

# Optimizing ligand conformations in flexible protein targets: a multi-objective strategy

Esteban López-Camacho<sup>1</sup> María Jesús García-Godoy<sup>1</sup> José García-Nieto<sup>1</sup> Antonio J. Nebro<sup>1</sup> José F. Aldana-Montes<sup>1</sup>



## Abstract

Finding the orientation of a ligand (small molecule) with the lowest binding energy to the macromolecule (receptor) is a complex optimization problem, commonly called ligand–protein docking. This problem has been usually approached by minimizing a single objective that corresponds to the final free energy of binding. In this work, we propose a new multi-objective strategy focused on minimizing: (1) the root mean square deviation (RMSD) between the co-crystallized and predicted ligand atomic coordinates, and (2) the ligand–receptor intermolecular energy. This multi-objective strategy provides the molecular biologists with a range of solutions computing different RMSD scores and intermolecular energies. A set of representative multi-objective algorithms, namely NSGA-II, SMPSO, GDE3 and MOEA/D, have been evaluated in the scope of an extensive set of docking problems, which are featured by including HIV-proteases with flexible ARG8 side chains and their inhibitors. As use cases for biological validation, we have included a set of instances based on new retroviral inhibitors to HIV-proteases. The proposed multi-objective approach shows that the predictions of ligand’s pose can be promising in cases in which studies *in silico* are necessary to test new candidate drugs (or analogue drugs) to a given therapeutic target.

**Keywords** Molecular docking · Multi-objective optimization · Metaheuristics

## 1 Introduction

Ligand–protein docking is an optimization problem whose main objective is to predict a position and orientation of a small molecule (ligand) to a larger receptor molecule (macromolecule) with minimum binding energy. Molecular docking is based on two steps: (1) the prediction of the position and orientation of the ligand to the receptor and (2) the assessment of the ligand–receptor binding affinity that is obtained. This technique has been used to do studies *in silico* in drug discovery.

In the last years, molecular docking has been focused on minimizing the final free energy binding using single-

objective algorithms by using a force-field-based energy function (Ru et al. 2016; Peh and Hong 2016). In accordance with other past studies (López-Camacho et al. 2015), the differential evolution algorithm (DE) returned the best solutions in terms of energy and RMSD score compared to steady-state genetic algorithm (ssGA), generational genetic algorithm (gGA) and PSO (particle swarm optimization). A recent approach is DockThor (Leonhart et al. 2018), which involves a single-objective approach based on minimizing the final binding energy. DockThor was compared with a set of mono-objective approaches presented in the state of the art, including those of jMetalCpp (López-Camacho et al. 2014). The results showed that DockThor outperforms the selected jMetalCpp algorithm.

In parallel, a series of studies based on multi-objective approaches to solve this problem have emerged (Gu et al. 2015; Janson et al. 2008; Sandoval-Perez et al. 2013). In López-Camacho et al. (2016), we presented a new multi-objective approach to minimize two objectives based on the RMSD score and the intermolecular energy ( $E_{\text{inter}}$ ). The first one corresponds to a similarity measure between the reference and the co-crystallized ligands. The second one is an

---

✉ José F. Aldana-Montes  
jfam@lcc.uma.es

<sup>1</sup> Department of Computer Languages and Computing Sciences, Institute for Software Technologies and Software Engineering (ITIS), Biomedical Research Institute of Málaga (IBIMA), University of Málaga ETSI Informática, Campus de Teatinos, 29071 Málaga, Spain

aggregation of bound and unbound states of the molecular complex. This approach was applied to a set of 11 instances that involve flexible receptors and inhibitors. In that study, a set of representative multi-objective algorithms were compared and SMPSO (Nebro et al. 2009) was the optimizer that obtained the best overall results for the tested molecular docking instances.

In the present study, we follow our previous research line by performing a thorough experimentation, with the motivation of assessing whether our initial claims could be generalized or not in the context of a large set of problems. We have performed an extended research from our previous study (López-Camacho et al. 2016), involving 73 instances used as benchmark by a previous study such as Morris et al. (2009). The macromolecules are therapeutic targets in which flexibility was applied. The ligands have different sizes (small, medium and large), and the obtained results can be analyzed taking into account the ligand size.

Our motivation is then to provide molecular biologists with a set of trade-off solutions with different RMSD scores and  $E_{\text{inter}}$ . It is worth noting that the optimized ligand structure, even with outstanding RMSD scores (measured in Å), does not necessarily correspond to the already known best solution. From a pharmacological point of view, the macromolecule (the therapeutic target) might have other different ligand binding sites that do not correspond to the co-crystallized ligand binding site obtained from in silico studies, for example, the case of allosteric modulation, which refers to the alteration of the macromolecule functionality by the binding of a modulator (a ligand) at a distant binding site (Townsend et al. 2015). Other examples refer to those cases where the ligand docking avoids the enzyme–substrate interaction and its binding site is different from the co-crystallized ligand. In this context, the approach proposed in this paper provides a wide range of solutions that the practitioner can choose, i.e., a solution that corresponds to a ligand conformation with a higher RMSD score and lower  $E_{\text{inter}}$ . The ligand related to this solution can have a position close to the entrance of the macromolecule’s active site or to some of molecular sites with known allosteric modulation. For this approach, the search methods used are several representative multi-objective algorithms which are: SMPSO, GDE3, MOEA/D and NSGA-II. Finally, in order to show the usability of this multi-objective strategy, we have chosen a set of flexible instances that correspond to flexible macromolecule–ligand complexes. In these examples, several solutions from the front of non-dominated solutions have been chosen and analyzed in terms of ligand binding site and molecular interactions.

This paper is organized as follows: Sect. 2 describes the ligand–protein docking problem from a multi-objective formulation. Section 3 describes the optimization techniques evaluated. Experimental methodology is explained in Sect. 4,

while analysis of results and discussions are reported in Sect. 5. Section 6 contains conclusions and future work.

