Virtual Screening Based on Electrostatic Similarity and Flexible Ligands

Savíns Puertas-Martín[1,3](✉), Juana L. Redondo[1], Antonio J. Banegas-Luna[2], Ester M. Garzón[1], Horacio Pérez-Sánchez[2], Valerie J. Gillet[3], and Pilar M. Ortigosa[1]

[1] Supercomputing - Algorithms Research Group (SAL), University of Almería, Agrifood Campus of International Excellence, ceiA3, 04120 Almería, Spain
{savinspm,jlredondo,gmartin,ortigosa}@ual.es

[2] Structural Bioinformatics and High Performance Computing Research Group (BIO-HPC), Universidad Católica de Murcia (UCAM), 30107 Murcia, Spain
ajbanegas@alu.ucam.edu, hperez@ucam.edu

[3] Information School, University of Sheffield, Sheffield S1 4DP, UK
v.gillet@sheffield.ac.uk

Abstract. Virtual Screening (VS) is a technique aimed at reducing the time and budget required when working on drug discovery campaigns. The idea consists of applying computational procedures to prefilter databases to a subset of potential compounds, to be characterized experimentally in later phases.

The problem lies in the fact that the current VS methods make simplifications, meaning they are not exhaustive. One particular common simplification is to consider the molecules as rigid. Such an assumption greatly reduces the computational complexity of the optimization problem to be solved, but it may result in poor or inefficient predictions. In this work, we have extended the features of Optipharm, a recently developed piece of software, by applying a methodology that considers the flexibility of the molecules. The new OptiPharm has several strengths over its previous version. More precisely, (i) it includes a prefilter based on molecule descriptors, (ii) simulates molecule flexibility by computing different poses for each rotatable bond, (iii) reduces the search space dimension, and (iv) introduces circular limits for the angular variables to enhance searchability. As the results show, these improvements help OptiPharm to achieve better predictions.

Keywords: Ligand based virtual screening · Molecule's flexibility · Optimization

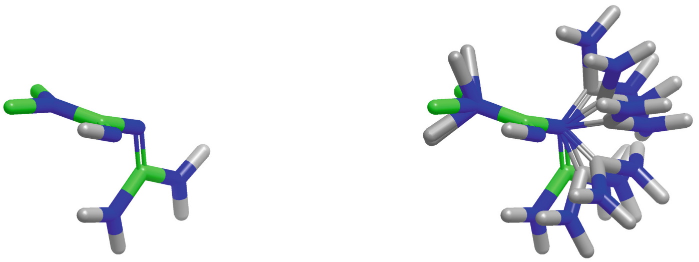
1 Introduction

Virtual Screening (VS) methods can be divided into structure-based (SBVS) and ligand-based (LBVS) methods [4,12,14] This work focuses on similarity LBVS methods [2,16]. In these techniques, the starting point is a source drug whose

O. Gervasi et al. (Eds.): ICCSA 2022 Workshops, LNCS 13378, pp. 127–139, 2022.
https://doi.org/10.1007/978-3-031-10562-3_10

shape, electrostatic potential, or other descriptor is known. This source ligand or crystal will be the target, and the virtual screening methods try to find the more similar molecules in an extensive database or chemolibrary. When calculating the electrostatic similarity between the target and a compound in the database, the more used methodology in the literature consists of optimizing in terms of shape by using Rapid Overlay of Chemical Structures (ROCS) [11], selecting a number N of compounds with the highest shape similarity values, and then evaluate them in terms of electrostatic similarity. Based on the assumption that a more realistic description of the compound bioactivity during the optimization procedure may help to obtain better predictions, a new version of OptiPharm was implemented in [10], which involves the direct optimization of the electrostatic similarity. As the results showed, the new methodology provided better predictions in electrostatic potential than the classical ones. In this work, we go a step forward and propose new improvements in OptiPharm aimed to reach even better predictions. To do so, we firstly include the flexibility of the molecules in the optimization procedure.

Protein flexibility is necessary for metabolism, transport, and function biological effects. Except for simple molecules such as O_2, both ligands and receptors are flexible molecules, which means that there is not a single three-dimensional representation of these molecules, but many. The conformational richness increases exponentially with the molecule's size, i.e., the more atoms (and therefore bonds, angles, and torsions) it possesses, the more degrees of freedom there are. These degrees of freedom are not additive but multiplicative, giving rise to many possible conformational states (see Fig. 1).



