hydroxy-3-nitrophenyl)-2-(4-oxo-3,4-dihydroquinazolin-2-yl)prop-2-enenitrile
mAU vs. Acquisition Time (min)
UV (3.72-4.66 min) P18125_C5_AF-399_15030257_02.d Subtract
mAU vs. Wavelength (nm)
+ESI Scan (rt: 3.81 min) Frag=380.0V P18125_C5_AF-399_15030257_02.d Subtract
* 335.0766
Counts vs. Mass-to-Charge (m/z)

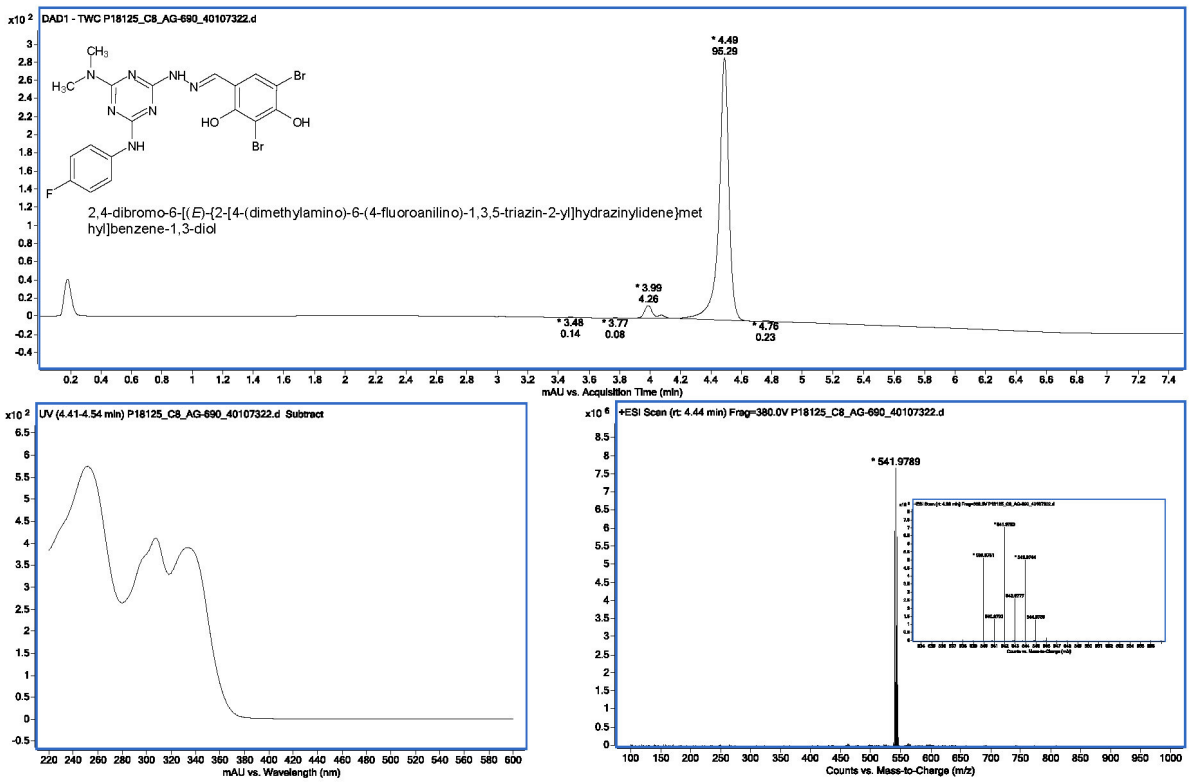
F)


DAD1 - TWC P18125_C6_AF-399_15285024.d
* 4.11
98.72
* 3.77
0.32
* 4.55
0.98
4-[5-(2-methoxyphenyl)-3-(4-methylphenyl)-4,5-dihydro-1 H-pyrazol-1-yl]-4-oxobutanoic acid
mAU vs. Acquisition Time (min)
UV (4.07-4.16 min) P18125_C6_AF-399_15285024.d Subtract
mAU vs. Wavelength (nm)
+ESI Scan (rt: 4.11 min) Frag=380.0V P18125_C6_AF-399_15285024.d
* 367.1663
409.1745
159.0916
Counts vs. Mass-to-Charge (m/z)

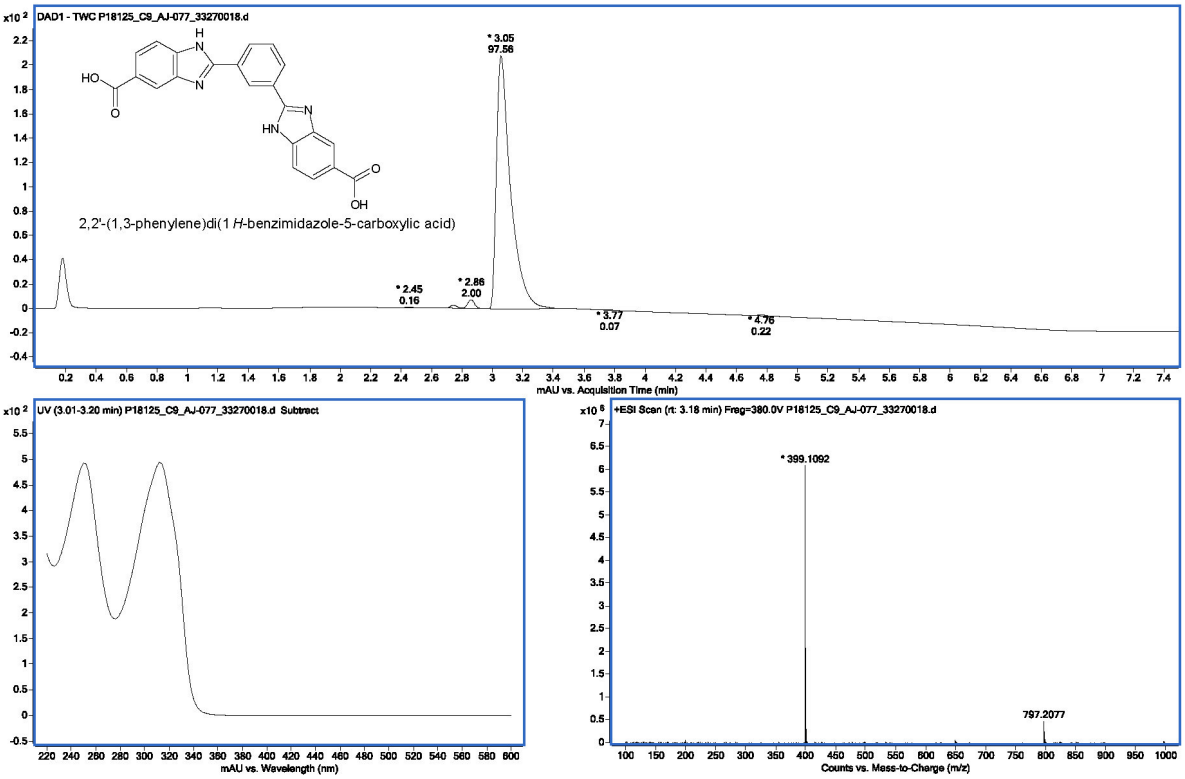
G)


DAD1 - TWC P18125_C7_AK-968_41025519 .d
5-{(1E)-1-[2-(9H-xanthene-9-carbonyl)hydrazinylidene]ethyl}thiophene-2-carboxylic acid
* 4.01
99.64
* 4.78
0.36
mAU vs. Acquisition Time (min)
UV (3.96-4.26 min) P18125_C7_AK-968_41025519 .d Subtract
mAU vs. Wavelength (nm)
+ESI Scan (rt: 4.12 min) Frag=380.0V P18125_C7_AK-968_41025519 .d Subtract
* 393.0950
* 181.0654
785.1707
Counts vs. Mass-to-Charge (m/z)

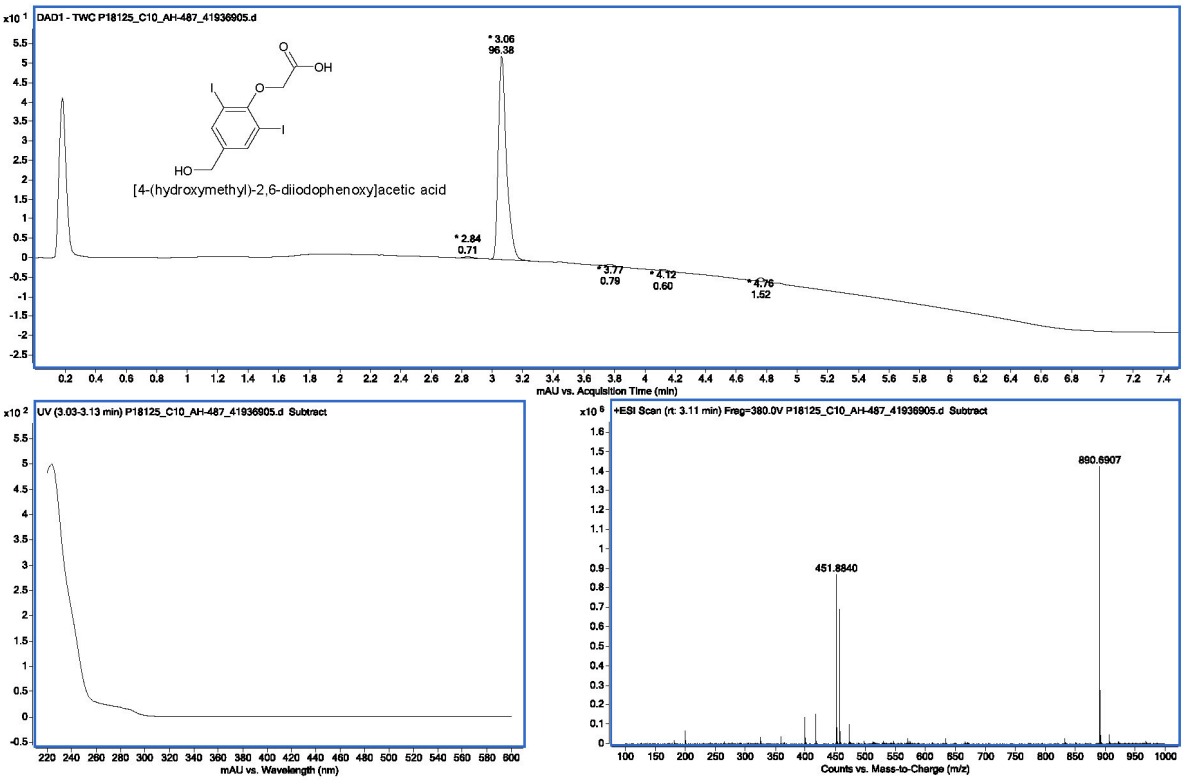
H)