## 2 Multi-objective ligand–protein docking

Previously to describe the ligand–protein docking problem, we include a series of formal definitions concerning multi-objective optimization. Without loss of generality, we assume that minimization is the goal for all the objectives.

**Definition 1** (*Multi-objective Optimization Problem*) Find a vector  $\mathbf{x}^* = [x_1^*, x_2^*, \dots, x_n^*]$  which satisfies the  $m$  inequality constraints  $g_i(\mathbf{x}) \geq 0, i = 1, 2, \dots, m$ , the  $p$  equality constraints  $h_i(\mathbf{x}) = 0, i = 1, 2, \dots, p$ , and minimizes the vector function  $\mathbf{f}(\mathbf{x}) = [f_1(\mathbf{x}), f_2(\mathbf{x}), \dots, f_k(\mathbf{x})]^T$ , where  $\mathbf{x} = [x_1, x_2, \dots, x_n]^T$  is the vector of decision variables.

The set of all values satisfying the constraints defines the *feasible region*  $\Omega$  and any point  $\mathbf{x} \in \Omega$  is a *feasible solution*. We seek the *Pareto optima*.

**Definition 2** (*Pareto Optimality*) for every  $\mathbf{x} \in \Omega$  and  $I = \{1, 2, \dots, k\}$  either  $\forall i \in I (f_i(\mathbf{x}) = f_i(\mathbf{x}^*))$  or there is at least one  $i \in I$  such that  $f_i(\mathbf{x}) > f_i(\mathbf{x}^*)$ .

This definition states that  $\mathbf{x}^*$  is Pareto optimal if no other feasible vector  $\mathbf{x}$  exists which would improve some criteria without causing a simultaneous worsening in at least one other criterion.

**Definition 3** (*Pareto Dominance*) A vector  $\mathbf{u} = (u_1, \dots, u_k)$  is said to dominate  $\mathbf{v} = (v_1, \dots, v_k)$  (denoted by  $\mathbf{u} \preceq \mathbf{v}$ ) if and only if  $\mathbf{u}$  is partially less than  $\mathbf{v}$ .  $\forall i \in \{1, \dots, k\}, u_i \leq v_i \wedge \exists i \in \{1, \dots, k\} : u_i < v_i$ .

### 2.1 Ligand–protein docking

The main objective in this problem is to find a ligand’s conformation ( $L$ ) to a receptor ( $R$ ) that results with minimum energy. An energy scoring function is then calculated to evaluate the ligand–receptor interaction according to three components, which represent degrees of freedom:

- The first one corresponds to the translation of the ligand, which is represented by the three position values ( $x, y, z$ ) in the axis of the Cartesian coordinate space;
- The second one includes the ligand’s orientation, which is modeled as a four variables quaternion including the angle slope ( $\theta$ );
- The third one represents the dihedral angles referring to the free rotation of torsion of the ligand and side chains of the receptor’s residues.

Each problem solution is encoded by a real-value vector of  $7 + n$  variables, in which the first three values are used for ligand translation, the following four variables represent the ligand and/or macromolecule orientation, and the last  $n$  variables represent the dihedral angles of ligand and receptor positions. In addition, a grid-based methodology is implemented (Morris et al. 2009) to perform the energy computation of molecular conformations.

More specifically, the  $E_{\text{inter}}$  is one of the force-field-based energy function terms used in AutoDock 4.2. This energy term is calculated as follows:

$$E_{\text{inter}} = Q_{\text{bound}}^{R-L} + Q_{\text{unbound}}^{R-L} \quad (1)$$

$Q_{\text{bound}}^{R-L}$  and  $Q_{\text{unbound}}^{R-L}$  are the states of bound and unbound of the ligand–receptor complex, respectively.

$$\begin{aligned} Q = & W_{\text{vdw}} \sum_{i,j} \left( \frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^6} \right) \\ & + W_{\text{hbond}} \sum_{i,j} E(t) \left( \frac{C_{ij}}{r_{ij}^{12}} - \frac{D_{ij}}{r_{ij}^{10}} \right) \\ & + W_{\text{elec}} \sum_{i,j} \frac{q_i q_j}{\epsilon(r_{ij}) r_{ij}} \\ & + W_{\text{sol}} \sum_{i,j} (S_i V_j + S_j V_i) e^{(-r_{ij}^2/2\sigma^2)} \end{aligned} \quad (2)$$

These energetic evaluation terms ( $Q$ ) take into account dispersion/repulsion (vdw), hydrogen bonds (hbond), electrostatics (elec) and desolvation (sol). In Eq. 2, constants for van der Waals, hydrogen bonds, torsional forces, electrostatic interactions and desolvation are included by means of weights  $W_{\text{vdw}}$ ,  $W_{\text{hbond}}$ ,  $W_{\text{conf}}$ ,  $W_{\text{elec}}$ , and  $W_{\text{sol}}$ , respectively. Value  $r_{ij}$  is the interatomic distance. Lennard–Jones parameters taken from the Amber force field are represented by  $A_{ij}$  and  $B_{ij}$  values, whereas  $C_{ij}$  and  $D_{ij}$  indicate the maximum well depth of potential energies between two atoms. The angle-dependent directionality is represented by  $E(t)$ . The third term computes a Coulomb approach for electrostatics, and the fourth term involves the volume ( $V$ ) of surrounding atoms of a given one weighted by  $S$ , with an exponential factor computing atom distances. A further explanation of the  $E_{\text{inter}}$  computation can be obtained from Morris et al. (2009).

The RMSD is a quantitative measure of similarity between two superimposed atomic coordinates of the co-crystallized ligand’s conformation and the predicted position of the ligand. In a nutshell, the lower the RMSD score, the better the solution is (if the co-crystallized ligand is considered as reference).

