Optimizing electrostatic similarity for virtual screening: A new methodology

S. Puertas-Martín,*† J. L. Redondo,† H. Pérez-Sánchez,‡ and P. M. Ortigosa†

†Department of Computer Science, University of Almería, Almería, ceiA3, Spain
‡Bioinformatics and High Performance Computing Research Group (BIO-HPC), Computer Engineering Department, Universidad Católica San Antonio de Murcia (UCAM), Spain

E-mail: savinspm@ual.es

Abstract

Ligand Based Virtual Screening (LBVS) methods are widely used in drug discovery as filters for subsequent in-vitro and in-vivo characterization. This means, increasing accuracy of LBVS approaches may have a huge impact on increasing chances of success. Since the databases processed in drug discovery campaigns are enormously large, this pre-selection process requires the use of fast and precise methodologies. The similarity between compounds can be measured using different descriptors such as shape, pharmacophore or electrostatic similarity. The latter is the goal of this work, i.e., we want to improve the process of obtaining the compounds most similar to a query in terms of electrostatic similarity. To do so, the current and widely proposed methodology in the literature is based on the use of ROCS to assess the similarity of compounds in terms of shape and then evaluate a small subset of them with ZAP for prioritization regarding electrostatic similarity. This paper proposes an alternative methodology that consists of directly optimizing electrostatic similarity and works with the entire database of compounds without using shape cut-offs. For this purpose, a new and improved version of the OptiPharm software has been developed. OptiPharm implements a parameterizable metaheuristic algorithm able to solve any optimization problems directly related to the involved molecular conformations. We show that our new method completely outperforms the classical proposal widely used in the literature. Accordingly, we are able to conclude that many of the compounds proposed with our novel approach could not be discovered with the classical one. As a result, this methodology opens up new horizons in Drug Discovery.
Introduction

The constant increase in the size of the databases used in Drug Discovery requires efficient techniques and methods that can be used to select the compounds most similar to a query molecule and at the lowest possible cost. One of these techniques is Virtual Screening (VS). VS is an in-silico technique that allows large libraries with millions of compounds to be processed in order to find new compounds related to a pharmacological query based on one or more features.\(^1\)-\(^4\) This represents a great advantage over experimental methods such as High-throughput screening (HTS) in terms of efficiency, budget, time and development cost.\(^5\) The resulting compounds from VS are subsequently acquired and empirically tested in the laboratory. In addition, VS techniques are often used as a pre-filter for HTS.\(^6\) All these advantages have increased the popularity of these techniques, which have experienced great advances over the last two decades. The interested reader is referred to previous works\(^7\)-\(^11\) for a description of different methods and tools currently used on VS.

However, there is still room for improvement regarding the accuracy of VS predictions so as not to discard promising compounds, or to reduce the time and error of calculations that compute the different features of the studied compounds.\(^12\) VS applied to the electrostatic similarity of compounds is a clear example of this. Contrary to what happens when VS is applied to select the most similar compounds in shape or pharmacophore properties, where the tools base their predictions on scoring functions that measure these particular features,\(^7,13,14\) the predictions in this field are not exclusively based on this descriptor, but on both the similarity of the three dimensional shape and electrostatic similarity.\(^15\)-\(^27\)

Broadly speaking, all the previous works follow the same methodology, although they may differ in the selection procedure used to determine the compounds proposed as best predictions. Essentially, they initially optimize the compounds in the database against the query in terms of shape by using ROCS,\(^28\) they prioritize the top \(N\) compounds with the highest shape similarity values and then evaluate them in terms of electrostatic similarity. It is noteworthy that in many different studies, the compounds with the highest values of shape are proposed as the most promising ones, while in just a few, only compounds where the combination of shape and potential is greater than a certain threshold are selected for further experimental validation. In this work, we will focus on comparison with the former approach, since it is the most widely used. The value of \(N\) is not fixed, as it depends on the particular study. Usually, \(N\) is less than 10% of the total compounds in the database.\(^18,23,24\) A search for the best compounds basing on shape pre-filtering may be counterproductive, since the selection of a low value of \(N\) can rule many promising compounds out, which may have a significant impact on the final results.

In this work, we propose a new approach that involves the direct optimization of the electrostatic similarity, which can provide a more realistic description of compound bioactivity. This has been possible thanks to a novel software, called OptiPharm, able to optimize the similarity between two given molecules for any scoring function provided. Not only that, but OptiPharm also possesses other advantages, since it can be parameterized according to different aims such as carrying out an exhaustive search for a compound or, alternatively, obtaining a re-
result of good quality but in a short time. It means OptiPharm can be adapted to different sorts of problems which is an advantage regarding the virtual screening problem, where compounds have different sizes and complexities. The interested reader is referred to the original work\textsuperscript{13} for an in-depth description of this method. Additionally, they have software available with the web tool BRUSELAS\textsuperscript{29} (http://bio-hpc.eu/software/Bruselas). In this work, OptiPharm has maximized electrostatic similarity by including an objective function measuring it as a unique modification. In the near future new objectives such as the pharmacophoric scoring function or the quantum mechanical based field scoring function will be also included. As results show, new algorithms such as OptiPharm are required since the solutions found are better and hence total time and budget are reduced in the discovery process. Furthermore, most of the research has always followed the same methodology, so the new results obtained are practically unexplored.

The rest of the paper is organized as follows. Section \textit{Methods} describes the two methods used for virtual screening based on electrostatic similarity, as the literature approach as the novel proposal. Additionally, a brief description about the mathematical formulation of the scoring functions is included. Section \textit{Results} describes the framework where the experiments have been carried out and the main results obtained. Finally, the conclusions and lines for future research are summarized in the last section.