DAD1 - TWC P18125_C8_AG-690_40107322.d
2,4-dibromo-6-[(E)-{2-[4-(dimethylamino)-6-(4-fluoroanilino)-1,3,5-triazin-2-yl]hydrazinylidene}methyl]benzene-1,3-diol
* 4.49
95.29
* 3.99
4.26
* 3.48
0.14
* 3.77
0.08
* 4.76
0.23
mAU vs. Acquisition Time (min)
UV (4.41-4.54 min) P18125_C8_AG-690_40107322.d Subtract
mAU vs. Wavelength (nm)
+ESI Scan (rt: 4.44 min) Frag=380.0V P18125_C8_AG-690_40107322.d
* 541.9789
Counts vs. Mass-to-Charge (m/z)

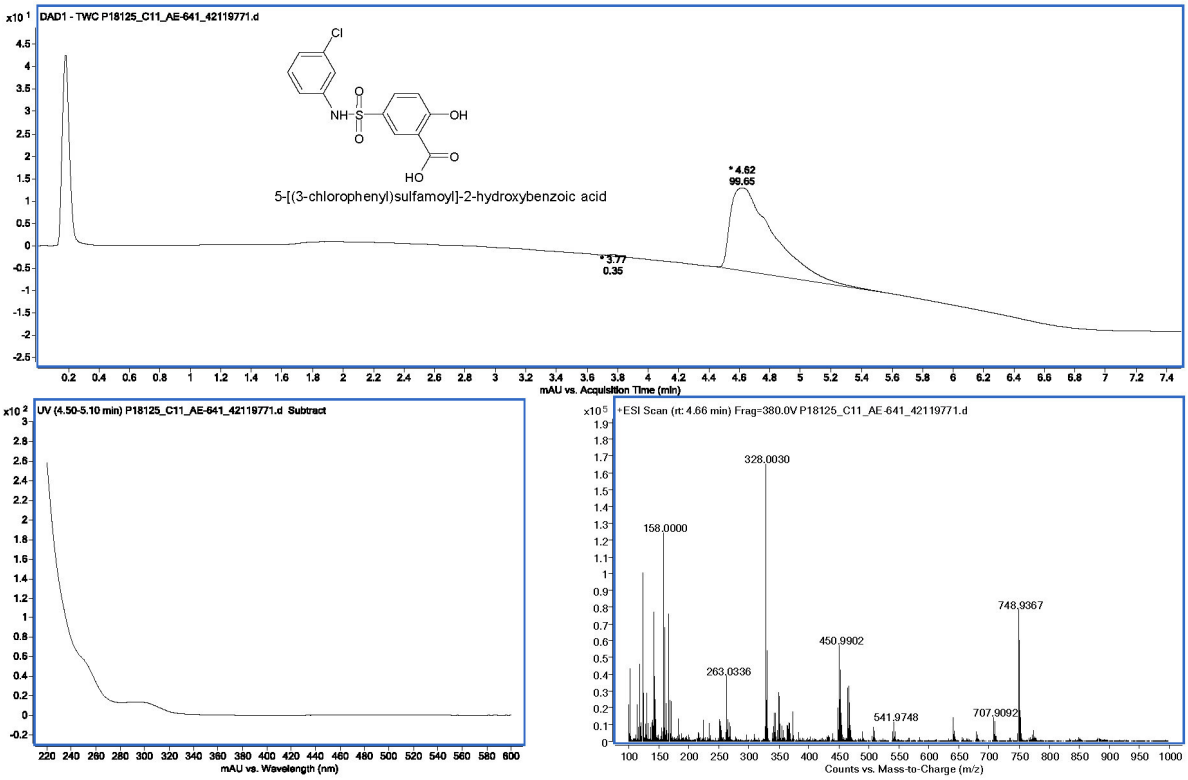
I)


DAD1 - TWC P18125_C9_AJ-077_33270018.d
2,2'-(1,3-phenylene)di(1 H-benzimidazole-5-carboxylic acid)
* 3.05
97.56
* 2.45
0.16
* 2.86
2.00
* 3.77
0.07
* 4.76
0.22
mAU vs. Acquisition Time (min)
UV (3.01-3.20 min) P18125_C9_AJ-077_33270018.d Subtract
mAU vs. Wavelength (nm)
+ESI Scan (rt: 3.18 min) Freq=380.0V P18125_C9_AJ-077_33270018.d
* 399.1092
797.2077
Counts vs. Mass-to-Charge (m/z)

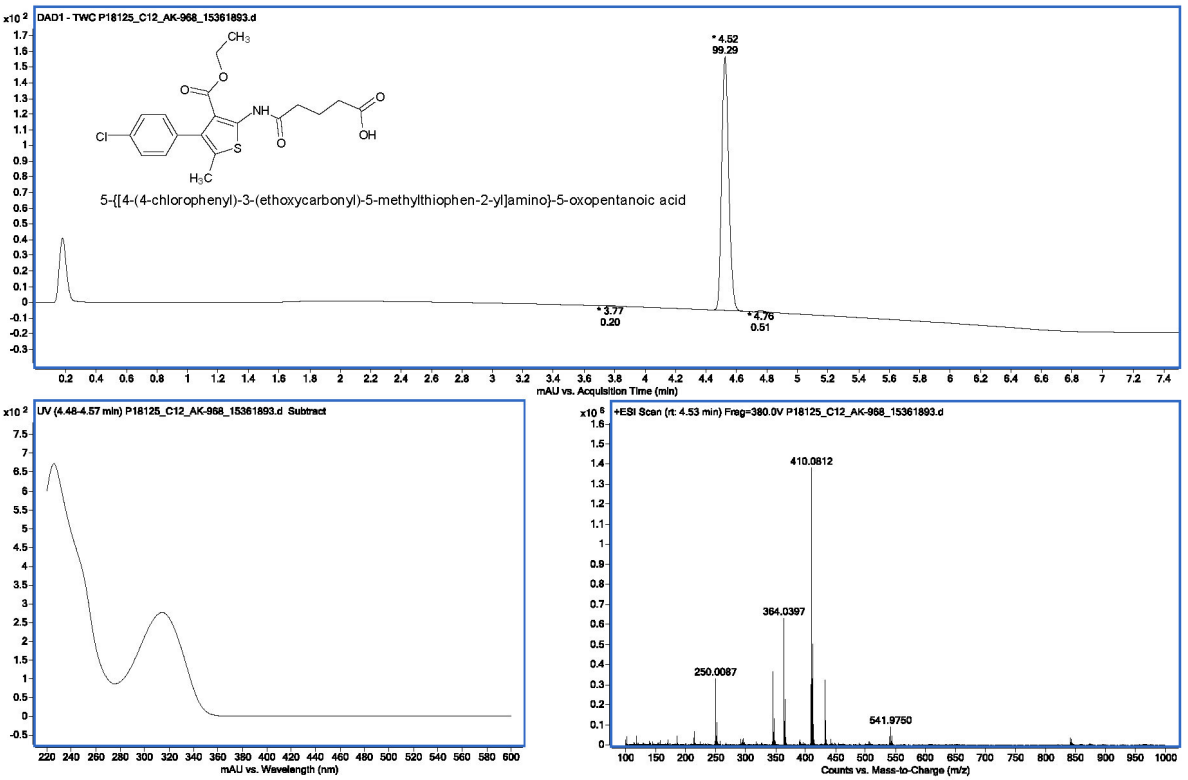
J)


DAD1 - TWC P18125_C10_AH-487_41936905.d
[4-(hydroxymethyl)-2,6-diiodophenoxy]acetic acid
* 3.06
96.38
* 2.84
0.71
* 3.77
0.79
* 4.12
0.60
* 4.76
1.52
mAU vs. Acquisition Time (min)
UV (3.03-3.13 min) P18125_C10_AH-487_41936905.d Subtract
mAU vs. Wavelength (nm)
+ESI Scan (rt: 3.11 min) Freq=380.0V P18125_C10_AH-487_41936905.d Subtract
890.6907
451.8840
Counts vs. Mass-to-Charge (m/z)

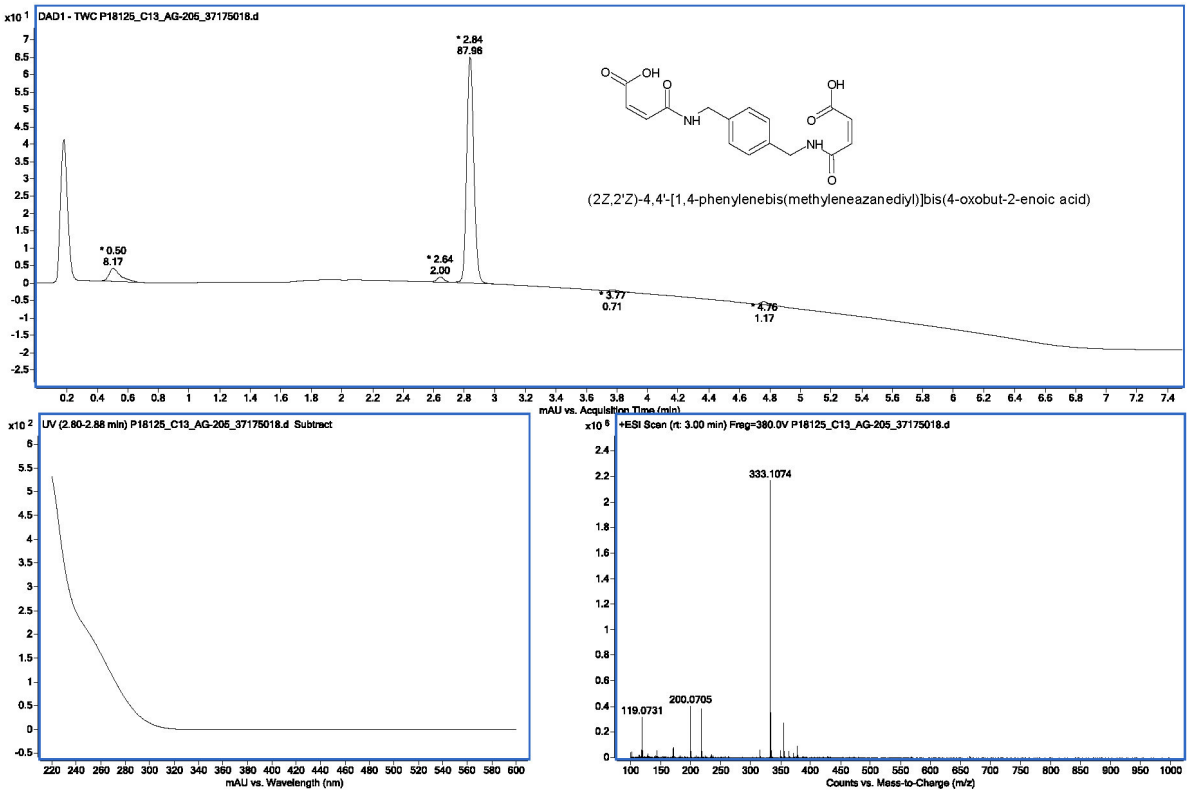
K)