According to the literature (Trott and Olson 2010; Kufareva and Ruben 2010), the RMSD cutoff of 2 Å is used as a criterion of the correct bound structure prediction. Ligands’

conformations with an RMSD score above 2 Å indicate a prediction that is not very accurate. However, it is worth noting that other docking solutions with high RMSD but low  $E_{\text{inter}}$  values would indicate alternative ligand–macromolecule interactions. These solutions should be indeed considered, since they can be interesting from a pharmacological point of view, as mentioned in Sect. 1.

$$\text{RMSD} = \sqrt{\frac{1}{N} \sum_i^n d_i^2} \quad (3)$$

where the averaging is performed over the  $n$  pairs of equivalent atoms and  $d_i$  is the distance between the two atoms in the  $i$ -th pair.

## 2.2 Proposed multi-objective formulation for ligand–protein docking

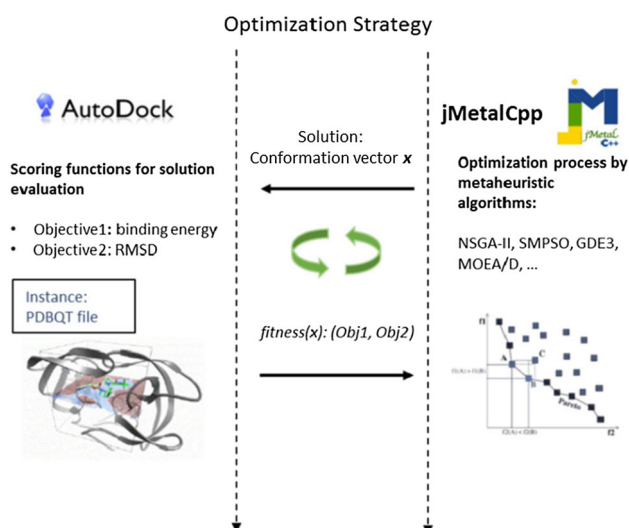
As mentioned, our goal is to formulate molecular docking as a bi-objective optimization problem by considering the two terms separately, although following a Pareto dominance scheme as follows:

- Objective 1: The  $E_{\text{inter}}$  as one of the energy function terms used in AutoDock 4.2, and expressed in Eq. 1.
- Objective 2: The RMSD (expressed in Eq. 3) as quantitative measure of similarity between two superimposed atomic coordinates of the co-crystallized ligand’s conformation and the predicted position of the ligand.

In this way, we can now take advantage of specific learning models of Pareto optimality-based techniques to deal with the molecular docking, which will result in sets of non-dominated solutions with different choices of energy and RMSD. In addition, we avoid the use of additional weight factors that could bias the search procedure to one of the different terms, as usually is seen in single-objective approaches.

## 3 Optimization strategy

Our optimization strategy includes two main processes acting in parallel: an optimization algorithm and a molecular docking procedure. The optimization part is carried out through a multi-objective algorithm provided by jMetalCpp (López-Camacho et al. 2014), an open-source framework of single/multi-objective optimization metaheuristics. The molecular binding procedure is performed by AutoDock 4.2 (Morris et al. 2009), a widely used tool for virtual drug discovery involving rigid and flexible macromolecule–ligand docking simulations.



**Fig. 1** Overall scheme of the optimization strategy. Solutions generated by multi-objective algorithms in jMetalCpp are then evaluated by AutoDock and finally sent back to jMetalCpp. Once the stop condition has been met, the best solution is returned to the AutoDock code which in turn generates an output file with the results. These results can be visualized with any tool that is suitable for visualizing AutoDock format (DLG) like AutoDockTool (ADT)

Therefore, jMetalCpp and AutoDock are integrated in such a way that the algorithms available in jMetalCpp can communicate with AutoDock to cooperate on the molecular docking optimization. This integration process consists of running AutoDock and jMetalCpp in two different threads inside the same task. This way, as the internal memory is shared, the two threads can communicate with each other and they synchronize using mutexes. This approach is efficient and flexible, allowing any of the algorithms included in jMetalCpp to be easily used for solving docking problems.

Figure 1 illustrates the optimization process, which algorithmic steps are also organized through the pseudocode of Algorithm 1. Therefore, after population and Pareto front are initialized (lines 2 to 4), while a given stop condition is not met (e. g., maximum number of iterations), a given metaheuristic algorithm in jMetal is executed to iteratively generate new solutions (line 6). Whenever the binding quality of a new solution has to be numerically quantified, it is sent to AutoDock to be evaluated in terms of binding energy and RMSD. After this evaluation, AutoDock returns the corresponding objective values to the algorithm, which are assigned to the evaluated solution (line 7). After this, the Pareto front is updated (line 8). This process is repeated until reaching a given stop condition. As a result, those non-dominated solutions found so far by the multi-objective algorithms are returned (Line 10). Finally, AutoDock is instanced to generate the output data concerning the optimized docking. Thus, the final results follow the standard format tailored to AutoDock users.