(a) A molecule of the target DB00331 that has some rotable bonds.

(b) A set of conformation generated from the rigid DB00331 molecule.

Fig. 1. A rigid molecule (a) can generate different conformations (b). An example for the DB00331 target from the DrugBank database is shown here. The Target structure has been painted green for both figures. (Color figure online)

For this reason, most of the studies are based on ligands where flexibility is considered to assume the protein to be almost rigid or with partial flexibility, so that they only rotate a maximum of the possible rotatable bonds [1,5,7]. In some cases, the solution is to perform the Virtual Screening considering the molecule as rigid, and then apply a process where the flexibility is studied for the number of rotatable bonds allowed by the algorithm [5]. This process sometimes consists of varying in a discrete way each of the angles of the rotatable bonds to find the best solution. Following any of these methods, the computational time is reduced, but many solutions keep unexplored. What we propose in this work is a previous analysis of the molecules. First, many descriptors are calculated, and the most representatives are selected to compute the difference between the target and the molecules in the database. Then best molecules are filtered and selected as flexible molecules and are applied a conformational generation process. This methodology allows exploring all the conformation of each compound widely but saves time by discarding uninteresting molecules.

Apart from considering the flexibility of the molecules, the new OptiPharm incorporates mechanisms of interest that enhance the search and helps to reduce the computational cost. As the results will show, all these improvements help provide better predictions in the molecules.

The rest of the paper is organized as follows. Section 2 describes the scoring function considered in this study and resumes the main ideas of OptiPharm, focusing on the new procedures and strategies developed. Section 3 summarizes the computational and scientific context taken into consideration for the experiments. Finally, Sects. 4 and 5 show the main results and conclusions inferred.

2 Methods

2.1 Electrostatic Similarity Scoring Function

The electrostatic similarities are obtained by numerical solution of the Poisson equation [3], viz:

$$\nabla\{\epsilon(r)\nabla\phi(r)\} = -\rho_{mol}(r) \tag{1}$$

where $\phi(r)$ 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}:

$$E_{AB} = \int \phi^A(r)\phi^B(r)\Theta^A(r)\Theta^B(r)\mathbf{dr} \approx h^3 \sum_{ijk} \phi^A_{ijk}\phi^B_{ijk}\Theta^A_{ijk}\Theta^B_{ijk} \tag{2}$$

where Θ is a masking function to ensure potentials interior to the compound are not considered part of the comparison. The integral appearing in (2) is a volume integral, computed using a grid-spacing parameter, h.

Notice that the accuracy obtained from (2) depends on the number of atoms in the two compared molecules. To measure the similarity between compounds, regardless of the number of atoms that they are composed of and the descriptor used, the Tanimoto Similarity [6] value is computed as follows:

$$Tc_E = \frac{E_{AB}}{E_{AA} + E_{BB} - E_{AB}} \tag{3}$$

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.

2.2 OptiPharm Algorithm

OptiPharm is a recent software designed explicitly for LBVS problems. It implements a global evolutionary optimizer capable of calculating the similarity between two compounds, a target and a query. To do so, it uses different methods in the optimization process to gradually adjust the position of the query while the target fixes its position. The interested reader is referred to [9,10] for an in-depth description of the original algorithm. In this work, we present a new version of OptiPharm. In the following, we briefly describe the new contributions.

To explore the solution space, OptiPharm works with a user-defined population of size M, which applies reproduction, selection, and improvement methods to each member of the population. A member or solution of this population represents the rotation and translation of the query molecule. Originally ten parameters were used to represent this modification, which means to work in a 10-dimensional search space. This paper presents a new version of OptiPharm, where the search space is reduced to 6 dimensions. The main change consists of replacing the use of quaternions with a semi-sphere parametrization, which simplifies the definition of the rotation axis. Consequently, searchability is enhanced due to the reduction of the search space dimension. Nevertheless, not only that, this new system avoids the repetition of the same rotation axis already explored.