DAD1 - TWC P18125_C11_AE-641_42119771.d
5-[(3-chlorophenyl)sulfamoyl]-2-hydroxybenzoic acid
* 4.62
99.65
* 3.77
0.35
mAU vs. Acquisition Time (min)
UV (4.50-5.10 min) P18125_C11_AE-641_42119771.d Subtract
mAU vs. Wavelength (nm)
+ESI Scan (rt: 4.66 min) Frag=380.0V P18125_C11_AE-641_42119771.d
328.0030
158.0000
748.9367
450.9902
263.0336
541.9748
707.9092
Counts vs. Mass-to-Charge (m/z)

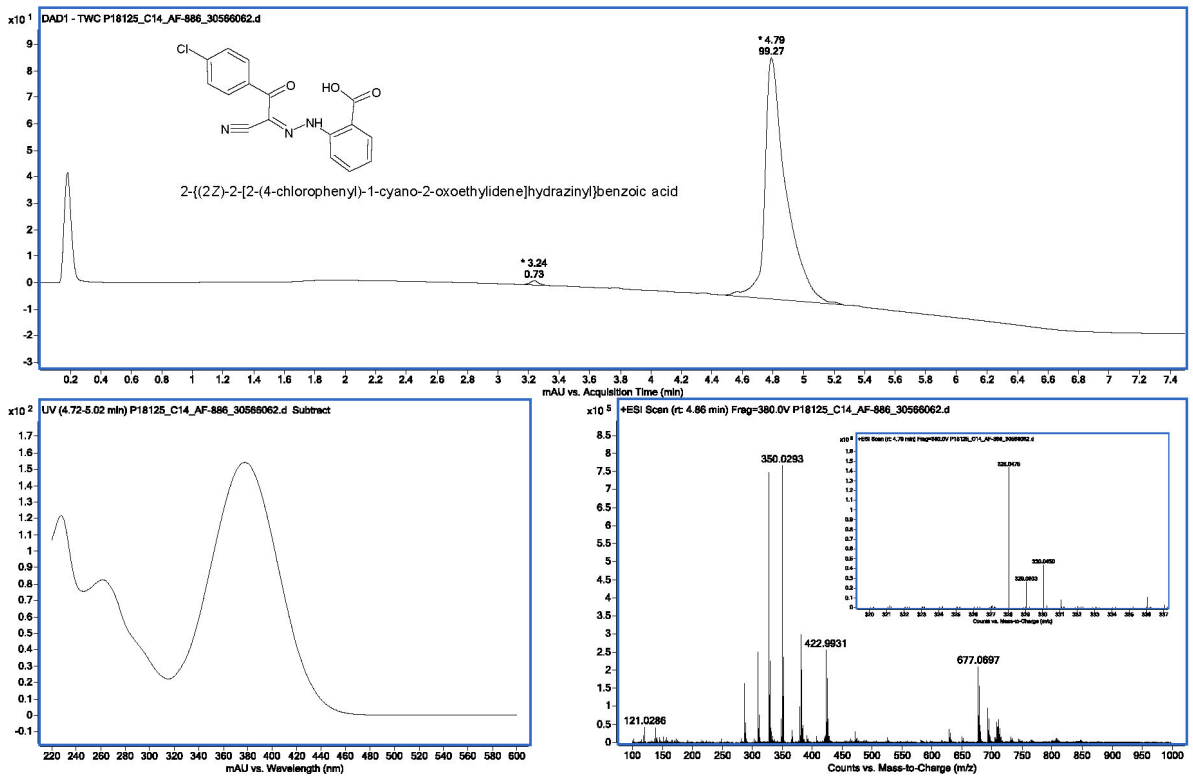
L)


DAD1 - TWC P18125_C12_AK-968_15361893.d
5-{[4-(4-chlorophenyl)-3-(ethoxycarbonyl)-5-methylthiophen-2-yl]amino}-5-oxopentanoic acid
* 4.52
99.29
* 3.77
0.20
* 4.76
0.51
mAU vs. Acquisition Time (min)
UV (4.48-4.57 min) P18125_C12_AK-968_15361893.d Subtract
mAU vs. Wavelength (nm)
+ESI Scan (rt: 4.53 min) Freq=380.0V P18125_C12_AK-968_15361893.d
410.0812
364.0397
250.0087
541.9750
Counts vs. Mass-to-Charge (m/z)

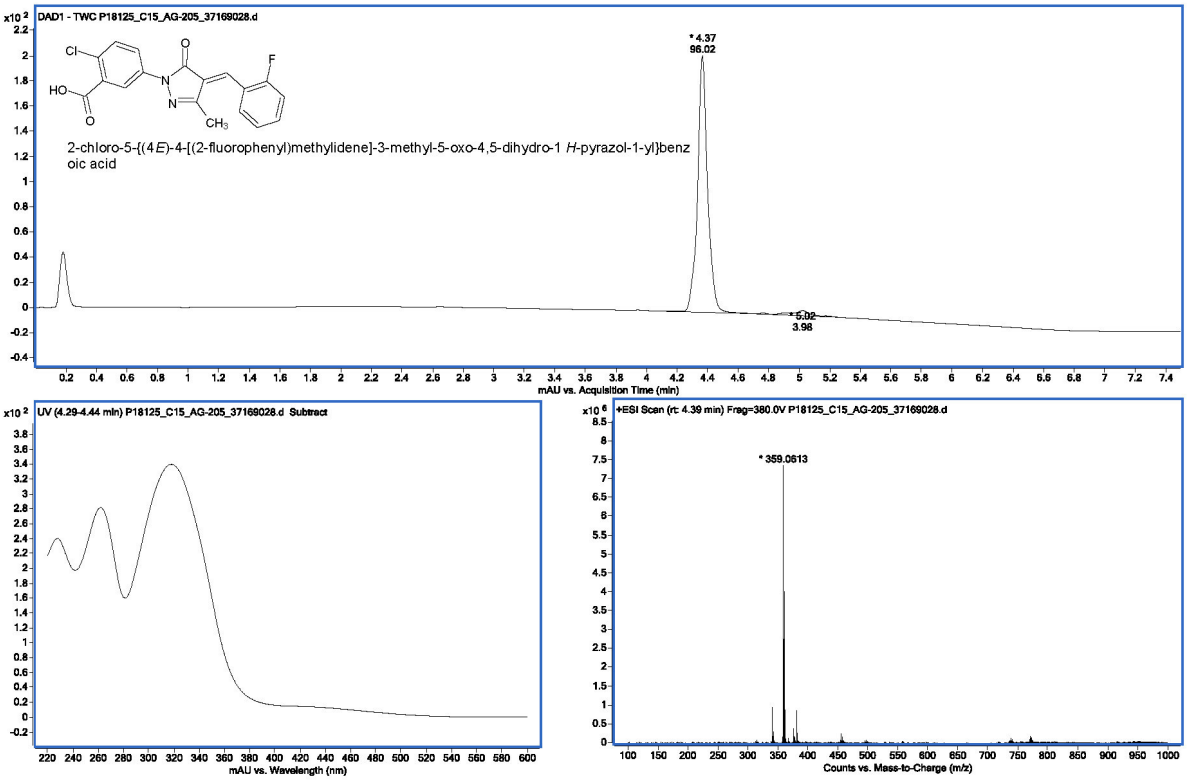
M)


DAD1 - TWC P18125_C13_AG-205_37175018.d
* 2.84
87.96
* 0.50
8.17
* 2.64
2.00
* 3.77
0.71
* 4.76
1.17
(2Z,2'Z)-4,4'-[1,4-phenylenebis(methyleneazanediyl)]bis(4-oxobut-2-enoic acid)
mAU vs. Acquisition Time (min)
UV (2.80-2.88 min) P18125_C13_AG-205_37175018.d Subtract
mAU vs. Wavelength (nm)
+ESI Scan (rt: 3.00 min) Freg=380.0V P18125_C13_AG-205_37175018.d
333.1074
200.0705
119.0731
Counts vs. Mass-to-Charge (m/z)