\section*{Methods}

In this section, we describe the two optimization methodologies for Ligand Based Virtual Screening (LBVS) based on electrostatic similarity. The former is currently the method most frequently used in the literature. In short, it computes a sublist of molecules with the highest three-dimensional shape similarity. Usually, such a sublist is only composed of less than 10\% of the total number of compounds in the database. From the reduced list, the compound(s) with the greatest electrostatic similarity is/are selected. The second one involves the solving of an optimization problem guided by a function measuring the electrostatic similarity. Subsections \textit{Literature approach: The LBVS method guided by molecular shape (LBVS-S)} and \textit{The new approach: A LBVS method guided by electrostatic similarity (LBVS-E)} briefly describe both techniques. For the sake of completeness, it is included here a first subsection devoted to defining the mathematical functions used to guide the searching processes. The figures in which the values of these objective functions are graphically represented have been created with VIDA v4.4.0\textsuperscript{30} using the default configuration.

\textbf{Scoring functions to measure similarity between compounds.}

\textbf{Shape Similarity}

The shape similarity of two compounds is calculated as follows:
\[ V_{AB} = \sum_{i \in A, j \in B} V_{ij} = \sum_{i \in A, j \in B} p_i p_j K_{ij} \left( \frac{\pi}{\alpha_i + \alpha_j} \right)^3 \]  

where \( p_i \) and \( p_j \) are set to 2.7, \( \alpha_i \) and \( \alpha_j \) obtain the van der Waals value for each atom and

\[ K_{ij} = \exp \left( -\frac{\alpha_i \alpha_j R_{ij}^2}{\alpha_i + \alpha_j} \right) \]

where \( R_{ij} \) is the distance between atoms \( i \) and \( j \).

Notice that the accuracy obtained from Equation 1 depends on the number of atoms in the two compared molecules, i.e., the higher this number, the longer the value of \( V_{AB} \) as an absolute value. To be able to measure the level of similarity between compounds, regardless of the number of atoms that they are composed of and the descriptor used, the Tanimoto Similarity value is computed as follows:

\[ T_c = \frac{V_{AB}}{V_{AA} + V_{BB} - V_{AB}} \]

where \( V_{AB} \) is the A molecule overlaid onto B molecule. \( V_{AA} \) and \( V_{BB} \) is the overlap of the molecules A and B, respectively. Equation 3 has a value in the range \([0, 1]\), where 0 means there is no overlapping and 1 means the shape of both molecules is the same.

**Electrostatic similarity**

The electrostatic similarities are obtained by numerical solution of the Poisson equation, viz:

\[ \nabla \{ \epsilon(r) \nabla \phi(r) \} = -\rho_{mol}(r) \]

where \( \nabla \) is the electrostatic potential, \( \epsilon(r) \) is the dielectric constant, and \( \rho_{mol}(r) \) is the molecular charge distribution. Electrostatic similarity between two compounds is compared by determining \( E_{AB} \):

\[ E_{AB} = \int \rho^A(r) \rho^B(r) \Theta^A(r) \Theta^B(r) d\vec{r} \approx h^3 \sum_{ijk} \phi_{ijk}^A \phi_{ijk}^B \Theta_{ijk}^A \Theta_{ijk}^B \]

where \( \Theta \) is a masking function to ensure potentials inside the compound are not considered part of the comparison. The integral appearing in Equation 5 is a volume integral, computed using a grid-spacing parameter, \( h \).

Again the accuracy obtained by Equation 5 depends on the number of atoms in the compared molecules. As such, similarly to what was done previously, the Tanimoto Similarity value has been computed as follows:

\[ T_{cE} = \frac{E_{AB}}{E_{AA} + E_{BB} - E_{AB}} \]

where \( E_{AB} \) is the A molecule overlaid onto B molecule. \( E_{AA} \) and \( E_{BB} \) is the overlap of the molecules A and B, respectively. In this case, Equation 6 has a value in the range \([-0.33, 1]\), where -0.33 means the charges of both compounds have the same
value but opposite loads, 0 means there is no overlapping, and 1 means the charges of both molecules are the same.

**Literature approach: The LBVS method guided by molecular shape (LBVS-S)**

This method bases its predictions on a previous pre-filtering process consisting of identifying the \( N \) candidate compounds from the database with the highest shape similarity. After that, for each selected compound, the electrostatic similarity is calculated at the optimum superimposition obtained in the previous stage. Finally, the molecule with the highest electrostatic similarity value is selected as the one for the solution.

In this work, we have used the tool ROCS\textsuperscript{28} to optimize the shape similarity between two molecules. ROCS is a parametrized piece of software used to maximize volume overlapping similarity and utilizes the previously described Equation 3 to represent molecules by means of Gaussian functions.\textsuperscript{33,34} Electrostatic similarity has been calculated using the ZAP Toolkit (see Equation 6). This software has been downloaded without modification from.\textsuperscript{35} It is worth mentioning that ROCS and ZAP are, by far, the most widely used tools in the literature for VS based on shape and electrostatic similarity.\textsuperscript{36–40} For this reason they have been selected as part of this study; i.e. a fair and complete study must be carried out by making a comparison with the state-of-the-art methods.

**The new approach: A LBVS method guided by electrostatic similarity (LBVS-E)**

Our main aim when using this approach is to obtain the compound(s) with the highest electrostatic similarity values. Thus, an optimization problem must be defined with this aim in mind. Broadly speaking, any tool, method or algorithm used will be better guided towards the optima if the objective function is a numerical model representing the real objective. Until now, the methods proposed were based on the idea that if two compounds have the same shape, they can probably have the same electrostatic properties. Consequently, they solve a shape similarity optimization problem, instead of focusing on the electrostatic similarity, which may be more useful from the drug discovery point of view. The new approach presented here is based on the idea that the scoring function used to guide the optimization method must be the one based on electrostatic similarity, and not any other. If not, the search may converge to a sub-optimal solution.\textsuperscript{41–43} To prove our hypothesis, OptiPharm,\textsuperscript{13} a recently proposed tool which implements an evolutionary optimization algorithm specifically designed to work on LBVS problems, is used. It is not limited to a particular objective function; on the contrary, OptiPharm admits any scoring function. In this work, to be sure that we are comparing in the same terms and equal conditions, the electrostatic similarity between two compounds has been computed by using the source code of the ZAP Toolkit, also downloaded from https://docs.eyesopen.com/toolkits/python/zaptk/thewayofzap.html.\textsuperscript{35} We
Figure 1: Number of compounds included on the FDA database, according to their number of atoms.

have only included a slight modification related to the function that reads compounds. Specifically, the method open belonging to oemolstream ifs has been replaced by openstring. The idea behind this change is to reduce the computing time, i.e. open involves writing to and reading from a local file while openstring allows reading from a c++ string directly.