### Algorithm 1 Pseudocode of Common Algorithmic Template for Multi-objective Molecular Docking

---

```

1: alg ∈ {NSGA-II, SMPSO, GDE3, MOEA/D}
2: pop ← initializePopulation()
3: evaluateFitness(pop, Autodock()) // Autodock thread
4: pareto_front ← initializeParetoFront(pop)
5: while not StopCondition() do
6:   pop ← performNewAlgorithmIteration(pop, alg)
7:   evaluateFitness(pop, Autodock())
8:   pareto_front ← updateParetoFront(pareto_front, pop)
9: end while
10: return pareto_front

```

---

For this study, as we are interested on evaluating our multi-objective approach for molecular docking, we have focused on four representative multi-objective metaheuristics in the state of the art. In concrete, we have selected the most widely used algorithm in the field (NSGA-II), a swarm intelligence optimizer (SMPSO), a solver based on differential evolution (GDE3), and an algorithm based on decomposition (MOEA/D). A brief description of these techniques is given as follows:

- NSGA-II (Deb et al. 2002). The *Non-dominated Sorting Genetic Algorithm II* performs a generational strategy to obtain a new offspring population from the original one, by means of the application of genetic operators: selection, crossover and mutation. It uses a non-dominated sorting procedure based on Pareto ranking together with crowding distance density estimator to promote convergence and diversity, respectively. In spite of being one of the classical EMO algorithms in the literature, it still shows a prominent performance in multiple benchmarking and real-world problems.
- SMPSO (Nebro et al. 2009) (*Speed Modulation Particle Swarm Optimization*) is a multi-objective particle swarm optimization algorithm. Its main feature is a bounding strategy of particles velocities to guide new particle's positions within the problem search space of variables. SMPSO uses the polynomial mutation as perturbation factor to avoid fast convergence. It uses an external archive to store those non-dominated solutions obtained during the search.
- GDE3 is the *Generalized Differential Evolution 3* (Kukkonen and Lampinen 2005), which follows the algorithmic design of NSGA-II, although using differential mutation and selection operators (of DE), instead of those typically used by NSGA-II. GDE3 uses an adapted version of crowding distance to generate better distributed sets of solutions.
- MOEA/D (Zhang and Li 2007) (*Multi-objective Evolutionary Algorithm Based on Decomposition*) has become the classical representative archive-less optimization



technique, where a multi-objective problem is decomposed into a set of single-objective subproblems that are optimized simultaneously. In particular, we have focused in this study on MOEA/D-DE (Li and Zhang 2009), which is a variant that applies differential evolution as variation operators. In addition, this algorithm applies a polynomial mutation to avoid premature convergence, hence to improve its exploration capability.

## 4 Experimentation

This section is devoted to explain the set of ligand–protein instances used for benchmarking the evaluated algorithms. The experimental methodology is also explained, as well as the set of parameters used to configure the algorithms.

### 4.1 Ligand–protein instances

In order to perform a through experimentation, we have selected an extensive benchmark of 73 complexes comprising both ligand and receptor flexibility. These complexes have been selected to cover an heterogeneous range of molecular samples, comprising from small-size ligand to large inhibitors. Our aim is to obtaining unbiased conclusions concerning which algorithm performs the best on a wide and heterogeneous set of complexes, hence to be suggested as optimizer to the decision-maker.

Other past studies carried out with these molecular instances (Morris et al. 2009) showed that those complexes involving small ligands usually represent more difficult problems, since the flexibility added to the receptor side chains (ARG8) increases the space of ligand interaction. The ligand–protein instances have been extracted from the PDB database,<sup>1</sup> so they have been lightly adapted for being used in our docking simulations for solution evaluation.

The benchmark of ligand–protein instances selected, with the PDB accession code, the X-ray crystal structures names, and the structure resolution (Å), is shown in Table 1. For all of them, the torsional degrees of freedom (flexibility) are 10 for ligands and 6 for macromolecules (receptors). These torsion movements are selected to allow the fewest number of atoms to move around the ligand core. This implies a number of 23 problem variables ( $n$ ) in solution vectors, consisting of: 3 for translation, 4 for rotation quaternion, and 16 for torsional degrees.

### 4.2 Methodology

A series of 31 independent executions have been launched for each combination of algorithm and molecular instance,

<sup>1</sup> In URL <http://www.rcsb.org/pdb/home/home.do>.

**Table 1** X-ray crystal structure coordinates taken from the PDB database and used in docking experiments. They consist of 73 complexes with accession codes from the PDB database. The range of resolution (Å) of each subgroup is also shown in the last column

Ligand type	PDB code	Resolution (Å)
Small size	1A9M, 1AAQ, 1B6L, 1B6M, 1BDL, 1BDQ, 1BDR, 1GNM, 1GNN, 1GNO, 1HBV, 1HEG, 1HIH, 1HPV, 1HSG, 1HTE, 1KZK, 1SBG, 1TCX, 1ZIH, 1ZIR, 3AID	1.09–2.8
Medium size	1B6J, 1B6K, 1B6P, 1D4K, 1D4L, 1HEF, 1HPS, 1HXW, 1IZH, 1IZI, 1JLD, 1K6C, 1K6P, 1K6T, 1K6V, 1MTR, 1MUI, 2BPX, 4HVP, 4PHV, 5HVP	1.75–2.8
Large size	1A94, 1HIV, 1HOS, 1HTG, 1HVI, 1HVJ, 1HVK, 1HVL, 1HVS, 1HWR, 1ODY, 1VIJ, 1VIK, 3TLH, 7HVP, 9HVP	1.8–2.8
Cyclic urea	1BV7, 1BV9, 1BWA, 1BWB, 1DMP, 1G35, 1HPO, 1MES, 1MEU, 1PRO, 1QBR, 1QBT, 1QBU, 7UPJ	1.8–2.5

summing up a total number of:  $31 \times 73 \times 4 = 9052$  runs in the experimentation phase. To carry out the experiments, we have used a Condor<sup>2</sup> middleware platform for distribute computing, which comprises close to 400 cores acting task scheduler, each task devoted to one independent run.

The resulting distributions have been computed in terms of the median and interquartile range (IQR). We have focused on two main quality indicators to assess performance: hypervolume ( $I_{HV}$ ) and unary additive Epsilon ( $I_{\epsilon+}$ ) (Deb 2001). The first indicator quantifies the convergence and diversity of the resulting Pareto fronts, whereas the second one ( $I_{\epsilon+}$ ) measures the convergence degree. For each molecular instance, a reference Pareto front has been computed with all the non-dominated solutions obtained in all the executions of all the algorithms. This is a useful practice when calculating the two indicators, since we are dealing with a real-world optimization problem for which the Pareto fronts are not known.