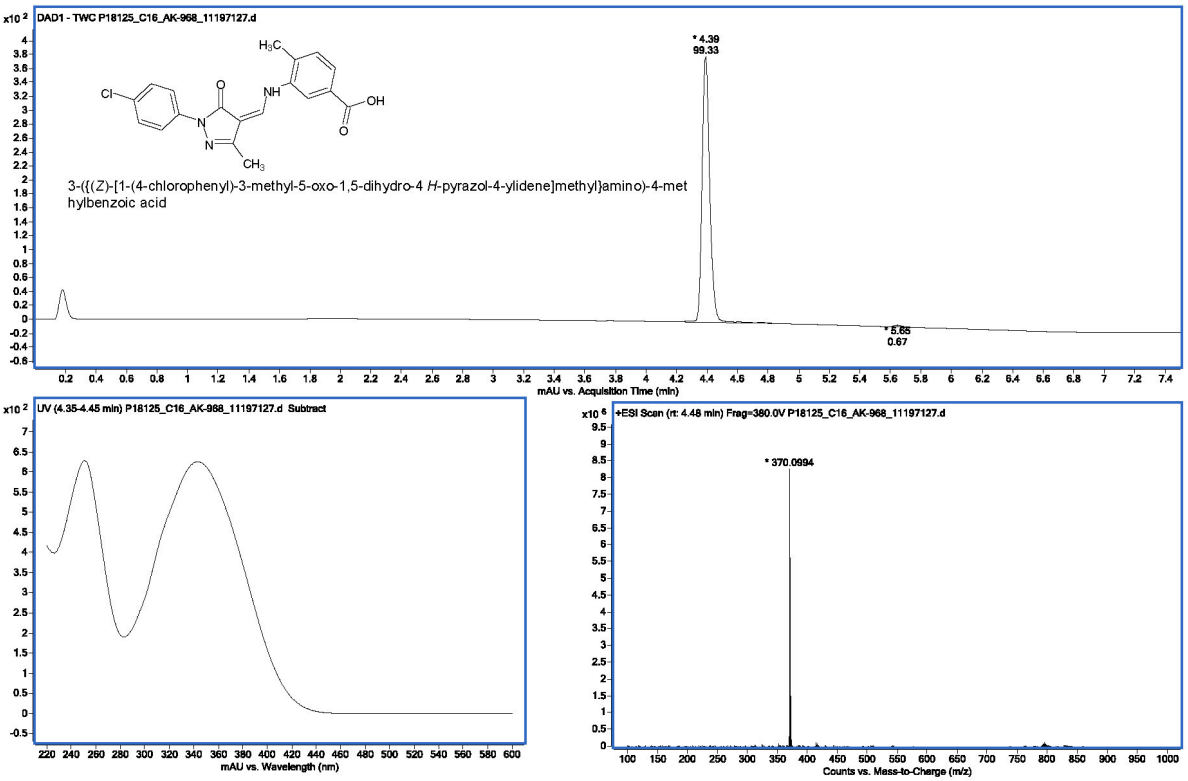
N)


DAD1 - TWC P18125_C14_AF-886_30566062.d
* 4.79
99.27
* 3.24
0.73
2-{(2Z)-2-[2-(4-chlorophenyl)-1-cyano-2-oxoethylidene]hydrazinyl}benzoic acid
mAU vs. Acquisition Time (min)
UV (4.72-5.02 min) P18125_C14_AF-886_30566062.d Subtract
mAU vs. Wavelength (nm)
+ESI Scan (rt: 4.86 min) Freg=380.0V P18125_C14_AF-886_30566062.d
350.0293
422.9931
677.0697
121.0286
Counts vs. Mass-to-Charge (m/z)

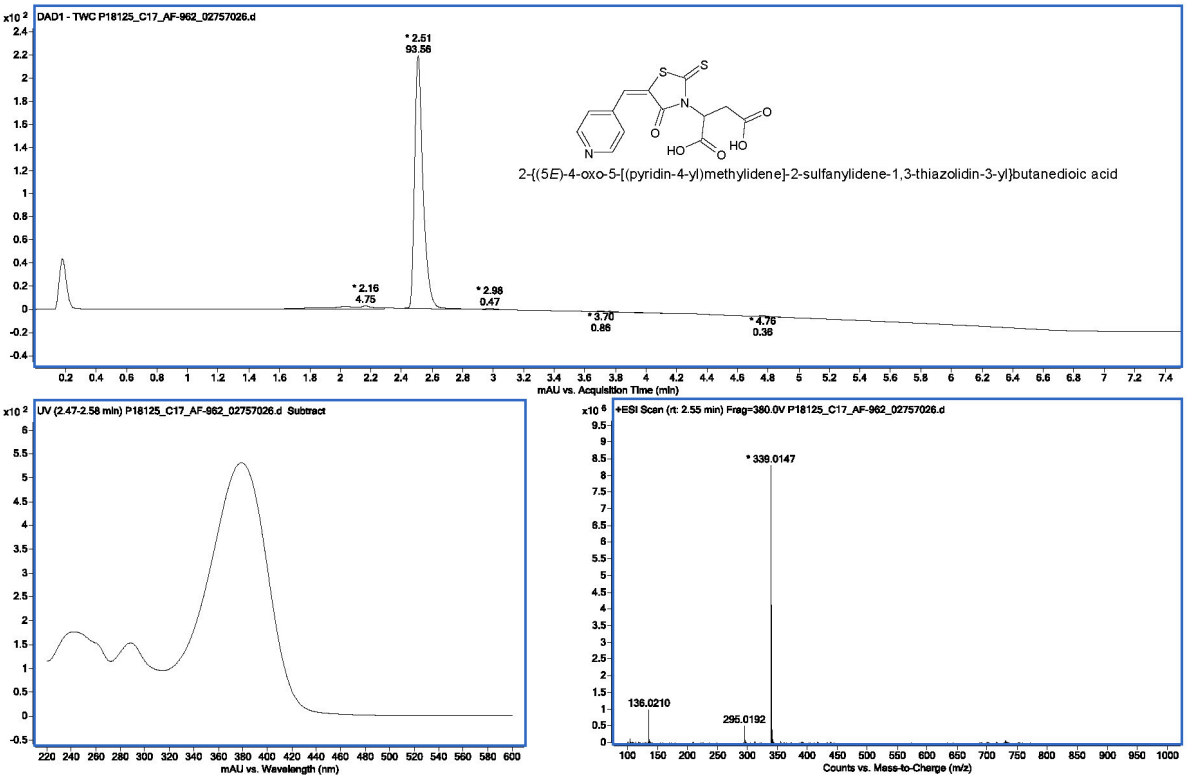
O)


DAD1 - TWC P18125_C15_AG-205_37169028.d
2-chloro-5-{(4E)-4-[(2-fluorophenyl)methylidene]-3-methyl-5-oxo-4,5-dihydro-1 H-pyrazol-1-yl}benz
oic acid
* 4.37
96.02
* 5.02
3.98
mAU vs. Acquisition Time (min)
UV (4.29-4.44 min) P18125_C15_AG-205_37169028.d Subtract
mAU vs. Wavelength (nm)
+ESI Scan (rt: 4.39 min) Frag=380.0V P18125_C15_AG-205_37169028.d
* 359.0513
Counts vs. Mass-to-Charge (m/z)

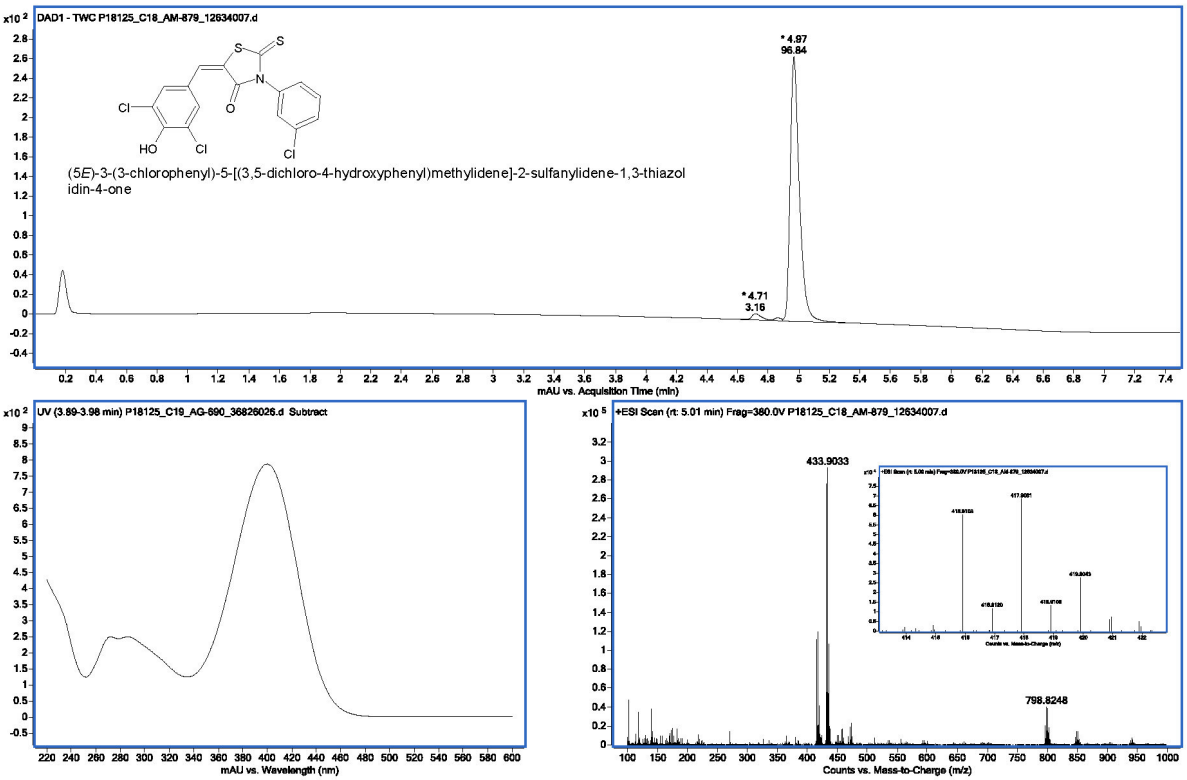
P)


DAD1 - TWC P18125_C16_AK-968_11197127.d
3-({(Z)-[1-(4-chlorophenyl)-3-methyl-5-oxo-1,5-dihydro-4 H-pyrazol-4-ylidene]methyl}amino)-4-met
hylbenzoic acid
* 4.39
99.33
* 5.65
0.67
mAU vs. Acquisition Time (min)
UV (4.35-4.45 min) P18125_C16_AK-968_11197127.d Subtract
mAU vs. Wavelength (nm)
+ESI Scan (rt: 4.48 min) Frag=380.0V P18125_C16_AK-968_11197127.d
* 370.0994
Counts vs. Mass-to-Charge (m/z)