**Hardware setup**

All the experiments in this work have been executed using a Bullx R424-E3, which consists of 2 Intel Xeon E5 2650v2 (16 cores), 128 GB of RAM memory and 1 TB HDD (http://hpca.ual.es/en/infraestructure) along with the cluster Eagle (https://wiki.man.poznan.pl/hpc/index.php?title=Eagle).

**Benchmarks**

In this work, a database provided by The Food and Drug Administration has been used (FDA). The Food and Drug Administration is a federal agency of the United States Department of Health and Human Services responsible for protecting and promoting public health by controlling, among other things, prescription and over-the-counter pharmaceutical drugs (medications). This agency provides a data set containing 1751 compounds, which represents approved medicines that can be safely used on humans in the USA. This database is useful since in the high similarity cases it would directly contribute to drug re-purposing. This is of relevant utility given the clear trend regarding re-purposing drugs observed over the last 5 years.44–46

The version of the database used in this work was obtained from DrugBank v5.0.147 and necessary mol2 files for the VS calculations were set up by using AmberTools by removing salts and neutralizing their protonation state, computing partial charges by MMFF94 force field, adding hydrogen atoms and minimizing energies (default parameters).49
A comprehensive computational analysis may cover a representative sample of the database. The compounds included in the FDA database have different attributes, one of the most relevant for the study at hand being the number of atoms. In this work, a selection of 50 compounds has been made in the following way: the compounds in the database have been sorted by the number of atoms, including hydrogen, and then divided into 24 intervals (see Figure 1). From each sector, at least one compound was chosen at random and proportional to the number of compounds in the sector.

Results

Influence of the size list of top-ranked compounds in the LBVS-S method.

As previously mentioned, the LBVS method guided by molecular shape, bases its predictions on pre-filtering of the $N$ best compounds in terms of superimposition score. In this subsection, a study has been conducted to know how the value of $N$ affects the final results from the point of view of electrostatic similarity. In particular, the LBVS-S has been performed on the selected 50 queries and for five different values of $N$, i.e. $N$ has been set to 175, 438, 876, 1313 and 1751.

Figure 2 illustrates the main steps of the LBVS-S method for the Query DB01213 and $N = 1751$, i.e. the total number of compounds in the FDA set. Initially, the Query is compared to each compound TargetS from the database to obtain their optimum position and corresponding shape similarity value $Tc_S$. As previously mentioned, this stage is carried out by using ROCS. Afterwards, compounds are sorted ($Rk_S$) in decreasing order by $Tc_S$. The $N$ best compounds are selected and evaluated to measure the corresponding electrostatic similarity value $Tc_E^{Eval}$. Notice that the evaluation of the electrostatic similarity considers the pose obtained with the shape similarity optimization. The compound with the highest $Tc_E^{Eval}$, called BestComp throughout this paper, is selected as the best prediction. Finally, as an additional and unconsidered stage in the LBVS-S method, we have computed the optimized superposition between the BestComp and the Query by using OptiPharm. The corresponding $Tc_E$ value is then provided.

To get an overview of the results, average values of the BestComp found for the 50 queries and each value for $N$ have been computed, and shown in Table 1. In particular, the average position $Av(Rk_S)$ in the sorted list where the BestComp were located have been computed, together with the following: their mean number of atoms $Av(N_S)$, their average shape similarity value $Av(Tc_S)$, their corresponding electrostatic similarity value $Av(Tc_E^{Eval})$ when they are evaluated and finally their mean electrostatic similarity when they are optimized $Av(Tc_E)$.

As can be seen, the predictions seem to improve in term of electrostatic similarity as the number $N$ of selected molecules in the sorted list increases (see columns $Av(Tc_E^{Eval})$ and $Av(Tc_E)$). In accordance with these results, the posterior comparison between LBVS-S and LBVS-E methods has been carried out by setting $N = 1751$. 
Optimizing shape similarity of the candidate compounds. Sorting candidate compounds by $T_c$. Selecting the N best candidates and evaluating Electrostatic similarity. Optimizing electrostatic similarity of the best compound. 

Selecting the compound with the highest electrostatic similarity. 

Optimization of Shape similarity by ROCS. Evaluation of Electrostatic similarity by ZAP. Optimization of Electrostatic similarity by OptiPharm. 

Figure 2: Performance of the LBVS-S method for a particular case where Query = DB01213 and $N = 1751$ using the FDA database. 

Table 1: Influence of the parameter $N$ in the results obtained by the LBVS-S method. For each value of $N$, the following average values from the 50 queries, are shown: position in the shape ranking ($Av(R_{kS})$), number of atoms ($Av(N_S)$), shape similarity score ($Av(T_{cS})$), electrostatic similarity evaluation score ($Av(T_{cE}^{Eval})$) and electrostatic optimized similarity value ($T_{cE}$). 

| $N$  | $Av(R_{kS})$ | $Av(N_S)$ | $Av(T_{cS})$ | $Av(T_{cE}^{Eval})$ | $Av(T_{cE})$ |
|------|--------------|-----------|--------------|----------------------|--------------|
| 175  | 73           | 53        | 0.627        | 0.451                | 0.559        |
| 438  | 162          | 50        | 0.587        | 0.486                | 0.568        |
| 876  | 287          | 51        | 0.564        | 0.495                | 0.569        |
| 1313 | 324          | 50        | 0.559        | 0.497                | 0.570        |
| 1751 | 362          | 49        | 0.554        | 0.497                | 0.569        |

Performance comparison between LBVS-S and LBVS-E methods. 

To analyze the performance of both methods, we have conducted a study in which the selected 50 molecular queries are processed with reference to the FDA database.
Figure 3: Performance of the LBVS-E method for a particular case where Query = DB01213 is compared to the FDA database.