This new mechanism provides improved freedom of exploration. In addition to reducing input parameters, the new version incorporates some problem knowledge, such as a mechanism to keep the angular variables between 0 and 2π in a continuous circular. So, if during the search an angle α takes a value greater than 2π, it is updated to the $\alpha - 2\pi$ value. In the previous version of OptiPharm, this value was updated to the maximum value of 2π.

2.3 Methodology

Procedure for Rigid Molecules

The process is trivial when working with rigid molecules and will be referred to throughout this paper under the name *Rigid*. As explained in the previous section, OptiPharm allows getting the best overlapping between two molecules, the target, and the query, to maximize the electrostatic similarity score. Consequently, when this procedure is repeated for each molecule in the database, their similarity score can be known. After that, the last step sorted the molecules by their similarity value. This procedure returns the most similar ones of interest since they can be successful potential drugs because they are the most similar to the target. Figure 2 shows this process to obtain a ranked list of compounds.

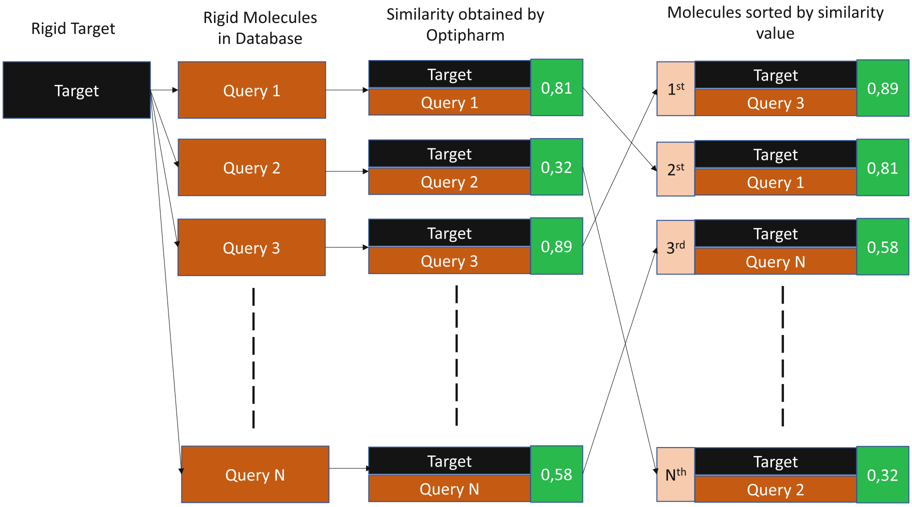


Fig. 2. Procedure to rank rigid molecules.

Procedure for Molecular Conformations

Working with flexible molecules with some rotational bonds implies modifying the methodology to obtain the similarity between a target molecule and a query molecule in the database. Multiple alternative conformers of this molecule are constructed by modifying the rotatable bonds with various rotation angles. This procedure simulates the flexibility in the rotatable bonds of a given molecule. However, the number of molecules in the database grows dramatically, and consequently, so does time.

A solution to this problem is to first discard those not promising compounds in the database and then generate conformations to the remaining molecules. This process, which we have called *Flexible* throughout the document, is explained in Fig. 3. In this figure, first, the descriptors are calculated for each molecule in the database. In this work, more than 4,000 different descriptors are obtained. However, many are not relevant or are repetitive, so different machine learning metrics are applied to discard them. In particular, variance and correlation are applied, and those descriptors whose values are 0 in most cases have been removed. This filter reduces the number of descriptors to a more limited number.

The obtained descriptors are then used to filter the molecules in the database. As can be seen in Fig. 4, the Euclidean distance between the query molecule and each of the targets is calculated. Once they are available, the compounds are ordered from smallest to largest by the distance value, and the best M compounds are selected based on an empirical cutoff value. Finally, several conformations are generated for the selected query compounds and the target.