### 4.3 Parameter setup

For the common parameters of all the evaluated algorithms, we have chosen a population size of 150 individuals (particles in the case of SMP SO) and 1,500,000 function evaluations as stop condition. These values were chosen, as they are the default settings used by AutoDock in the reference study (Norgan et al. 2011).

<sup>2</sup> In URL: <http://research.cs.wisc.edu/htcondor/>.

In the case of specific parameters of each algorithm, a different parameter setup has been set by following recommendations of our previous studies (López-Camacho et al. 2015, 2016), as follows:

- NSGA-II: The distribution indexes for SBX crossover and mutation operators are  $\eta_c = 20$  and  $\eta_m = 20$ , respectively. For these two operators,  $p_c = 0.9$  is the crossover probability and  $p_m = 1/n$  the mutation probability, with  $n$  the number of decision variables. It applies binary tournament selection.
- SMPSO: In this algorithm,  $\varphi_1$  and  $\varphi_2$  are the acceleration coefficients, which are set to 1.5. The inertia weight is  $\omega = 0.9$ , and polynomial mutation is applied to a set of particles (1/6) in the swarm.
- GDE3 (variant *rand/1/bin*), the differential mutation and crossover parameters  $\mu$  and  $C_r$  take a value of 0.5.
- MOEA/D  $\mu$  is set to 0.5 and  $C_r$  is set to 1.0. The polynomial mutation is configured with the same settings applied in NSGA-II.

## 5 Results

In this section, the obtained results are presented and analyzed. First, we compare the performance of the algorithms. A second analysis with selected solutions is also carried to provide insights into biological validation.

### 5.1 Performance comparisons

In order to evaluate the performance of each algorithm, we have taken in consideration the selected quality indicators. The first indicator we applied is the  $I_{HV}$ , which is calculated by obtaining the sum of the contributed volume of each point of an approximated front with regard to a given reference point. Therefore, the higher the converge and diversity degrees, the better the  $I_{HV}$  value is. The medians and interquartile ranges of the obtained solutions in terms of  $I_{HV}$ , for the set of 73 docking instances and for the algorithms compared, are shown in Table 2. In accordance with these results, SMPSO obtains the best hypervolume for 63 out of the 73 instances, and the second best results in 5 instances. MOEA/D obtains the best results in 2 instances and the second best results in 31 instances. GDE3 obtained the best results in 8 instances and the second best results in 2 instances. It is observable that many cells have a  $I_{HV}$  equal to zero. This happens when all the obtained solutions are out of the limits of the reference point, which indicates we are facing a very hard optimization problem.

Table 3 shows the median and interquartile ranges computed in terms of  $I_{\epsilon+}$ . Similarly to the previous observations,

SMPSO obtains the best results in 60 of the 73 instances and the second best in 6 instances. MOEA/D obtains the best results in 6 instances and the second best in 61 instances. Finally, GDE3 obtains the best results in 7 instances although the second best in 6 instances.

Moreover, with the aim at providing the results obtained with statistical confidence, we have used a formal procedure by means of nonparametric statistical tests, since in some cases the distribution of results did not follow the conditions of normality and heteroscedasticity required (Sheskin 2007). In this way, the analyses and comparisons are focused on the distributions of each of the two metrics studied. To this end, we have used Friedman’s ranking and Holm’s post hoc multi-comparison tests to check which algorithms are statistically worse than the best ranked technique, which is used as control one.

In accordance with this procedure, Table 4 shows the results of these tests, where SMPSO has the best Friedman’s ranking value of  $I_{HV}$  (with a value of 1.20), followed by MOEA/D, GDE3, and NSGA-II. This means that SMPSO is set as the control technique for  $I_{HV}$  in the post hoc Holm test, and hence, it is compared with the remaining algorithms. In this regard, the adjusted  $P$ -values (Holm<sub>Ap</sub> in Table 4) resulting from these comparisons are for the last three algorithms (MOEA/D, GDE3 and NSGA-II), lower than the confidence level, set to  $\alpha = 0.05$  in this study. In the case of  $I_{\epsilon+}$ , SMPSO also shows the best ranking and is statistically better than MOEA/D, GDE3 and NSGA-II, for the benchmark of molecular instances used here.

From a graphical point of view, Figs. 2, 3, 4 and 5 plot the approximated reference Pareto fronts from all executions (in continuous line), with regard to the Pareto fronts of each compared algorithm, for all the molecular instances with small size, medium size large size and cyclic urea, respectively. In these graphs, each solution from the four studied algorithms is represented using different shapes and colors as shown in the plot legend. An interesting observation can be made with regards to the convergence region of fronts for each algorithm, where, for example, MOEA/D seems to converge to the region promoting solutions with low energies, while SMPSO is more focused on regions biased to the RMSD objective. This gives us insights into the different learning procedures induced by SMPSO and MOEA/D, which search in different regions of the problem landscape, therefore obtaining complementary solutions covering different parts in the Pareto fronts approximated.