Notice that comparing a query with itself always reaches the maximum similarity value, both for electrostatic potential as well as for shape. Subsequently, these results were removed when ranking the compounds. In other words, the compounds given as a result are not the most similar ones, but the second compounds in the ranked list. Additionally, as previously mentioned, the traditional method has been carried out considering the total number of compounds in the database $N = 1751$, so as to increase the probability of finding better predictions.

To illustrate how we generate the later summarizing tables, a sample of the results obtained by both methods when comparing a query to the molecules in the dataset is studied. In particular, the instance Query $= DB01213$ is analyzed. Notice that this is the example used to illustrate the stages of the LBVS-S method in Figure 2. After that, the same instance is considered to exemplify the performance of the LBVS-E method (see Figure 3). Notice that this Query has been selected because it is small and it helps to see the main ideas of the paper very easily by using figures. However, the conclusions inferred from the associated results can be extrapolated to any other Query. As can be observed, the LBVS-E technique solves an optimization problem to determine the electrostatic similarity, $Tc_E$, between the pharmaceutical Query and every Target_E in the database. Afterwards, the list of compounds is sorted by the $Tc_E$ value and the one located in first position, $Rk_E = 1$, is selected as the best prediction. Finally, to complete the study, optimization is carried out to calculate the shape similarity $Tc_S$ between the chosen compound and the Query.
Query compound: DB01213

| LBVS-S. Best compound: DB00184 |
|--------------------------------|
| ![DB00184](image1) |
| $T_c^S = 0.621$ |
| $T_c^{Eval} = 0.500$ |
| ![DB01213](image2) |
| ![DB00184](image3) |
| ![DB03255](image4) |
| $T_c^E = 0.609$ |

| LBVS-EP. Best compound: DB03255 |
|--------------------------------|
| ![DB03255](image5) |
| $T_c^S = 0.963$ |
| $T_c^{Eval} = 0.880$ |
| ![DB00184](image6) |
| ![DB01213](image7) |
| ![DB03255](image8) |
| $T_c^E = 0.810$ |

Figure 4: Results of LBVS-S and LBVS-E where Query = DB01213. The Query compound is colored green. Query electrostatic fields are colored deep blue and red. Best compounds are shown in grey and their electrostatic potential fields in light blue and pink.

For the sake of clarity and comparison, the results shown in Figures 2 and 3 are summarized in Table 2. The meaning of the columns as well as the particular values in the tables, are the ones previously explained and shown in each figure. The last row corresponds to the values associated with the best predictions. As can be observed, each method obtains a different compound as a top solution. LBVS-S provides the DB00184 molecule with a $T_c^S = 0.621$ and a $T_c^{Eval} = 0.500$. At the same time, LBVS-E proposes the DB03255 compound as being the most similar to the query with $T_c^E = 0.810$ and $T_c^{Eval} = 0.880$. As such, the LBVS-E method has not only obtained a more similar compound in terms of electrostatic potential, but also in shape. In Figure 4 the final position for each case is shown.

Once the specific case of DB01213 has been explained in detail, the results of the 50 queries have been summarized in Table 3. Columns $Rk_{E}^{Eval}$ and $Rk_{E}$ have been removed in this table because their values are always 1. The last row summarizes the average of the results.

As evidenced, LBVS-E obtains on average $T_c^E = 0.738$, which is higher than that given by LBVS-S, $T_c^{Eval} = 0.497$. Similar conclusions can be inferred when comparing the $T_c^E$ average values for both methods. Additionally, when the results are analyzed individually, we can see that LBVS-E provides solutions with higher $T_c^E$ values than those achieved by LBVS-S. In fact, in 48 out of 50 cases, LBVS-E obtains a different compound than that reached by LBVS-S.

Regarding shape similarity, it is possible to infer that, on average, the methods are equivalent in terms of accuracy of the predictions, i.e. LBVS-S obtains an average value of $T_c^s = 0.554$ while LBVS-E reaches a mean value of $T_c^s = 0.505$. Furthermore, analyzing the obtained results individually, we can see that in 2 out of 50 cases, LBVS-E offers better or equivalent predictions than that achieved by...
Table 2: Summary of the results obtained for both LBVS-S and LBVS-E methods for the query compound DB01213. The column notation, the colors included and the corresponding results come from Figures 2 and 3, i.e. they maintain the same meaning as shown previously for those pictures. The last row indicates the results associated with the top solution selected for each method.

| Query | $N_Q$ | $Rk_S$ | Target$_S$ | $S$ | $Tc_s$ | $Rk_{EvalS}$ | Tc$_{EvalS}$ | $Tc_E$ | $N_E$ | Tc$_{EvalE}$ | $Tc_E$ | $E$ |
|-------|-------|--------|------------|-----|-------|--------------|-----------|-------|------|-------------|-------|-----|
| DB01213 | 12    | 1      | DB03255    | 13  | 0.963 | 111          | 0.140     |       |      |             |       |     |
|       |       | 2      | DB06637    | 13  | 0.961 | 242          | 0.081     | 2     | DB06637    | 13  | 0.798 | 0.885 |
|       |       | 3      | DB01065    | 9   | 0.943 | 1722         | -0.172    | 3     | DB00763    | 13  | 0.796 | 0.845 |
| 183   | DB00184 | 26    | 0.621      | 1   | 0.500 |             |           | 110   | DB00184    | 26  | 0.609 | 0.319 |
|       |       |       |             |     |       |             |           |       |      |             |       |     |

LBVS-S in terms of shape (see columns $Tc_s$ in LBVS-S and $Tc_{EvalS}$ in LBVS-E). It means that cases exist where two compounds can be very similar in terms of electrostatic potential, although they can be very different in terms of three-dimensional shape. It means that those solutions could not be obtained by using the methodology followed by the traditional LBVS-S method, since it only focuses on the compounds with the highest similarity in shape.

Making a somewhat more detailed approach for compounds smaller than 50 atoms, which means the first 23 query compounds in the table, there are 5 cases where the difference is less than 0.05 (DB00529, DB00173, DB00331, DB00915 and DB01352) and in another 3 cases the difference is 0.1 (DB01043, DB07615 and DB01268). Considering the values of these 7 cases in which the shape LBVS-E is smaller than that of LBVS-S, the average difference is 0.048, while the mean gain in electrostatic similarity for those 7 compounds is 0.271. In large compounds, which includes 27 queries, there are only two cases with similar characteristics, which are compounds DB09236 with a difference of 0.07 and DB06699 with a difference of 0.013, both of them for shape similarity. In view of these results, the LBVS-E method seems to be justified when proposing new solutions for small compounds.