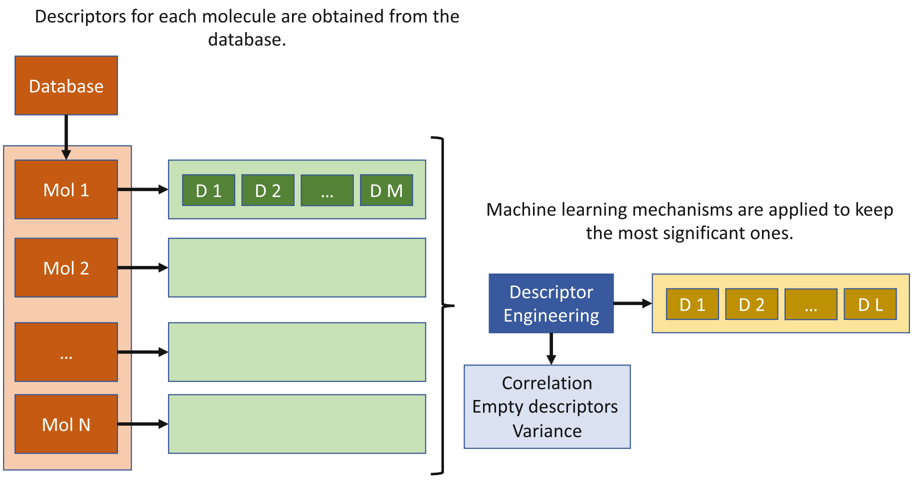

Descriptors for each molecule are obtained from the database.
Database
Mol 1
Mol 2
...
Mol N
D 1
D 2
...
D M
Machine learning mechanisms are applied to keep the most significant ones.
Descriptor Engineering
Correlation
Empty descriptors
Variance
D 1
D 2
...
D L

Fig. 3. Descriptor selection procedure.

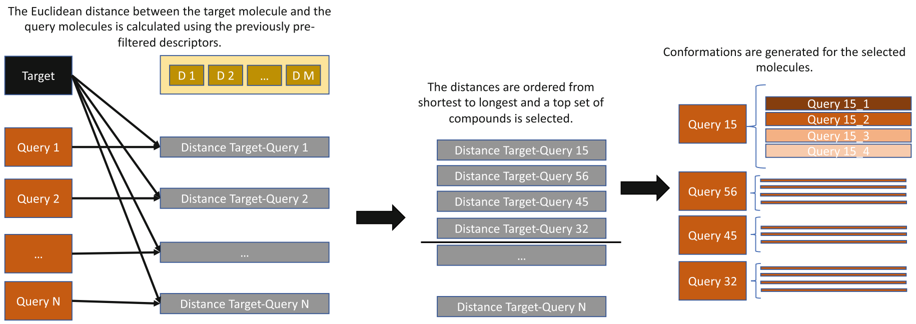


Fig. 4. Molecule filtering process based on Euclidean distance and subsequent generation of conformations of selected molecules.

After generating multiple conformations for target and query, an optimization procedure is run using OptiPharm. Figure 5 shows such a procedure using an example where only three conformations have been generated for both the target and the query.

As shown in Fig. 5, an extensive comparison is performed, which involves running $nt \times nq$ times OptiPharm algorithm instead of just one for rigid molecules. In this exhaustive comparison, nt represents the number of conformations of the target molecule, and nq is the number of conformations of the query molecule. Once the maximum similarity of each of the comparisons has been calculated, the algorithm searches for the highest value and provides it as the final similarity result between the two flexible molecules.

Once the flexible target has been compared with all molecules in the database, they are ordered based on their conformation computed similarity value.