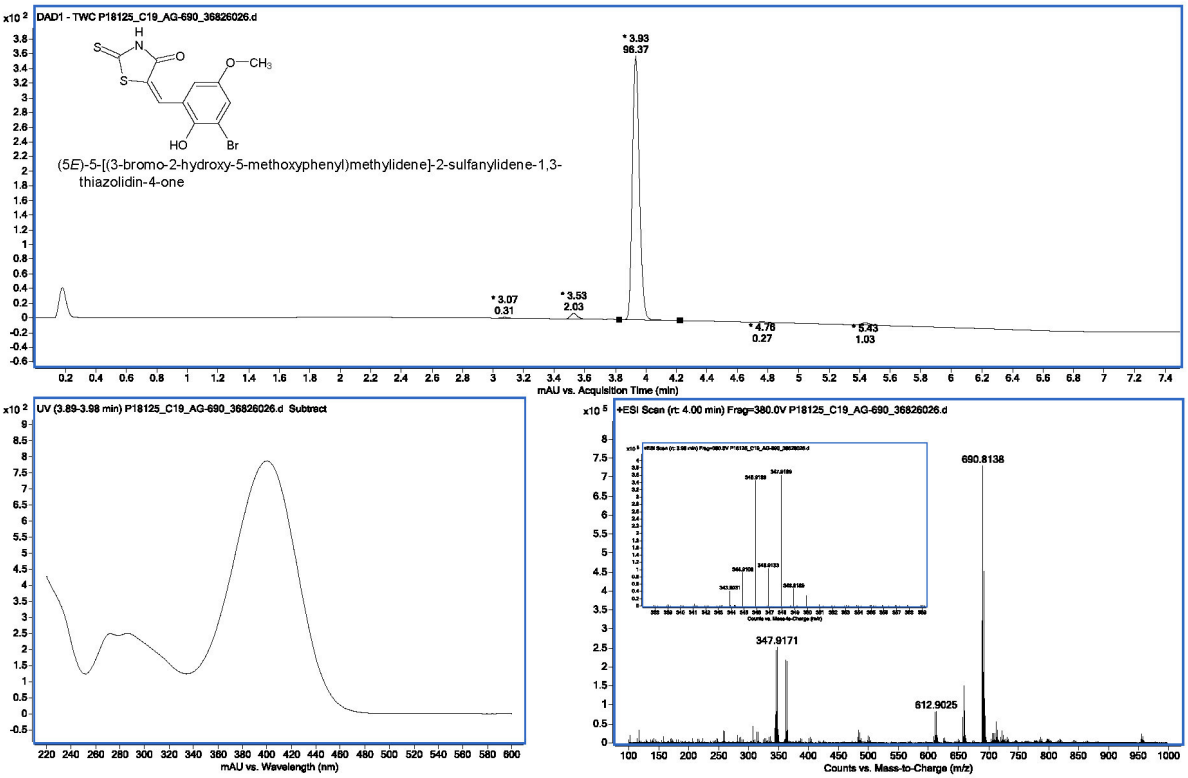
Q)


DAD1 - TWC P18125_C17_AF-962_02757026.d
* 2.51
93.56
* 2.16
4.75
* 2.98
0.47
* 3.70
0.86
* 4.76
0.36
mAU vs. Acquisition Time (min)
2-[(5E)-4-oxo-5-[(pyridin-4-yl)methylidene]-2-sulfanylidene-1,3-thiazolidin-3-yl]butanedioic acid
UV (2.47-2.58 min) P18125_C17_AF-962_02757026.d Subtract
mAU vs. Wavelength (nm)
+ESI Scan (rt: 2.55 min) Frag=380.0V P18125_C17_AF-962_02757026.d
* 339.0147
136.0210
295.0192
Counts vs. Mass-to-Charge (m/z)

R)


DAD1 - TWC P18125_C18_AM-879_12634007.d
* 4.97
96.84
* 4.71
3.16
mAU vs. Acquisition Time (min)
(5E)-3-(3-chlorophenyl)-5-[(3,5-dichloro-4-hydroxyphenyl)methylidene]-2-sulfanylidene-1,3-thiazol
idin-4-one
UV (3.89-3.98 min) P18125_C19_AG-890_36826026.d Subtract
mAU vs. Wavelength (nm)
+ESI Scan (rt: 5.01 min) Frag=380.0V P18125_C18_AM-879_12634007.d
433.9033
798.8248
Counts vs. Mass-to-Charge (m/z)

S)


DAD1 - TWC P18125_C19_AG-690_36826026.d
(5E)-5-[(3-bromo-2-hydroxy-5-methoxyphenyl)methylidene]-2-sulfanylidene-1,3-thiazolidin-4-one
* 3.93
96.37
* 3.07
0.31
* 3.53
2.03
* 4.76
0.27
* 5.43
1.03
mAU vs. Acquisition Time (min)
UV (3.89-3.98 min) P18125_C19_AG-690_36826026.d Subtract
mAU vs. Wavelength (nm)
+ESI Scan (rt: 4.00 min) Frag=380.0V P18125_C19_AG-690_36826026.d
690.8138
347.9171
612.9025
Counts vs. Mass-to-Charge (m/z)

T)

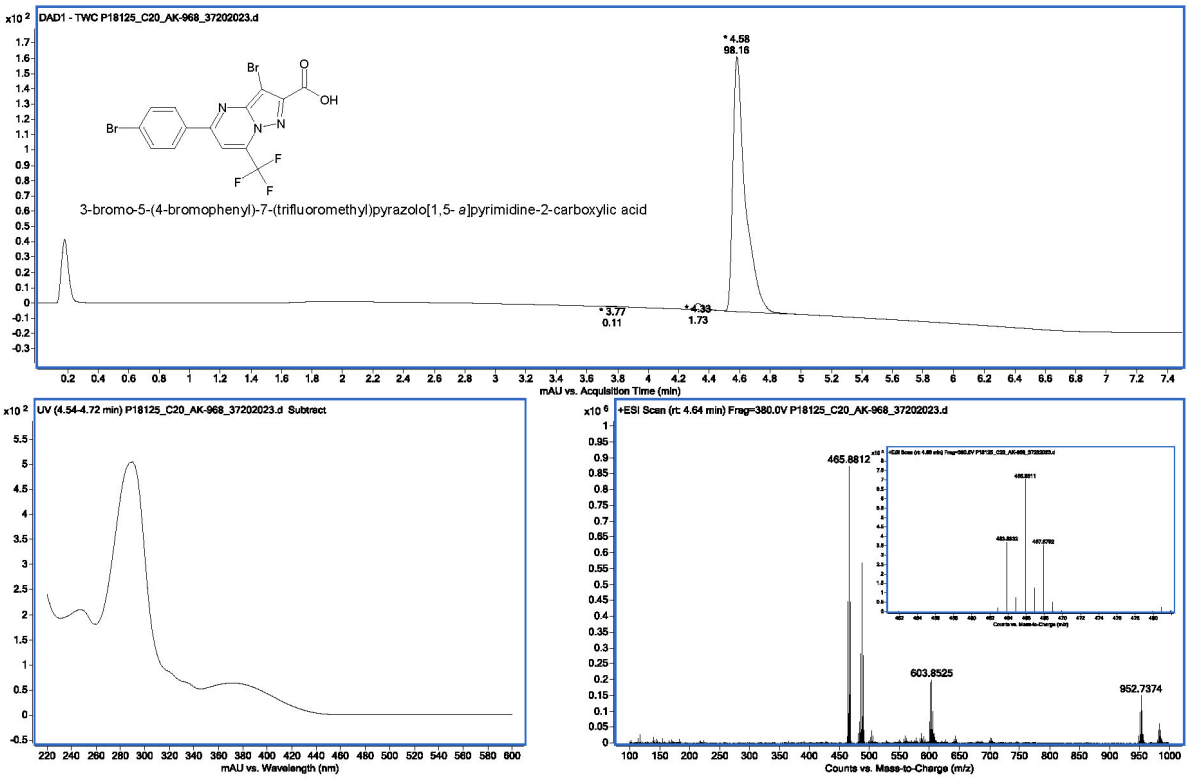
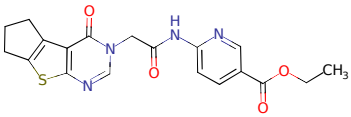
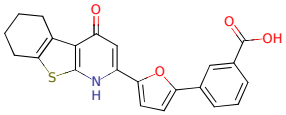

DAD1 - TWC P18125_C20_AK-968_37202023.d
3-bromo-5-(4-bromophenyl)-7-(trifluoromethyl)pyrazolo[1,5-a]pyrimidine-2-carboxylic acid
* 4.58
98.16
* 3.77
0.11
* 4.33
1.73
mAU vs. Acquisition Time (min)
UV (4.54-4.72 min) P18125_C20_AK-968_37202023.d Subtract
mAU vs. Wavelength (nm)
+ESI Scan (rt: 4.64 min) Frag=380.0V P18125_C20_AK-968_37202023.d
465.8812
603.8525
952.7374
Counts vs. Mass-to-Charge (m/z)

Figure S4. This figure shows the LC-UV/Vis and Q-TOF spectra for the 20 compounds whose bioactivities have been tested. Panels A-T refer to compounds **1**-**20**, respectively. The following information is provided for each tested compound: **a)** the UV/Vis relative area percentage (calculated without considering the solvent peak) on top of each chromatographic peak (with the corresponding retention time); **b)** the total wavelength chromatograms (220-600 nm) from the LC-UV/Vis; **c)** the UV/Vis spectrum from 220 to 600 nm; and **d)** the mass spectrum from 100 to 1000 m/z.