However, not all the improvements are related to electrostatic fields. The optimization of electrostatic potential using OptiPharm might allow a better solution to be found in terms of shape too. Compounds DB01119 and DB1213 in Table 3 are some outstanding examples. For example, in the case of Query = DB01119, the best compound found by LBVS-S is DB00828 with $Tc_S = 0.655$ and $Tc_{EvalS} = 0.519$. Moreover, LBVS-E’s best compound is DB00173. It has a better $Tc_E$, i.e. 0.829, but also the position of those compounds after the electrostatic optimization is improved, $Tc_{EvalE} = 0.779$. 

11
Table 3: Rows are sorted by the number of atoms of queries. For each query, the same procedure explained in Table 2 is followed. The last row summarizes the average values for each column.

| Query  | $N_Q$ | $Rk_S$ | $Target_S$ | $S$ | $T_c$ | $Eval$ | $T_c$ | $Eval$ | $T_c$ | $Eval$ | $T_c$ | $Eval$ |
|--------|-------|--------|------------|-----|-------|-------|-------|-------|-------|-------|-------|-------|
| DB00529 | 10    | 316    | DB05266    | 35  | 0.496 | 0.437 | 0.593 | 0.468 | 0.614 |
| DB01213 | 12    | 182    | DB00184    | 26  | 0.621 | 0.500 | 0.609 | 0.880 | 0.963 |
| DB00173 | 15    | 102    | DB00851    | 23  | 0.792 | 0.546 | 0.536 | 0.777 | 0.830 |
| DB00172 | 17    | 24     | DB00128    | 18  | 0.881 | 0.469 | 0.696 | 0.814 | 0.769 |
| DB00331 | 20    | 380    | DB00961    | 40  | 0.598 | 0.599 | 0.697 | 0.559 | 0.649 |
| DB01119 | 21    | 513    | DB00828    | 15  | 0.655 | 0.519 | 0.613 | 0.779 | 0.829 |
| DB02513 | 25    | 27     | DB1275     | 20  | 0.872 | 0.526 | 0.569 | 0.745 | 0.805 |
| DB00915 | 28    | 125    | DB00160    | 13  | 0.684 | 0.404 | 0.543 | 0.673 | 0.924 |
| DB01352 | 29    | 1      | DB00306    | 32  | 0.926 | 0.947 | 0.983 | 0.901 | 0.926 |
| DB01365 | 30    | 180    | DB1191     | 33  | 0.738 | 0.902 | 0.960 | 0.628 | 0.824 |
| DB00657 | 33    | 47     | DB06770    | 16  | 0.788 | 0.396 | 0.517 | 0.546 | 0.614 |
| DB00478 | 34    | 30     | DB00752    | 21  | 0.787 | 0.508 | 0.637 | 0.609 | 0.861 |
| DB01043 | 34    | 27     | DB09045    | 21  | 0.765 | 0.400 | 0.478 | 0.711 | 0.861 |
| DB00380 | 35    | 601    | DB00731    | 50  | 0.619 | 0.380 | 0.407 | 0.453 | 0.655 |
| DB00693 | 37    | 1034   | DB04575    | 59  | 0.525 | 0.362 | 0.429 | 0.391 | 0.783 |
| DB09185 | 37    | 243    | DB1233     | 43  | 0.739 | 0.839 | 0.960 | 0.429 | 0.650 |
| DB07615 | 40    | 71     | DB04552    | 28  | 0.704 | 0.861 | 0.866 | 0.610 | 0.574 |
| DB09219 | 40    | 123    | DB00321    | 44  | 0.698 | 0.347 | 0.329 | 0.450 | 0.462 |
| DB00674 | 42    | 279    | DB00375    | 23  | 0.688 | 0.505 | 0.653 | 0.662 | 0.495 |
| DB00887 | 45    | 209    | DB00232    | 31  | 0.642 | 0.401 | 0.454 | 0.378 | 0.576 |
| DB01198 | 45    | 273    | DB00209    | 59  | 0.648 | 0.748 | 0.768 | 0.334 | 0.491 |
| DB01155 | 48    | 1      | DB01165    | 46  | 0.858 | 0.671 | 0.818 | 0.899 | 0.835 |
| DB00246 | 50    | 467    | DB00268    | 44  | 0.542 | 0.843 | 0.852 | 0.877 | 0.604 |
| DB00381 | 53    | 525    | DB00573    | 32  | 0.577 | 0.285 | 0.278 | 0.377 | 0.524 |
| DB00876 | 54    | 576    | DB1002    | 49  | 0.516 | 0.395 | 0.505 | 0.532 | 0.524 |
| DB09237 | 54    | 380    | DB09092    | 44  | 0.580 | 0.759 | 0.824 | 0.902 | 0.447 |
| DB00254 | 55    | 1100   | DB00753    | 25  | 0.521 | 0.626 | 0.836 | 0.868 | 0.574 |
| DB01268 | 57    | 902    | DB00914    | 54  | 0.518 | 0.792 | 0.765 | 0.883 | 0.564 |
| DB01196 | 60    | 7      | DB00783    | 44  | 0.741 | 0.397 | 0.385 | 0.527 | 0.195 |
| DB01621 | 66    | 274    | DB00268    | 44  | 0.552 | 0.821 | 0.845 | 0.867 | 0.454 |
| DB09236 | 66    | 459    | DB00607    | 51  | 0.509 | 0.406 | 0.438 | 0.664 | 0.551 |
| DB00632 | 69    | 537    | DB00511    | 123 | 0.348 | 0.067 | 0.246 | 0.997 | 0.137 |
| DB08903 | 69    | 6      | DB01433    | 58  | 0.621 | 0.840 | 0.867 | 0.588 | 0.464 |
| DB01419 | 70    | 380    | DB09209    | 61  | 0.431 | 0.854 | 0.879 | 0.933 | 0.423 |
| DB00320 | 80    | 204    | DB00438    | 23  | 0.515 | 0.367 | 0.396 | 0.563 | 0.245 |
| DB00728 | 91    | 1383   | DB06204    | 40  | 0.399 | 0.688 | 0.761 | 0.874 | 0.101 |
| DB00503 | 98    | 655    | DB00206    | 84  | 0.371 | 0.256 | 0.243 | 0.401 | 0.280 |
| DB01232 | 100   | 639    | DB06480    | 52  | 0.389 | 0.691 | 0.741 | 0.791 | 0.387 |
| DB00309 | 110   | 385    | DB1603     | 45  | 0.455 | 0.241 | 0.297 | 0.467 | 0.534 |
| DB04786 | 120   | 4      | DB09158    | 82  | 0.377 | 0.424 | 0.708 | 0.910 | 0.120 |
| DB09114 | 130   | 117    | DB00595    | 57  | 0.376 | 0.273 | 0.506 | 0.876 | 0.190 |