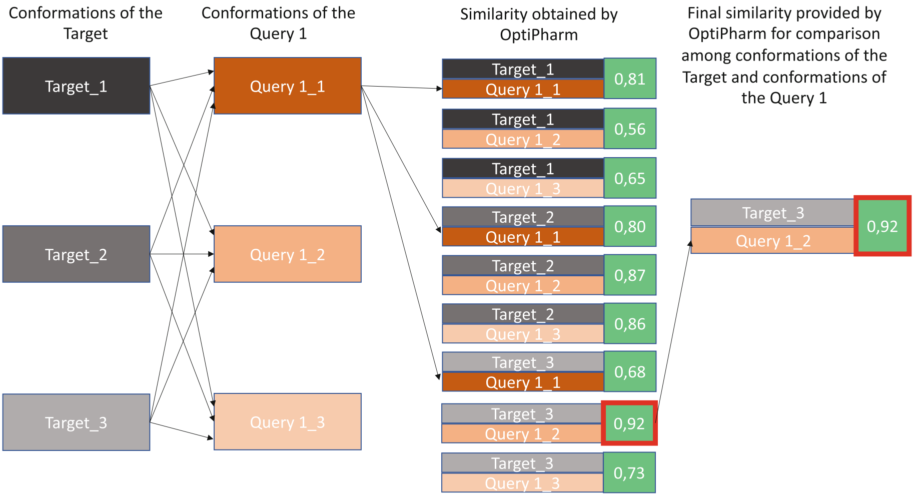


Fig. 5. Procedure for obtaining maximum similarity when working with conformations of molecules.

Consequently, new query compounds with a high similarity value can be identified while they are not detected when working with rigid molecules.

3 Materials

Hardware. The experiments of this work have been carried out using a cluster of 8x Bull Sequana X440-A5: 2 AMD EPYC Rome 7642 (48 cores) and 512 GB of RAM memory and 240 GB SSD.

FDA Database. The database used in this work was obtained from Drugbank, v.5.0.1 [15]. Specifically, a subset of 1,751 molecules validated by the Food and Drug Administration (FDA) has been used. The FDA is a federal agency of the U.S. 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 (medicines). The original database has been downloaded from https://go.drugbank.com.

Software. The new version of OptiPharm, described in Sect. 2.2 is the optimization algorithm used to find the maximum similarity between two compounds. It has been configurated to consider the hydrogen atoms of each molecule. In addition, all the heavy atom radii have been set to 1.7Å. Furthermore, all compound pairs are centered and aligned. Consequently, the molecule centroids have been located at the coordinates center of the search space. Finally, each molecule has been aligned so that its longest axis has been oriented at X-axis and the shortest along the Z-axis. The input parameter set used in OptiPharm have been:

$N = 200,000$ function evaluations, $M = 5$ starting poses, $t_{max} = 5$ iterations, and $R = 1$ as the smallest possible radius.

Additionally, software OMEGA [8] has been the generator selected to obtain the conformations of targets and queries in the database. The maximum number of conformations for a given compound was limited to 500, though the obtained number was smaller in many compounds due to a small number of rotatable bonds.

Finally, Dragon (6.0.38) has been used to calculated 4885 descriptors for each molecule in the database.

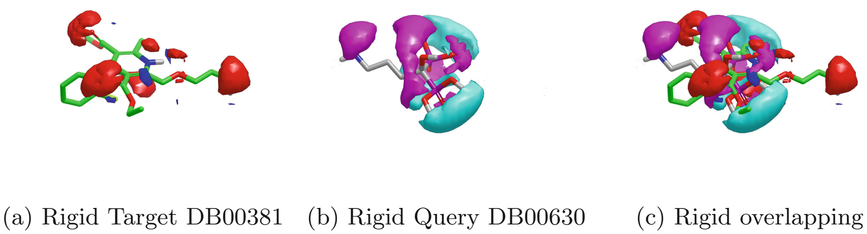
4 Results

In this section, we will show the results obtained by the new methodology, and we will compare them with the ones obtained for the original OptiPharm in [10]. For illustration, we will only depict the outcomes obtained for the molecules DB00381 and DB00876.

Table 1. Top-10 most similar compounds in electrostatic to the target DB00381.

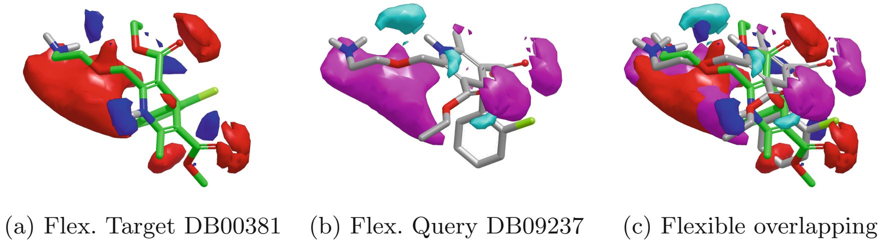
Rigid		*Flexible*				
Query	Tc_E	Target conformation	Query conformation	Tc_E	Rk_R	Tc_E^R
DB00630	0.377	32	DB09237_43	0.762	1406	0.118
DB00409	0.377	32	DB01214_481	0.727	1384	0.121
DB00751	0.374	32	DB00622_175	0.718	999	0.207
DB00933	0.374	32	DB00557_212	0.704	1186	0.178
DB00998	0.370	32	DB00383_159	0.700	398	0.264
DB00334	0.367	32	DB01244_44	0.689	1549	0.105
DB00891	0.359	32	DB00979_23	0.683	517	0.254
DB00611	0.358	32	DB00571_483	0.679	240	0.280
DB00540	0.358	32	DB01359_440	0.666	1268	0.153
DB00647	0.357	32	DB00748_58	0.665	260	0.278