A)

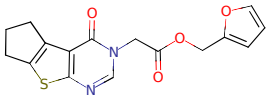


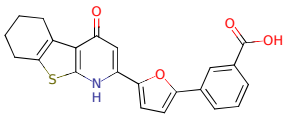


1 Reaxys RN: 26332088

Tanimoto: 0.32

B)

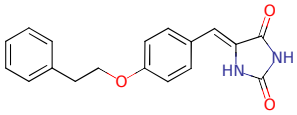


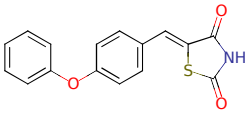


2 Reaxys RN: 26332088

Tanimoto: 0.32

C)

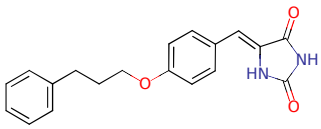


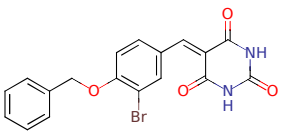


3 Reaxys RN: 9058052

Tanimoto: 0.43

D)

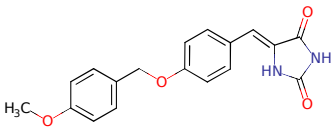


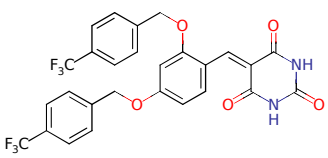


4 Reaxys RN: 21891758

Tanimoto: 0.46

E)

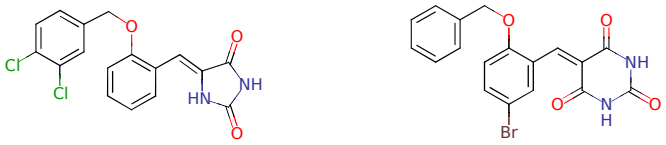




5 Reaxys RN: 21891764

Tanimoto: 0.43

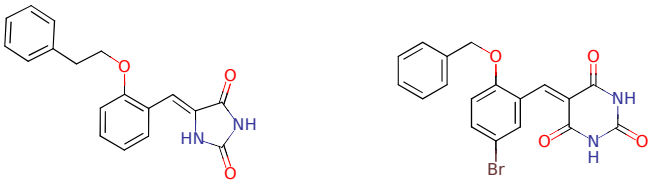
F)



6 Reaxys RN: 21891759

Tanimoto: 0.58

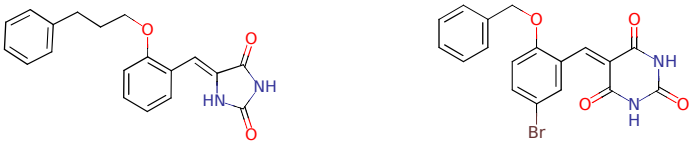
G)



7 Reaxys RN: 21891759

Tanimoto: 0.52

H)

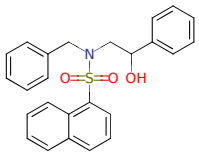


8 Reaxys RN: 21891759

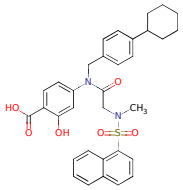
Tanimoto: 0.52

Figure S5. In each panel, a hit compound reported by Ma *et al.*[31] is represented in 2D, together with the most similar active compound that was used for clustering in the current study. In each case, the hit compound is labeled with its corresponding code in the original publication and the most similar active compound is labeled with its Reaxys[39] Registry Number. The Tanimoto similarity value resulting from the comparison of the OpenEyePath FP of both compounds is shown below. MarvinSketch[45] was used to draw the structures.

A)



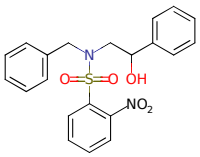
115



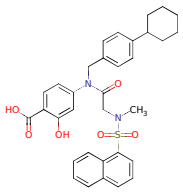
Reaxys RN: 21793264

Tanimoto: 0.41

B)



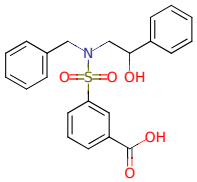
116



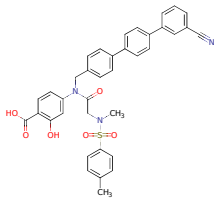
Reaxys RN: 21793264

Tanimoto: 0.34

C)



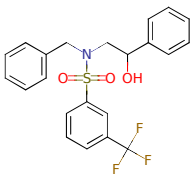
117



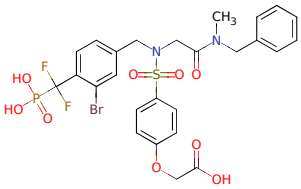
Reaxys RN: 22595724

Tanimoto: 0.33

D)



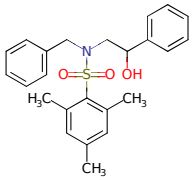
118



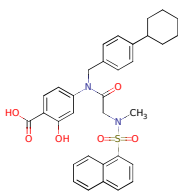
Reaxys RN: 10224258

Tanimoto: 0.33

E)



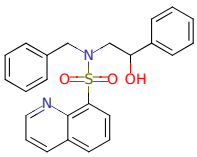
119

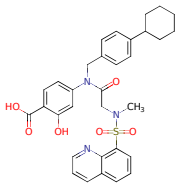


Reaxys RN: 21793264

Tanimoto: 0.28

F)

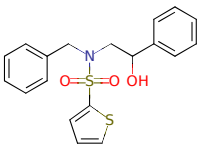


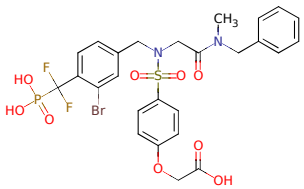


120 Reaxys RN: 21793265

Tanimoto: 0.41

G)

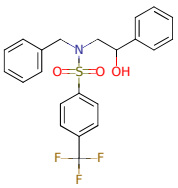


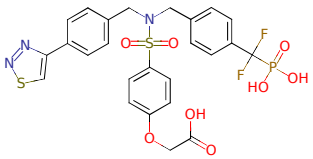


121 Reaxys RN: 10224258

Tanimoto: 0.25

H)

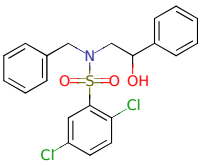


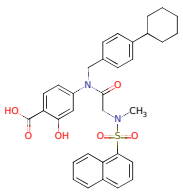


127 Reaxys RN: 10222316

Tanimoto: 0.38

I)

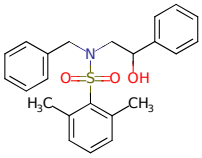


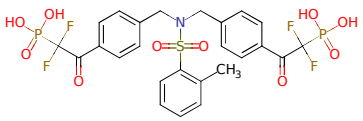


130 Reaxys RN: 21793264

Tanimoto: 0.34

J)



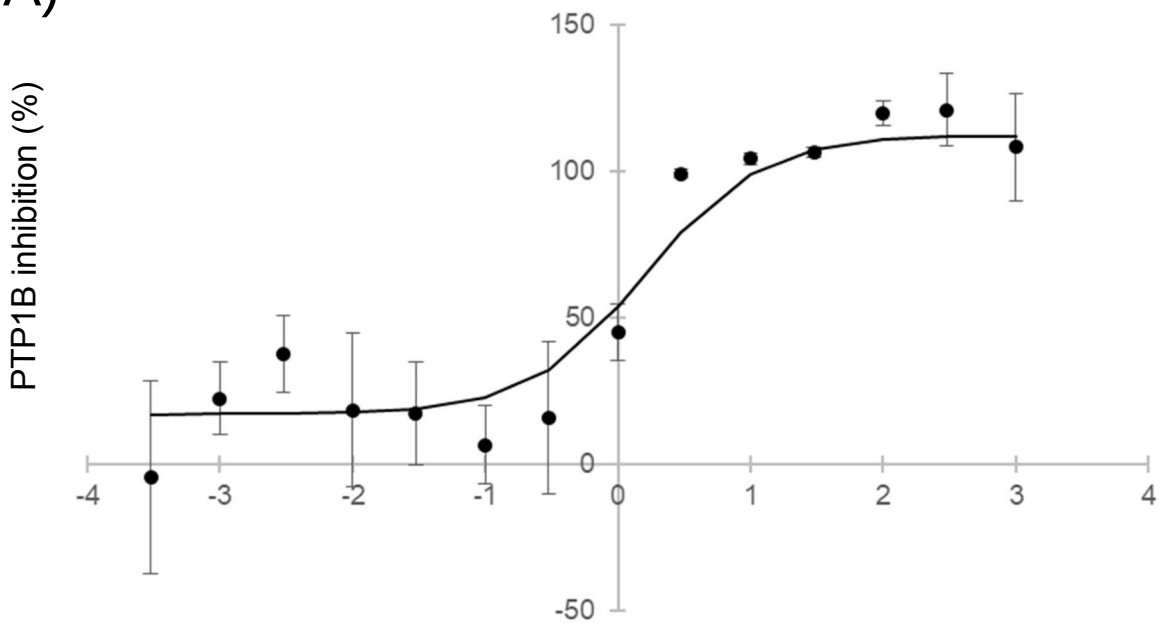


131 Reaxys RN: 25330566

Tanimoto: 0.31

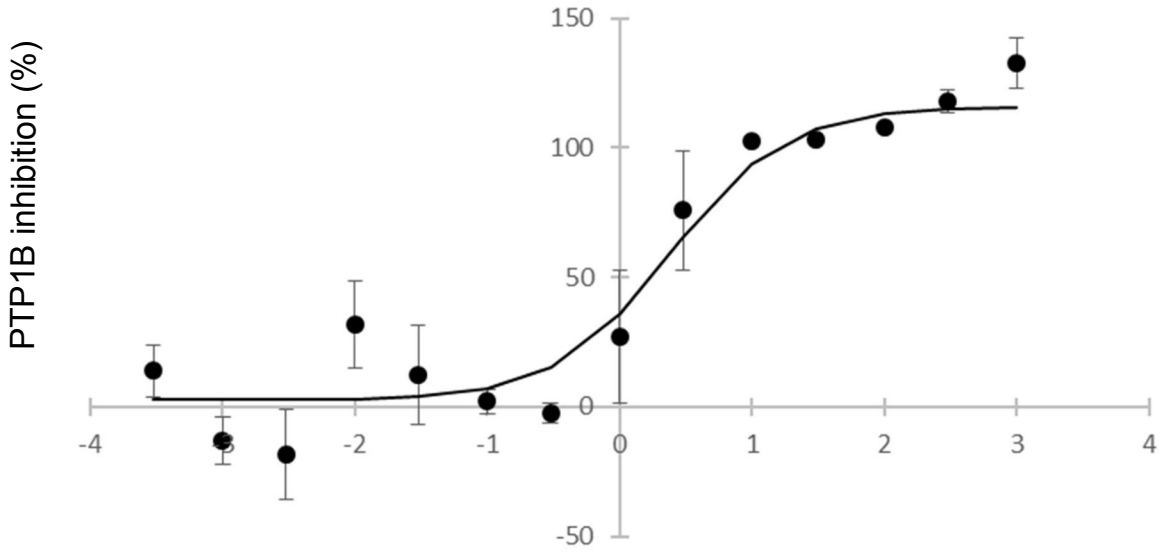
Figure S6. In each panel, a hit compound reported by Balaramnavar *et al*.[32] is represented in 2D, together with the most similar active compound that was used for clustering in the current study. In each case, the hit compound is labeled with its corresponding code in the original publication and the most similar active compound is labeled with its Reaxys[39] Registry Number. The Tanimoto similarity value resulting from the comparison of the OpenEyePath FP of both compounds is shown below. MarvinSketch[45] was used to draw the structures.