Continued on next page
ZAP Toolkit accuracy problem

The ZAP Toolkit has been widely used in the literature to calculate the electrostatic similarity score for two compounds.\(^2,15–27,50\)

In this subsection we would like to remark that the ZAP Toolkit can return an erroneous value, which was discovered when using Optipharm. During the optimization procedure, Optipharm can progressively separate two input compounds aimed to escape from local optima and explore the searching space in depth. In fact, it is possible to analyze cases where no overlap exists between the input molecules. During the analysis of the results, we discovered that cases exist where the ZAP Toolkit can overflow, mainly when situations such as the previously mentioned happen. See Figure 5 to see a particular example. Herein, compound DB01365 remains fixed on the left while compound DB00459 occupies three positions (red, blue and pink). The red compound obtains an electrostatic similarity value of 1. The light blue compound is displaced half a unit to the left, i.e., closer to the reference compound and its similarity value is 0.38. The pink compound is shifted 0.5 units to the right, that is, away from the reference compound. Its similarity value is 0. Calculations can be made using the Zap python script available at https://docs.eyesopen.com/toolkits/python/zaptk/thewayofzap.html in the Electrostatic Similarity section.

This problem has been solved in OptiPharm by considering the poses with the previously mentioned problem unfeasible. It means that they are no longer considered during the optimization process.

Conclusions

In this work, a new approach to solve the LBVS problem based on the electrostatic similarity has been put forward. More specifically, the software OptiPharm, recently proposed in the literature, has been used as alternative to the current state-of-the-art method. OptiPharm implements an evolutionary optimization algorithm able to include as objective, any scoring function. As such, it can directly solve the electro-
static similarity optimization problem by only including as an objective a function measuring such a descriptor. The method resulting from joining OptiPharm with the electrostatic similarity objective has been called LVBS-E. Conversely, the method proposed in the literature, which has been named LBVS-S throughout the paper, looks for a sublist of the top compounds with the highest shape similarity by using ROCS, to later evaluate their electrostatic similarity with ZAP. In this work, a study to analyze the influence of the number of compounds in such a sublist has been carried out. As the results have shown, the larger the number of molecules considered, the better the prediction obtained in terms of electrostatic similarity.

From this conclusion, a computational study has been carried out to compare the new method LBVS-E with the one in the literature LBVS-S. To increase the probability of finding good predictions, LBVS-S has been executed taking into account the whole database prior to the electrostatic similarity evaluation. Even so, LBVS-E performs better than LBVS-S, achieving better predictions in electrostatic potential for the 50 queries included in the study. Regarding the shape similarity, both methods behave in a similar fashion, on average obtaining compounds with similar shape similarity values. It is important to mention that the new methodology proposed in this paper is novel, which means that the predictions proposed have not been analyzed previously.

Finally, we have shown that ZAP can return erroneous values. This is an important discovery, since it is the most commonly used software in the literature to measure the electrostatic similarity.

In the future, we have plans to implement this objective function from scratch, but for the study at hand, we considered that it was more important to compare it with the state-of-the-art software. Please note that, this implementation will also decrease the execution time in OptiPharm, since from a computational point of view,
integrating a function into a program is a very time consuming task. Additionally, other functions measuring the pharmacophore similarity will be implemented. Additionally, we will analyze the problem from a multi-objective perspective, where shape an electrostatic similarity are optimized simultaneously.

Acknowledgement

This work was funded by grants from the Spanish Ministry of Economy and Competitiveness (RTI2018-095993-B-100, CTQ2017-87974-R), Junta de Andalucía (P12-TIC301, UAL18-TIC-A020-B), Fundación Séneca–Agencia de Ciencia y Tecnología de la Región de Murcia under Projects 20988/PI/18 and 20817/PI/18. Powered@NLHPC: This research was partially supported by the supercomputing infrastructure of the NLHPC (ECM-02). The authors also thankfully acknowledge the computer resources and the technical support provided by the Plataforma Andaluza de Bioinformática of the University of Málaga. This work was partially supported by the computing facilities of the Extremadura Research Centre for Advanced Technologies (CETA—CIEMAT), funded by the European Regional Development Fund (ERDF). CETA—CIEMAT belongs to CIEMAT and the Government of Spain. Savín Puertas Martín is a fellow of the Spanish ‘Formación de Profesorado Universitario’ program, financed by the Spanish Ministry of Education, Culture and Sport. Additionally, the authors would also like to thank N.C. Cruz for his technical support.

References

(1) Hamza, A.; Wei, N.-N.; Zhan, C.-G. Ligand-Based Virtual Screening Approach Using a New Scoring Function. Journal of Chemical Information and Modeling 2012, 52, 963–974.

(2) Boström, J.; Grant, J. A.; Fjellström, O.; Thelin, A.; Gustafsson, D. Potent Fibrinolysis Inhibitor Discovered by Shape and Electrostatic Complementarity to the Drug Tranexamic Acid. 2013.

(3) Kumar, A.; Zhang, K. Y. Application of Shape Similarity in Pose Selection and Virtual Screening in CSARdock2014 Exercise. Journal of Chemical Information and Modeling 2016, 56, 965–973.