Following our methodology, we first calculate the 4885 descriptors for all molecules in the database. Subsequently, this group is reduced to 757 by applying the statistical metrics. With these numbers, it is computed the Euclidean distance between each molecule and the target. Later, the molecules are shorted according to that distance in ascending order. Only those molecules ranked within 10% of the shortest distance were selected. For the 175 molecules that remain, several conformations are generated. In particular, the sub-database obtained for each target consists of 47,983 and 38,833 conformations for the targets DB00381 and DB00876, respectively. In addition, target DB00381 resulted in 383 conformations, and DB00876 in 154.



(a) Rigid Target DB00381 (b) Rigid Query DB00630 (c) Rigid overlapping

Fig. 6. Maximum similarity solution for Target DB00381 when working with rigid molecules. Figures (a) and (b) represent the target and query compounds and their electrostatic fields. Figure (c) represents the optimal overlapping between the two compounds where Tc_E is maximum.



(a) Flex. Target DB00381 (b) Flex. Query DB09237 (c) Flexible overlapping

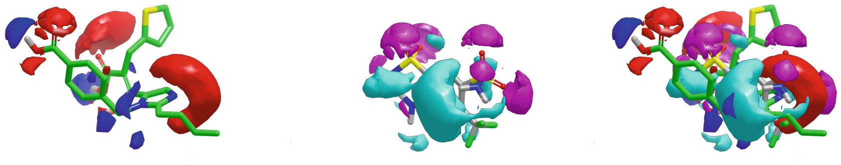
Fig. 7. Maximum similarity solution for Target DB00381 when working with flexible molecules. Figures (a) and (b) represent the more similar conformations of target and query compounds and their electrostatic fields. Figure (c) represents the optimal overlapping between the two compounds where Tc_E is maximum.

Tables 1 and 2 compare the main results obtained for the methodologies *Rigid* and *Flexible* for Targets DB00381 and DB00876. In particular, they show the 10 queries with the greatest similarity provided for both versions. More precisely, for *Rigid* methodology, we provide its name, *Query*, and the corresponding similarity value Tc_E. For *Flexible*, we indicate the pair target-query that has obtained the best match, i.e. we identify those two molecules by also indicating their corresponding conformation number. This information is depicted in columns *Target conformation* and *Query conformation*. Finally, we show at column Tc_E the scoring function value obtained for each match. For the sake of comparison, we also indicate in column Rk_R the position that the *Query conformation* occupies in the list obtained by *Rigid*, and its corresponding scoring value in column Tc_E^R.

The results show an improvement in the quality of the solutions. As can be seen in Table 1, the most similar compound found following the *Flexible* methodology (0.762) improves twice the value of the *Rigid* (0.377). Moreover, the most similar compounds in the rankings are different, i.e., DB00630 for the *Rigid* and DB09237 for the *Flexible*. Additionally, this result can be seen

Table 2. Top-10 most similar compounds in electrostatic to the target DB00876.