### 5.2 Analysis on ligand binding sites and molecular interactions

In order to show the applicability of this multi-objective strategy, we have chosen a set of complexes that include HIV-protease therapeutic targets. The 1BDR and 1QBU cor-

**Table 2** Median and interquartile range of  $I(HV)$  for each algorithm and instance

	SMPSO	GDE3	MOEA/D	NSGA-II
1A9M	5.48e-01 (3.5e-02)	0.00e+00 (0.0e+00)	0.00e+00 (3.3e-01)	0.00e+00 (0.0e+00)
1AAQ	6.86e-01 (2.5e-02)	0.00e+00 (7.3e-01)	5.31e-01 (7.6e-01)	0.00e+00 (0.0e+00)
1B6L	6.86e-01 (8.3e-02)	0.00e+00 (0.0e+00)	0.00e+00 (6.4e-01)	0.00e+00 (0.0e+00)
1B6M	4.77e-01 (5.5e-02)	0.00e+00 (0.0e+00)	0.00e+00 (2.6e-01)	0.00e+00 (0.0e+00)
1BDL	5.87e-01 (4.8e-02)	7.81e-01 (9.3e-02)	5.52e-01 (3.8e-01)	5.21e-01 (6.2e-01)
1BDQ	6.87e-01 (2.9e-02)	0.00e+00 (0.0e+00)	2.44e-01 (6.1e-01)	0.00e+00 (0.0e+00)
1BDR	5.15e-01 (6.8e-02)	0.00e+00 (0.0e+00)	4.80e-01 (5.7e-01)	0.00e+00 (0.0e+00)
1GNM	8.13e-01 (3.2e-02)	0.00e+00 (0.0e+00)	0.00e+00 (7.6e-01)	0.00e+00 (0.0e+00)
1GNN	6.99e-01 (4.2e-02)	0.00e+00 (0.0e+00)	0.00e+00 (3.5e-01)	0.00e+00 (0.0e+00)
1GNO	7.78e-01 (3.4e-02)	0.00e+00 (0.0e+00)	0.00e+00 (0.0e+00)	0.00e+00 (0.0e+00)
1HBV	6.07e-01 (4.5e-02)	0.00e+00 (0.0e+00)	4.27e-01 (6.5e-01)	0.00e+00 (0.0e+00)
1HEG	5.18e-01 (5.1e-02)	6.63e-01 (4.2e-02)	5.76e-01 (2.3e-01)	9.12e-02 (4.2e-01)
1HIH	3.54e-01 (5.6e-02)	1.20e-01 (4.2e-01)	1.55e-01 (5.6e-01)	0.00e+00 (6.7e-02)
1HPV	6.57e-01 (6.0e-02)	0.00e+00 (0.0e+00)	4.28e-01 (6.6e-01)	0.00e+00 (0.0e+00)
1HSG	4.71e-01 (7.1e-02)	0.00e+00 (0.0e+00)	0.00e+00 (0.0e+00)	0.00e+00 (0.0e+00)
1HTE	5.55e-01 (4.6e-02)	9.59e-02 (5.7e-01)	4.02e-01 (5.4e-01)	0.00e+00 (0.0e+00)
1KZK	0.00e+00 (0.0e+00)	1.13e-01 (7.0e-02)	4.04e-01 (2.3e-01)	0.00e+00 (6.2e-02)
1SBG	5.48e-01 (8.6e-02)	0.00e+00 (0.0e+00)	1.56e-01 (4.9e-01)	0.00e+00 (0.0e+00)
1TCX	5.62e-01 (8.7e-02)	0.00e+00 (0.0e+00)	5.19e-01 (5.7e-01)	0.00e+00 (0.0e+00)
1ZIH	8.35e-01 (4.0e-02)	0.00e+00 (0.0e+00)	0.00e+00 (8.9e-01)	0.00e+00 (0.0e+00)
1ZIR	5.21e-01 (1.4e-01)	0.00e+00 (0.0e+00)	0.00e+00 (7.7e-01)	0.00e+00 (0.0e+00)
3AID	6.67e-01 (2.8e-02)	0.00e+00 (0.0e+00)	5.17e-01 (7.2e-01)	0.00e+00 (0.0e+00)
1B6J	8.12e-01 (2.3e-02)	0.00e+00 (0.0e+00)	0.00e+00 (6.0e-01)	0.00e+00 (0.0e+00)
1B6K	9.47e-01 (9.9e-03)	0.00e+00 (0.0e+00)	0.00e+00 (7.9e-01)	0.00e+00 (0.0e+00)
1B6P	4.53e-01 (1.4e-01)	0.00e+00 (0.0e+00)	0.00e+00 (3.2e-03)	0.00e+00 (0.0e+00)
1D4K	6.25e-01 (5.5e-02)	0.00e+00 (0.0e+00)	0.00e+00 (0.0e+00)	0.00e+00 (0.0e+00)
1D4L	7.15e-01 (5.5e-02)	0.00e+00 (0.0e+00)	0.00e+00 (0.0e+00)	0.00e+00 (0.0e+00)
1HEF	8.90e-01 (1.0e-02)	9.16e-01 (2.7e-02)	9.00e-01 (3.8e-02)	8.76e-01 (6.4e-01)
1HPS	7.54e-01 (4.7e-02)	0.00e+00 (0.0e+00)	0.00e+00 (4.7e-01)	0.00e+00 (0.0e+00)
1HXW	5.57e-01 (6.0e-02)	6.08e-01 (4.8e-01)	4.88e-01 (5.2e-01)	0.00e+00 (5.0e-01)
1IZH	7.60e-01 (3.1e-02)	0.00e+00 (0.0e+00)	3.59e-01 (7.0e-01)	0.00e+00 (0.0e+00)
1IZI	7.26e-01 (4.4e-02)	0.00e+00 (0.0e+00)	5.21e-01 (8.4e-01)	0.00e+00 (0.0e+00)
1JLD	5.99e-01 (3.4e-02)	0.00e+00 (6.8e-01)	5.52e-01 (6.3e-01)	0.00e+00 (0.0e+00)
1K6C	5.72e-01 (3.9e-02)	0.00e+00 (0.0e+00)	0.00e+00 (4.1e-01)	0.00e+00 (0.0e+00)