A)



[compound **5**] in μM	Log([compound **5**] in μM])	PTP1B inhibition (%)			
		Replicate 1	Replicate 2	Replicate 3	Mean ± standard deviation
1000	3	118.7	118.9	87.0	108.2 ± 18.4
300	2.5	125.2	107.0	130.7	121.0 ± 12.4
100	2	122.1	115.0	122.7	120.0 ± 4.3
30	1.5	107.9	106.5	104.6	106.3 ± 1.6
10	1	103.0	103.2	106.4	104.2 ± 1.9
3	0.5	100.7	98.6	97.8	99.0 ± 1.5
1	0	36.1	55.2	43.3	44.9 ± 9.6
0.3	-0.5	6.9	-4.4	45.2	15.9 ± 26.0
0.1	-1	8.4	18.9	-7.6	6.6 ± 13.3
0.03	-1.5	28.6	26.2	-3.1	17.3 ± 17.6
0.01	-2	-6.7	45.6	16.4	18.4 ± 26.2
0.003	-2.5	34.8	26.2	51.8	37.6 ± 13.0
0.001	-3	14.0	36.5	16.4	22.3 ± 12.4
0.0003	-3.5	-27.6	18.9	n/a	-4.3 ± 32.9

B)

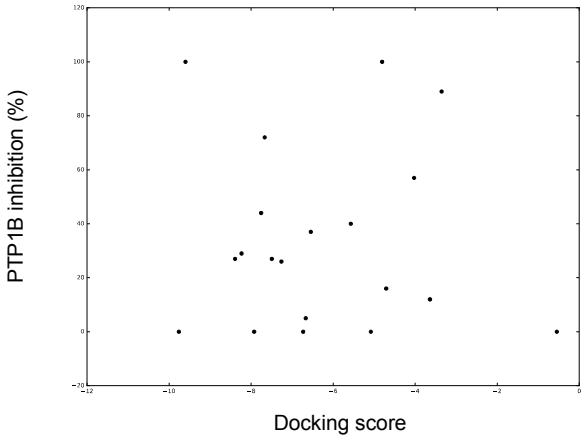


Log([compound **7**] in μM)

[compound **7**] in μM	Log([compound **7**] in μM])	PTP1B inhibition (%)			
		Replicate 1	Replicate 2	Replicate 3	Mean ± standard deviation
1000	3	137.3	121.8	139.2	132.7 ± 9.5
300	2.5	113.9	117.3	122.3	117.8 ± 4.2
100	2	108.9	108.6	107.0	108.2 ± 1.0
30	1.5	103.5	103.2	103.4	103.4 ± 0.1
10	1	102.5	102.3	102.5	102.4 ± 0.1
3	0.5	49.6	92.0	86.1	75.9 ± 23.0
1	0	18.3	7.2	55.7	27.1 ± 25.4
0.3	-0.5	1.3	-5.9	-2.4	-2.3 ± 3.6
0.1	-1	7.4	-1.0	-0.0	2.1 ± 4.6
0.03	-1.5	-9.0	28.0	18.4	12.5 ± 19.2
0.01	-2	32.4	14.4	47.9	31.6 ± 16.8
0.003	-2.5	1.8	-30.3	-25.8	-18.1 ± 17.4
0.001	-3	-17.8	-19.3	-2.7	-13.2 ± 9.2
0.0003	-3.5	6.7	21.1	n/a	13.9 ± 10.1

Figure S7. Panels A and B show the IC_{50} curves of compounds **5** and **7**, respectively, and the data used to represent them. In both cases, the percentages of inhibition of PTP1B for 14 different compound concentrations were determined using triplicates. IC_{50} values were calculated using a four parameter logistic regression.

A)



B)

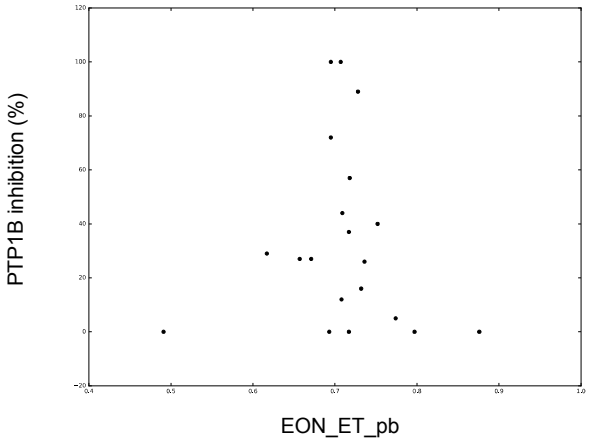
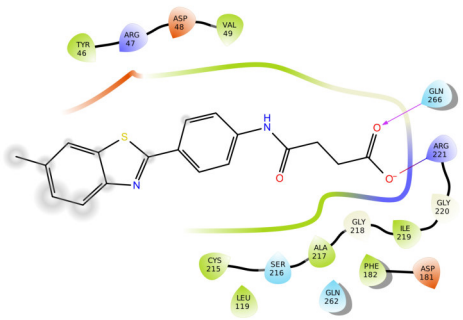


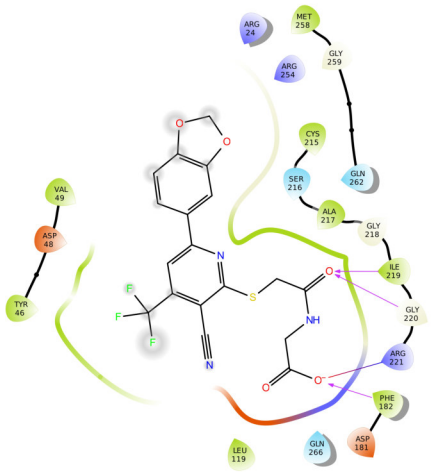
Figure S8. Representations of the docking score (panel A) and EON_ET_pb (panel B) against the percentages of PTP1B inhibition obtained for the 20 hit compounds at the concentration of 100 μM.

A)



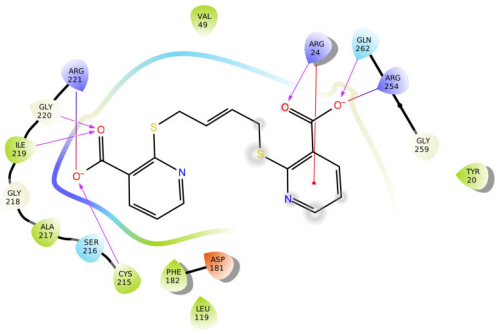
1 (AG-205/07930063)

B)



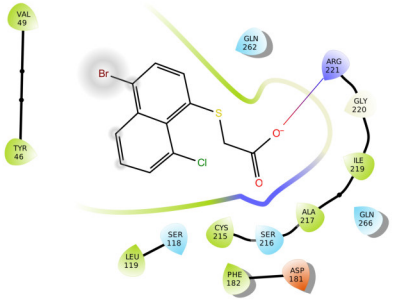
2 (AM-807/41462317)

C)



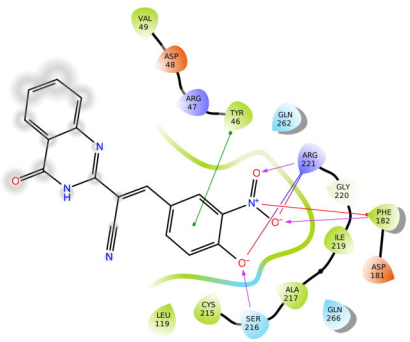
3 (AI-204/43362778)

D)



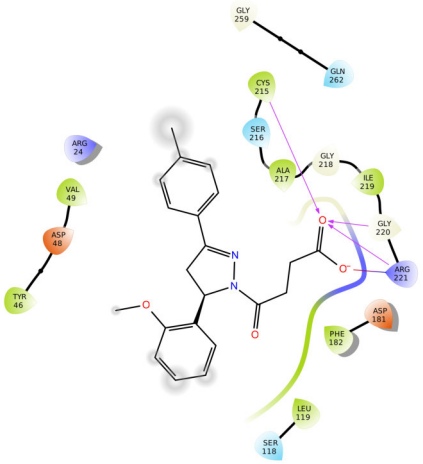
4 (AG-690/11836029)

E)



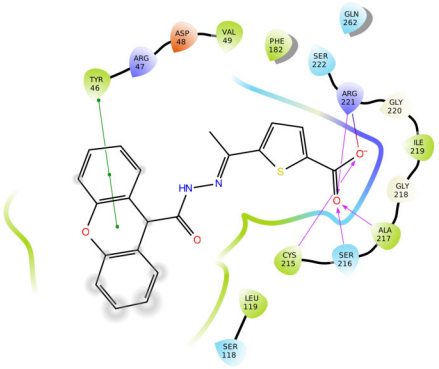
5 (AF-399/15030257)

F)



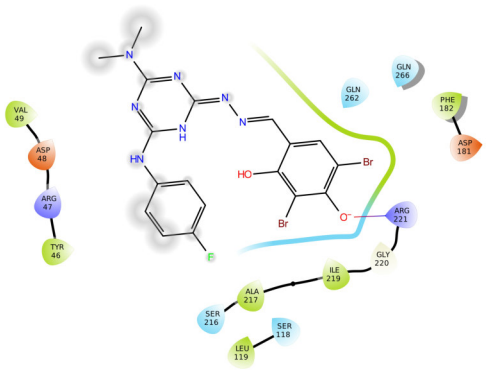
6 (AF-399/15285024)