(4) Wang, Z.; Lu, Y.; Seibel, W.; Miller, D. D.; Li, W. Identifying Novel Molecular Structures for Advanced Melanoma by Ligand-Based Virtual Screening. Journal of Chemical Information and Modeling 2009, 49, 1420–1427.

(5) Kar, S.; Roy, K. How far can virtual screening take us in drug discovery? Expert Opinion on Drug Discovery 2013, 8, 245–261.

(6) López-Ramos, M.; Perruccio, F. HPPD : Ligand- and Target-Based Virtual Screening on a Herbicide Target. J. Chem. Inf. Model. 2009, 50, 801–814.
(7) Lešnik, S.; Štular, T.; Brus, B.; Knez, D.; Gobec, S.; Janežič, D.; Konec, J. LiSiCA: A Software for Ligand-Based Virtual Screening and Its Application for the Discovery of Butryrylcholinesterase Inhibitors. *Journal of Chemical Information and Modeling*, 2015, 55, 1521–1528.

(8) Kalászi, A.; Szisz, D.; Imre, G.; Polgár, T. Screen3D: A Novel Fully Flexible High-Throughput Shape-Similarity Search Method. *Journal of Chemical Information and Modeling*, 2014, 1036–1049.

(9) Liu, X.; Jiang, H.; Li, H. SHAFTS: A hybrid approach for 3D molecular similarity calculation. 1. method and assessment of virtual screening. *Journal of Chemical Information and Modeling*, 2011, 51, 2372–2385.

(10) Dou, X.; Jiang, L.; Wang, Y.; Jin, H.; Liu, Z.; Zhang, L. Discovery of new GSK-3 \( \beta \) inhibitors through structure-based virtual screening. *Bioorganic & Medicinal Chemistry Letters*, 2018, 28, 160–166.

(11) Schmidt, T. C.; Cosgrove, D. A.; Boström, J. ReFlex3D: Refined Flexible Alignment of Molecules Using Shape and Electrostatics. *Journal of Chemical Information and Modeling*, 2018, acs.jcim.7b00618.

(12) Böhm, H.-J.; Stahl, M. *The Use of Scoring Functions in Drug Discovery Applications*; John Wiley & Sons, Ltd, 2003; pp 41–87.

(13) Puertas-Martín, S.; Redondo, J. L.; Ortigosa, P. M.; Pérez-Sánchez, H. OptiPharm: An evolutionary algorithm to compare shape similarity. *Scientific Reports*, 2019, 9, 1398.

(14) Yan, X.; Li, J.; Liu, Z.; Zheng, M.; Ge, H.; Xu, J. Enhancing Molecular Shape Comparison by Weighted Gaussian Functions. *Journal of Chemical Information and Modeling*, 2013, 53, 1967–1978.

(15) Tresadern, G.; Bemporad, D.; Howe, T. A comparison of ligand based virtual screening methods and application to corticotropin releasing factor 1 receptor. *Journal of Molecular Graphics and Modelling*, 2009, 860–870.

(16) Chu, S.; Gochin, M. Identification of fragments targeting an alternative pocket on HIV-1 gp41 by NMR screening and similarity searching. *Bioorganic & Medicinal Chemistry Letters*, 2013, 5114–5118.

(17) Kim, E.-S.; Cho, H.; Lim, C.; Lee, J.-Y.; Lee, D.-I.; Kim, S.; Moon, A. A natural piper-amide-like compound NED-135 exhibits a potent inhibitory effect on the invasive breast cancer cells. *Chemico-Biological Interactions*, 2015, 237, 58–65.

(18) Kossmann, B. R.; Abdelmalak, M.; Lopez, S.; Tender, G.; Yan, C.; Pommier, Y.; Marchand, C.; Ivanov, I. Discovery of selective inhibitors of tyrosyl-DNA phosphodiesterase 2 by targeting the enzyme DNA-binding cleft. *Bioorganic & Medicinal Chemistry Letters*, 2016, 3232–3236.
Woodring, J. L.; Bachovchin, K. A.; Brady, K. G.; Gallerstein, M. F.; Erath, J.; Tanghe, S.; Leed, S. E.; Rodriguez, A.; Mensa-Wilmot, K.; Sciotti, R. J.; Pollastri, M. P. Optimization of physicochemical properties for 4-anilinoquinazoline inhibitors of trypanosome proliferation. *European Journal of Medicinal Chemistry* 2017, 446–459.

Maccari, G.; Jaeger, T.; Moraca, F.; Biava, M.; Flohé, L.; Botta, M. A fast virtual screening approach to identify structurally diverse inhibitors of trypanothione reductase. *Bioorganic & Medicinal Chemistry Letters* 2011, 21, 5255–5258.

Kim, Y.-R.; Koh, H.-J.; Kim, J.-S.; Yun, J.-S.; Jang, K.; Lee, J.-Y.; Jung, J. U.; Yang, C.-S. Peptide inhibition of p22phox and Rubicon interaction as a therapeutic strategy for septic shock. *Biomaterials* 2016, 101, 47–59.

López-Ramos, M.; Perruccio, F. HPPD: Ligand- and Target-Based Virtual Screening on a Herbicide Target. *Journal of Chemical Information and Modeling* 2010, 801–814.

Hevener, K. E.; Mehboob, S.; Su, P.-C.; Truong, K.; Boci, T.; Deng, J.; Ghassemi, M.; Cook, J. L.; Johnson, M. E. Discovery of a Novel and Potent Class of *F. tularensis* Enoyl-Reductase (FabI) Inhibitors by Molecular Shape and Electrostatic Matching. *Journal of Medicinal Chemistry* 2012, 55, 268–279.

Kaoud, T. S.; Yan, C.; Mitra, S.; Tseng, C.-C.; Jose, J.; Taliaferro, J. M.; Tuohetahuntila, M.; Devkota, A.; Sammons, R.; Park, J.; Park, H.; Shi, Y.; Hong, J.; Ren, P.; Dalby, K. N. From in Silico Discovery to Intracellular Activity: Targeting JNK–Protein Interactions with Small Molecules. *ACS Medicinal Chemistry Letters* 2012, 3, 721–725.