1K6P	3.73e-01 (1.2e-01)	0.00e+00 (0.0e+00)	0.00e+00 (0.0e+00)	0.00e+00 (0.0e+00)
1K6T	4.45e-01 (2.2e-02)	0.00e+00 (0.0e+00)	0.00e+00 (0.0e+00)	0.00e+00 (0.0e+00)
1K6V	5.90e-01 (6.4e-02)	0.00e+00 (0.0e+00)	0.00e+00 (0.0e+00)	0.00e+00 (0.0e+00)
1MTR	5.53e-01 (7.9e-02)	0.00e+00 (0.0e+00)	0.00e+00 (6.6e-01)	0.00e+00 (0.0e+00)
1MUI	7.57e-01 (4.6e-02)	0.00e+00 (0.0e+00)	0.00e+00 (7.1e-01)	0.00e+00 (0.0e+00)
2BPX	8.53e-01 (8.2e-02)	0.00e+00 (0.0e+00)	0.00e+00 (0.0e+00)	0.00e+00 (0.0e+00)
4HVP	6.27e-01 (1.8e-01)	0.00e+00 (0.0e+00)	4.13e-01 (7.5e-01)	0.00e+00 (0.0e+00)
4PHV	6.64e-01 (9.3e-02)	0.00e+00 (0.0e+00)	0.00e+00 (7.3e-01)	0.00e+00 (0.0e+00)
5HVP	6.53e-01 (4.1e-02)	0.00e+00 (0.0e+00)	0.00e+00 (0.0e+00)	0.00e+00 (0.0e+00)
1A94	6.42e-01 (4.8e-02)	0.00e+00 (0.0e+00)	0.00e+00 (5.4e-01)	0.00e+00 (0.0e+00)
1HIV	6.13e-01 (9.2e-02)	0.00e+00 (0.0e+00)	0.00e+00 (3.8e-01)	0.00e+00 (0.0e+00)
1HOS	4.40e-01 (5.5e-01)	0.00e+00 (4.4e-03)	9.52e-01 (9.2e-02)	0.00e+00 (1.5e-01)
1HTG	5.16e-01 (4.4e-02)	0.00e+00 (0.0e+00)	0.00e+00 (0.0e+00)	0.00e+00 (0.0e+00)
1HVI	6.30e-01 (5.0e-02)	0.00e+00 (0.0e+00)	2.93e-01 (7.5e-01)	0.00e+00 (0.0e+00)
1HVJ	5.17e-01 (3.1e-02)	0.00e+00 (0.0e+00)	2.59e-01 (5.2e-01)	0.00e+00 (0.0e+00)
1HVK	5.46e-01 (1.5e-02)	0.00e+00 (0.0e+00)	2.57e-01 (2.4e-01)	0.00e+00 (0.0e+00)
1HVL	4.99e-01 (1.9e-02)	0.00e+00 (0.0e+00)	2.13e-01 (4.8e-01)	0.00e+00 (0.0e+00)
1HVS	3.92e-01 (5.3e-02)	0.00e+00 (0.0e+00)	1.43e-01 (2.6e-01)	0.00e+00 (0.0e+00)
1HWR	8.51e-01 (1.3e-02)	8.69e-01 (9.0e-01)	7.23e-01 (3.9e-01)	0.00e+00 (7.5e-01)
1ODY	5.84e-01 (1.4e-01)	0.00e+00 (0.0e+00)	8.45e-02 (5.2e-01)	0.00e+00 (0.0e+00)
1VIJ	6.58e-01 (3.9e-02)	0.00e+00 (0.0e+00)	4.78e-01 (6.5e-01)	0.00e+00 (0.0e+00)
1VIK	5.60e-01 (3.3e-02)	0.00e+00 (0.0e+00)	0.00e+00 (5.1e-01)	0.00e+00 (0.0e+00)
3TLH	4.88e-01 (3.8e-02)	6.57e-01 (4.1e-02)	5.37e-01 (1.7e-01)	1.89e-01 (5.3e-01)
7HVP	6.45e-01 (1.8e-02)	6.79e-01 (2.0e-01)	6.73e-01 (1.9e-01)	4.80e-03 (6.2e-01)
9HVP	9.16e-01 (1.4e-02)	0.00e+00 (0.0e+00)	5.66e-01 (7.3e-01)	0.00e+00 (0.0e+00)
1BV7	5.00e-01 (3.4e-02)	0.00e+00 (0.0e+00)	0.00e+00 (6.4e-01)	0.00e+00 (0.0e+00)
1BV9	4.62e-01 (6.5e-02)	0.00e+00 (0.0e+00)	1.91e-01 (5.5e-01)	0.00e+00 (0.0e+00)
1BWA	5.42e-01 (7.4e-02)	0.00e+00 (0.0e+00)	0.00e+00 (6.9e-01)	0.00e+00 (0.0e+00)
1BWB	7.50e-01 (4.4e-02)	0.00e+00 (0.0e+00)	3.69e-01 (7.3e-01)	0.00e+00 (0.0e+00)
1DMP	8.30e-01 (3.2e-02)	0.00e+00 (0.0e+00)	0.00e+00 (5.5e-01)	0.00e+00 (0.0e+00)
1G35	5.16e-01 (4.6e-02)	0.00e+00 (0.0e+00)	0.00e+00 (0.0e+00)	0.00e+00 (0.0e+00)
1HPO	6.31e-01 (4.5e-02)	0.00e+00 (0.0e+00)	0.00e+00 (4.9e-01)	0.00e+00 (0.0e+00)
1MES	3.66e-01 (9.7e-02)	0.00e+00 (0.0e+00)	1.04e-01 (4.7e-01)	0.00e+00 (0.0e+00)
1MEU	6.19e-01 (7.7e-02)	5.13e-01 (3.3e-01)	4.73e-01 (4.5e-01)	0.00e+00 (0.0e+00)
1PRO	4.39e-01 (4.4e-02)	0.00e+00 (0.0e+00)	1.21e-02 (4.6e-01)	0.00e+00 (0.0e+00)
1QBR	7.20e-01 (7.8e-02)	0.00e+00 (0.0e+00)	5.48e-01 (7.5e-01)	0.00e+00 (0.0e+00)
1QBT	3.63e-01 (8.3e-02)	0.00e+00 (0.0e+00)	0.00e+00 (4.2e-01)	0.00e+00 (0.0e+00)
1QBU	9.18e-01 (1.1e-02)	9.18e-01 (9.4e-01)	8.52e-01 (9.1e-01)	0.00e+00 (0.0e+00)
7UPJ	7.23e-01 (8.1e-02)	0.00e+00 (0.0e+00)	0.00e+00 (4.2e-01)	0.00e+00 (0.0e+00)