Rigid		*Flexible*				
Query	Tc_E	Target Conformation	Query conformation	Tc_E	Rk_R	Tc_E^R
DB00774	0.532	69	DB00338_126	0.861	1164	0.225
DB00880	0.530	35	DB08897_40	0.861	1577	0.159
DB01153	0.527	18	DB00736_54	0.860	888	0.261
DB00690	0.526	69	DB06766_6	0.835	1403	0.197
DB00897	0.522	120	DB00966_424	0.832	772	0.276
DB00819	0.513	69	DB01129_440	0.829	713	0.285
DB01101	0.512	35	DB04843_72	0.813	1378	0.201
DB00425	0.507	1	DB04880_2	0.800	502	0.321
DB01002	0.505	103	DB00642_47	0.775	704	0.286
DB00809	0.503	35	DB06274_105	0.522	618	0.301



(a) Rigid Target DB00876 (b) Rigid Query DB00774 (c) Rigid overlapping

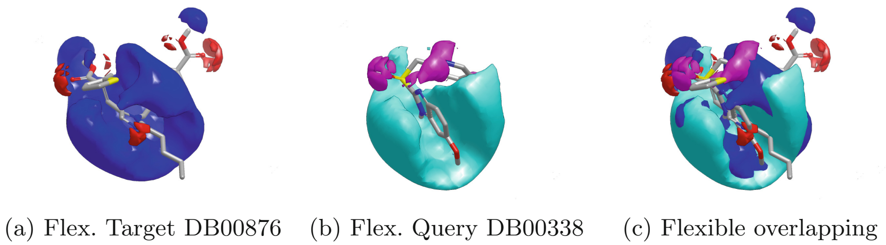
Fig. 8. Maximum similarity solution for Target DB00876 working with rigid molecules. Figures (a) and (b) represent the target and query compounds and their electrostatic fields. Figure (c) represents the optimal overlapping between the two compounds where Tc_E is maximum.

graphically in Figs. 6 and 7 where the molecules, their optimal overlapping, and electrostatic fields are represented using VIDA [13].

If the last two columns of *Flexible* are analyzed, it is clear that there is no good overlapping when rigid molecules are used. As seen in the Rk_R column, most of the top molecules in *Flexible* are below the position 1000th in the *Rigid* list.

Table 2 shows results along the same line as Table 1. In this study case, the top solution improves from 0.532 to 0.861 finding different compound as well. Figures 8 and 9 show graphically the most similar compounds for both methods.

The improvement obtained in the previous results is due to the solutions, including flexibility. However, as we previously stated, this increases the computational time considerably. However, the time has been ostensibly reduced thanks to the *Flexible* methodology employed in this work. The average time for an optimization process is 29 s for these target molecules in the



(a) Flex. Target DB00876 (b) Flex. Query DB00338 (c) Flexible overlapping

Fig. 9. Maximum similarity solution for Target DB00876 working with flexible molecules. Figures (a) and (b) represent the more similar conformations of target and query compounds and their electrostatic fields. Figure (c) represents the optimal overlapping between the two compounds where Tc_E is maximum.

current hardware. If we focus on the target DB00381 with a database of 38,833 conformations from the initial 175 molecules, each molecule generated on average 221 conformations, although the maximum could be 500 for each one). If this value is extrapolated to the whole database, $(1751 * 221 =)386,971$ conformations could be generated. So, the time saved with the filter applied by discarding unpromising compounds is 116 days per target, and whether 500 conformations are generated, 264 computation days would be saved.

5 Conclusions and Future Work

In this work, we have improved the software OptiPharm by considering molecule flexibility. Apart from that, the new version includes several mechanisms to reduce the computational effort. In particular, we have reduced the number of optimization parameters and the range of freedom in some of them. Consequently, the search space decreases, and the number of function evaluations needed to find the optimal similarity drops. Besides, we have analyzed and applied descriptors to filter the initial database. We have used statistical metrics such as variance, correlation, or Euclidean distance.

The results have shown that the new OptiPharm can obtain solutions with higher scoring values than the original one, meaning that new query compounds with a high similarity value can be identified. These compounds are not detected when working with rigid molecules. In addition, the descriptor filter allows to drastically reduce the run time, saving for the study at hand up to 264 days.

Future work proposes implementing a conformation generation algorithm as an internal procedure of OptiPharm and including new scoring functions.