Tiikkainen, P.; Markt, P.; Wolber, G.; Kirchmair, J.; Distinto, S.; Poso, A.; Kallioniemi, O. Critical Comparison of Virtual Screening Methods against the MUV Data Set. *Journal of Chemical Information and Modeling* 2009, 49, 2168–2178.

Massarotti, A.; Brunco, A.; Sorba, G.; Tron, G. C. ZINClick: A database of 16 million novel, patentable, and readily synthesizable 1,4-disubstituted triazoles. *Journal of Chemical Information and Modeling* 2014, 54, 396–406.

Oyarzabal, J.; Howe, T.; Alcazar, J.; Andrés, J. I.; Alvarez, R. M.; Dautzenberg, F.; Iturrino, L.; Martínez, S.; Van der Linden, I. Novel Approach for Chemotype Hopping Based on Annotated Databases of Chemically Feasible Fragments and a Prospective Case Study: New Melanin Concentrating Hormone Antagonists. *Journal of Medicinal Chemistry* 2009, 52, 2076–2089.

OpenEye, S. S. ROCS. *Santa Fe, NM. www.eyesopen.com*
(29) Banegas-Luna, A. J.; Cerón-Carrasco, J. P.; Puertas-Martín, S.; Pérez-Sánchez, H. BRUSELAS: HPC Generic and Customizable Software Architecture for 3D Ligand-Based Virtual Screening of Large Molecular Databases. *Journal of Chemical Information and Modeling* 2019, 59, 2805–2817.

(30) OpenEye, S. S. VIDA 4.4.0.4. Santa Fe, NM. www.eyesopen.com

(31) Jaccard, P. Distribution de la flore alpine dans le bassin des Dranses et dans quelques régions voisines. *Bulletin de la Société Vaudoise des Sciences Naturelles* 1901, 37, 241–272.

(32) Böttcher, C.; Belle, O. V.; Belle, B. *Theory of electric polarization*; Elsevier Scientific Pub. Co, 1974.

(33) Grant, J. A.; Pickup, B. T. A Gaussian Description of Molecular Shape. *The Journal of Physical Chemistry* 1995, 99, 3503–3510.

(34) Grant, J. A.; Gallardo, M. A.; Pickup, B. T. A fast method of molecular shape comparison: A simple application of a Gaussian description of molecular shape. *Journal of Computational Chemistry* 1996, 17, 1653–1666.

(35) OpenEye, S. S. Zap Toolkit. Santa Fe, NM. www.eyesopen.com

(36) Ellingson, B. A.; Skillman, A. G.; Nicholls, A. Analysis of SM8 and Zap TK calculations and their geometric sensitivity. *Journal of Computer-Aided Molecular Design* 2010, 24, 335–342.

(37) Thomas, D. G.; Chun, J.; Chen, Z.; Wei, G.; Baker, N. A. Parameterization of a geometric flow implicit solvation model. *Journal of Computational Chemistry* 2013, 34, 687–695.

(38) Hawkins, P. C. D.; Stahl, G. *Ligand-Based Methods in GPCR Computer-Aided Drug Design*; Springer New York, 2018; pp 365–374.

(39) Connelly, P. R.; Snyder, P. W.; Zhang, Y.; McClain, B.; Quinn, B. P.; Johnston, S.; Medek, A.; Tanoury, J.; Griffith, J.; Walters, W. P.; Dokou, E.; Knezić, D.; Bransford, P. The potency–insolubility conundrum in pharmaceuticals: Mechanism and solution for hepatitis C protease inhibitors. *Biophysical Chemistry* 2015, 196, 100–108.

(40) Gowthaman, R.; Lyskov, S.; Karanicolas, J. DARC 2.0: Improved Docking and Virtual Screening at Protein Interaction Sites. *PLOS ONE* 2015, 10, 1–24.

(41) Ivorra, B.; Ferrández, M.; Crespo, M.; Redondo, J.; Ortigosa, P.; Santiago, J.; Ramos, A. Modelling and optimization applied to the design of fast hydrodynamic focusing microfluidic mixer for protein folding. *Journal of Mathematics in Industry* 2018, 8, 4.

(42) Fernández, J.; Tóth, B.; Redondo, J.; Ortigosa, P.; Arrondo, A. A planar single-facility competitive location and design problem under the multi-deterministic choice rule. *Computers & Operations Research* 2017, 78, 305–315.
(43) Ferrández, M. R.; Redondo, J. L.; Ivorra, B.; Ramos, Á. M.; Ortigosa, P. M. Preference-based multi-objectivization applied to decision support for High-Pressure Thermal processes in food treatment. Applied Soft Computing 2019, 79, 326–340.

(44) Dakshanamurthy, S.; Issa, N. T.; Assefnia, S.; Seshasayee, A.; Peters, O. J.; Madhavan, S.; Uren, A.; Brown, M. L.; Byers, S. W. Predicting New Indications for Approved Drugs Using a Proteochemometric Method. Journal of Medicinal Chemistry 2012, 55, 6832–6848.

(45) Kumar, A.; Zhang, K. Y. J. Advances in the Development of Shape Similarity Methods and Their Application in Drug Discovery. Frontiers in Chemistry 2018, 6, 315.

(46) Yuan, S. et al. Structure-based discovery of clinically approved drugs as Zika virus NS2B-NS3 protease inhibitors that potently inhibit Zika virus infection in vitro and in vivo. Antiviral Research 2017, 145, 33–43.

(47) Wishart, D. S. et al. DrugBank 5.0: a major update to the DrugBank database for 2018. Nucleic Acids Research 2017, 46, D1074–D1082.

(48) Case, D. et al. AMBER 2017, 2017. San Francisco: University of California

(49) Halgren, T. A. Potential energy functions. Current opinion in structural biology 1995, 5, 205–210.

(50) Method for rapidly approximating similarities. Patent number: US8706427B2.
Optimizing electrostatic similarity for virtual screening: A new methodology.

Savíns Puertas-Martín,* Juana López Redondo, Horacio Pérez-Sánchez and Pilar Martínez Ortigosa