G)



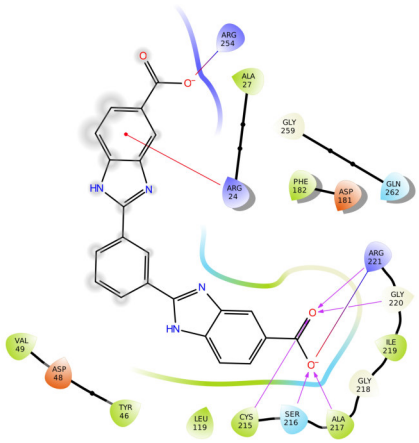
7 (AK-968/41025519)

H)



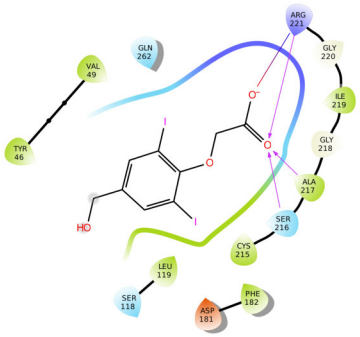
8 (AG-690/40107322)

I)



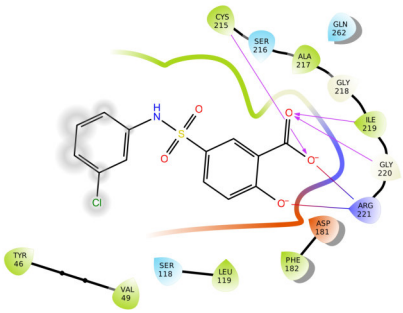
9 (AJ-077/33270018)

J)



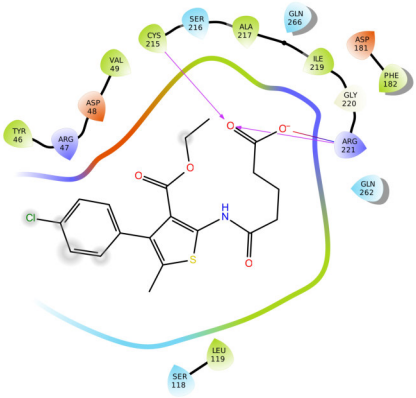
10 (AH-487/41936905)

K)



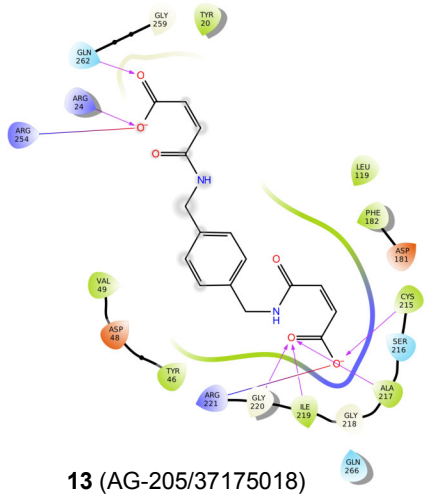
11 (AE-641/42119771)

L)



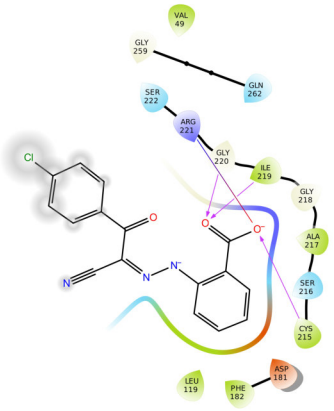
12 (AK-968/15361893)

M)



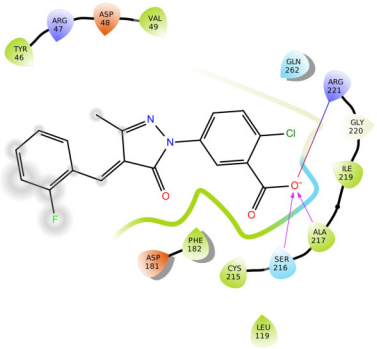
13 (AG-205/37175018)

N)



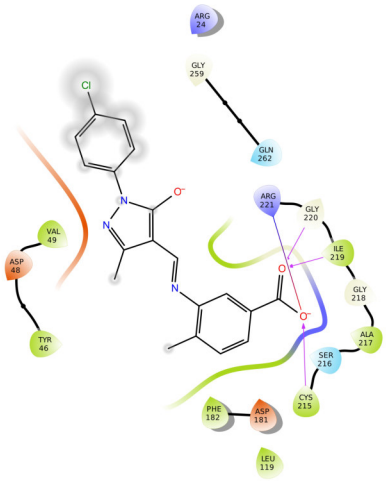
14 (AF-886/30566062)

O)



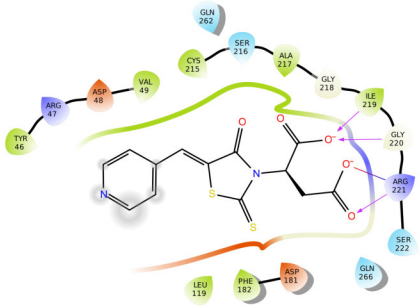
15 (AG-205/37169028)

P)



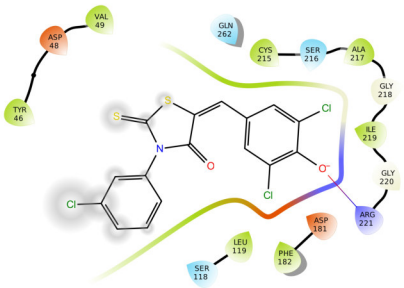
16 (AK-968/11197127)

Q)



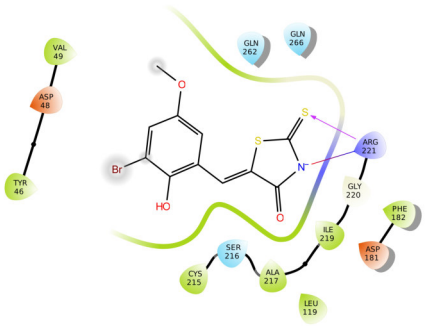
17 (AF-962/02757026)

R)



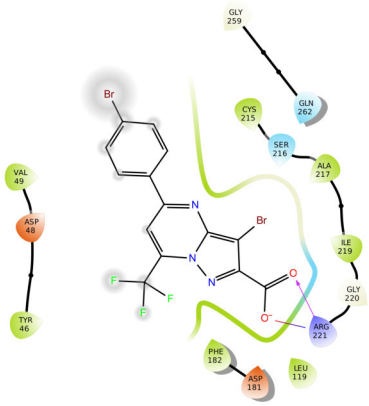
18 (AM-879/12634007)

S)



19 (AG-690/36826026)

T)



20 (AK-968/37202023)

Figure S9. Ligand interaction diagrams of the docked poses with the highest EON_ET_pb values of the 20 hit compounds. Negatively charged residues are colored in red, positively charged residues are colored in blue, hydrophobic residues are colored in green, polar residues are colored in cyan, glycine residues are colored in salmon, hydrogen bonds are represented as fuchsia arrows, salt bridges are represented as red and blue lines, π-π stacking interactions are erepresented as green lines, π-cation interactions are represented as red lines and atoms exposed to the solvent are circled in gray. The protein surface corresponding to the residues closest to the ligand is represented as a solid line colored by residue according to the presviously describred color scheme. This representation has been obtained with the help of Maestro[37] v10.7.

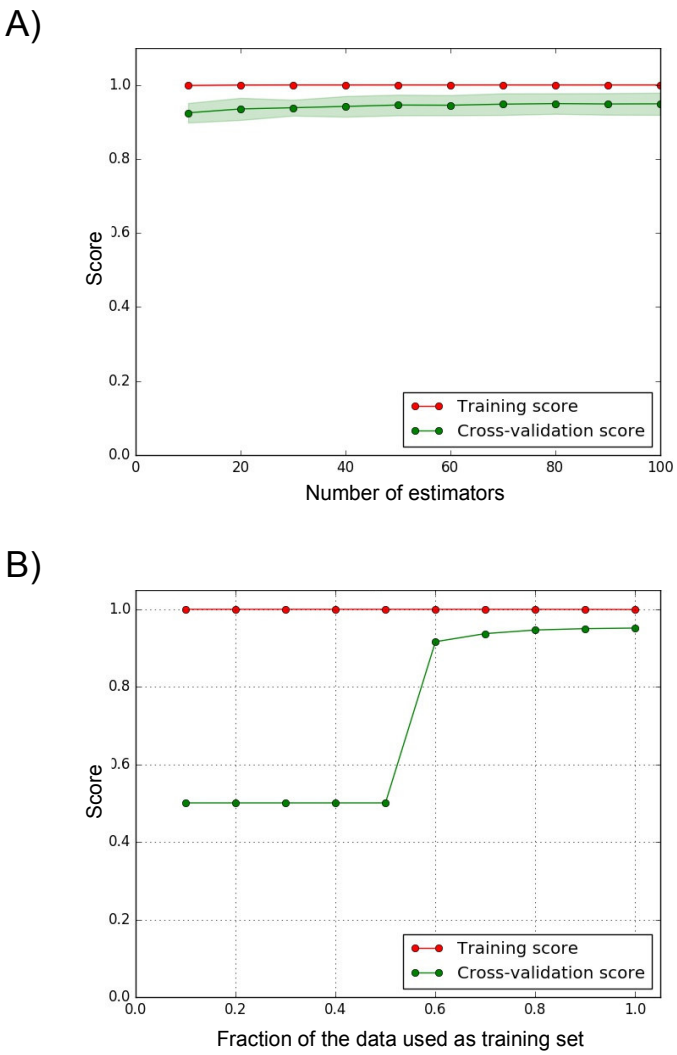


Figure S10. Representations used to define the RF model parameters and evaluate its performance. The training set is plotted in red and the validation set is plotted in green. In panel A, the validation curve is plotted varying the number of trees, showing that the estimator is not overfitted as the training and test scores are similar and that the number of trees chosen is adequate. In panel B, the learning curve is plotted using training and test sets of different sizes. It shows that the training set needs to comprise a high percentage of the data to reduce the error of variance.