Acknowledgement. This work was supported by the Spanish Ministry of Economy and Competitiveness through the CTQ2017-87974-R, RTI2018-095993-B-I00 and EQC2019-006418-P grants; by the Junta de Andalucía through the grant Proyectos de excelencia (P18-RT-1193), by the Programa Regional de Fomento de la Investigación (Plan de Actuación 2018, Región de Murcia, Spain) through the “Ayudas a la realización de proyectos para el desarrollo de investigación científica y técnica por grupos competitivos (20988/PI/18)" grant; by the University of Almeria throught the "Ayudas a proyectos de investigación I+D+I en el marco del Programa Operativo FEDER 2014-20" grant (UAL18-TIC-A020-B). Savíns Puertas Martín is a fellow of the "Margarita Salas" grant (RR_A_2021_21), financed by the European Union (NextGenerationEU).

References

1. Axenopoulos, A., Rafailidis, D., Papadopoulos, G., Houstis, E.N., Daras, P.: Similarity search of flexible 3D molecules combining local and global shape descriptors. IEEE/ACM Trans. Comput. Biol. Bioinf. **13**(5), 954–970 (2016). https://doi.org/10.1109/TCBB.2015.2498553
2. Bahi, M., Batouche, M.: Deep learning for ligand-based virtual screening in drug discovery. In: 2018 3rd International Conference on Pattern Analysis and Intelligent Systems (PAIS), pp. 1–5 (2018). https://doi.org/10.1109/PAIS.2018.8598488
3. Böttcher, C., Belle, O.V., Belle, B.: Theory of Electric Polarization. Elsevier, Amsterdam (1974). https://doi.org/10.1016/B978-0-444-41019-1.50006-7
4. Fatumo, S., Adebiyi, M., Adebiyi, E.: In silico models for drug resistance. In: Kortagere, S. (eds.) In Silico Models for Drug Discovery. Methods in Molecular Biology, vol. 993. Humana Press, Totowa (2013). https://doi.org/10.1007/978-1-62703-342-8_4
5. Hu, J., Liu, Z., Yu, D.J., Zhang, Y.: LS-align: an atom-level, flexible ligand structural alignment algorithm for high-throughput virtual screening. In: Bioinformatics, vol. 34, pp. 2209–2218. Oxford University Press (2018). https://doi.org/10.1093/bioinformatics/bty081
6. 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 **37**, 241–272 (1901)
7. Kalászi, A., Szisz, D., Imre, G., Polgár, T.: Screen3D: a novel fully flexible high-throughput shape-similarity search method. J. Chem. Inf. Model. **54**(4), 1036–1049 (2014). https://doi.org/10.1021/ci400620f
8. OMEGA 4.1.0.2: OpenEye Scientific Software: Santa Fe, NM, USA (2019). http://www.eyesopen.com
9. Puertas-Martín, S., Redondo, J.L., Ortigosa, P.M., Pérez-Sánchez, H.: OptiPharm: an evolutionary algorithm to compare shape similarity. Sci. Rep. **9**(1), 1398 (2019). https://doi.org/10.1038/s41598-018-37908-6
10. Puertas-Martín, S., Redondo, J.L., Pérez-Sánchez, H., Ortigosa, P.M.: Optimizing electrostatic similarity for virtual screening: a new methodology. In: Informatica, pp. 1–19 (2020). https://doi.org/10.15388/20-INFOR424
11. ROCS: OpenEye Scientific Software: Santa Fe, NM. http://www.eyesopen.com
12. Tanrikulu, Y., Krüger, B., Proschak, E.: The holistic integration of virtual screening in drug discovery. Drug Discov. Today **18**(7–8), 358–364 (2013). https://doi.org/10.1016/j.drudis.2013.01.007
13. VIDA 4.4.0.4: OpenEye Scientific Software: Santa Fe, NM. http://www.eyesopen.com
14. Vázquez, J., López, M., Gibert, E., Herrero, E., Luque, F.J.: Merging ligand-based and structure-based methods in drug discovery: an overview of combined virtual screening approaches. Molecules **25**(20), 4723 (2020). https://doi.org/10.3390/molecules25204723

15. Wishart, D.S., et al.: DrugBank 5.0: a major update to the DrugBank database for 2018. Nucl. Acids Res. **46**(D1), D1074–D1082 (2018). https://doi.org/10.1093/nar/gkx1037
16. Yang, Y., et al.: Ligand-based approach for predicting drug targets and for virtual screening against COVID-19. Brief. Bioinform. **22**(2), 1053–1064 (2021). https://doi.org/10.1093/bib/bbaa422