Best and second best median results have dark and light gray backgrounds, respectively

**Table 3** Median and interquartile range of  $I(\epsilon+)$  for each algorithm and instance

	SMPSO	GDE3	MOEA/D	NSGA-II
1A9M	3.90e-01 (4.6e-02)	2.79e+00 (2.7e-01)	1.22e+00 (9.7e-01)	2.42e+00 (1.5e+00)
1AAQ	2.11e-01 (5.2e-02)	1.61e+00 (1.5e+00)	2.82e-01 (2.0e+00)	6.03e+00 (5.9e+00)
1B6L	2.26e-01 (1.8e-01)	1.89e+00 (1.7e+00)	1.77e+00 (2.1e+00)	3.35e+00 (2.2e+00)
1B6M	4.19e-01 (6.9e-02)	3.23e+00 (8.7e-02)	1.47e+00 (1.8e+00)	3.46e+00 (1.0e+00)
1BDL	2.94e-01 (8.3e-02)	7.48e-02 (8.8e-02)	2.74e-01 (3.4e-01)	3.47e-01 (2.1e+00)
1BDQ	2.59e-01 (5.0e-02)	7.83e+00 (5.8e+00)	6.26e-01 (3.8e+00)	8.70e+00 (6.3e+00)
1BDR	4.75e-01 (7.2e-02)	1.61e+00 (1.1e+00)	5.08e-01 (4.7e-01)	2.02e+00 (8.8e-01)
1GNM	1.06e-01 (3.3e-02)	6.21e+00 (1.2e+00)	4.16e+00 (5.7e+00)	6.70e+00 (2.1e+00)
1GNN	2.17e-01 (7.2e-02)	5.43e+00 (1.9e-01)	4.52e+00 (4.3e+00)	5.51e+00 (8.8e-01)
1GNO	1.09e-01 (4.4e-02)	6.12e+00 (3.5e-02)	3.85e+00 (2.5e+00)	6.71e+00 (1.1e+00)
1HBV	3.36e-01 (7.8e-02)	1.92e+00 (4.2e-02)	4.02e-01 (1.8e+00)	3.53e+00 (2.0e+00)
1HEG	3.99e-01 (6.1e-02)	1.63e-01 (1.1e-01)	2.67e-01 (2.6e-01)	7.42e-01 (5.7e-01)
1HIH	5.85e-01 (7.4e-02)	7.23e-01 (1.5e+00)	6.65e-01 (1.5e+00)	2.32e+00 (2.3e+00)
1HPV	2.87e-01 (7.6e-02)	1.72e+00 (9.4e-01)	3.51e-01 (1.6e+00)	2.67e+00 (1.8e+00)
1HSG	4.07e-01 (1.3e-01)	4.88e+00 (5.1e-01)	3.31e+00 (2.2e+00)	4.45e+00 (1.7e+00)
1HTE	3.40e-01 (6.8e-02)	7.85e-01 (5.7e-01)	3.85e-01 (9.0e-01)	1.60e+00 (7.3e-01)
1KZK	1.14e+00 (1.4e-01)	8.87e-01 (7.0e-02)	5.74e-01 (2.4e-01)	1.04e+00 (2.6e-01)
1SBG	3.14e-01 (1.6e-01)	8.31e+00 (6.7e-01)	6.13e-01 (3.1e+00)	7.62e+00 (4.6e+00)
1TCX	3.11e-01 (1.8e-01)	5.01e+00 (3.3e+00)	2.28e-01 (7.5e-01)	5.45e+00 (3.8e+00)
1ZIH	1.07e-01 (7.2e-02)	5.36e+00 (8.1e-01)	2.24e+00 (4.2e+00)	8.54e+00 (4.2e+00)
1ZIR	4.41e-01 (1.6e-01)	5.33e+00 (1.1e+00)	1.02e+00 (2.8e+00)	6.58e+00 (2.2e+00)
3AID	3.00e-01 (3.1e-02)	3.77e+00 (1.5e+00)	3.50e-01 (1.2e+00)	4.12e+00 (2.6e+00)
1B6J	1.13e-01 (3.6e-02)	5.86e+00 (1.9e+00)	2.77e+00 (4.9e+00)	6.26e+00 (1.7e+00)
1B6K	2.63e-02 (9.4e-03)	5.55e+00 (4.4e+00)	1.14e+00 (3.1e+00)	7.21e+00 (1.9e+00)
1B6P	2.85e-01 (1.8e-01)	5.83e+00 (5.0e-01)	3.19e+00 (4.9e+00)	6.19e+00 (1.2e+00)
1D4K	2.81e-01 (9.2e-02)	8.22e+00 (1.1e-01)	5.24e+00 (5.5e+00)	8.49e+00 (2.7e+00)
1D4L	1.98e-01 (1.1e-01)	3.69e+00 (2.2e+00)	3.62e+00 (1.2e+00)	5.58e+00 (3.0e+00)
1HEF	7.93e-02 (5.8e-03)	5.85e-02 (1.2e-02)	5.68e-02 (1.3e-02)	6.66e-02 (6.5e-01)
1HPS	1.41e-01 (4.9e-02)	7.27e+00 (1.2e-01)	2.57e+00 (2.6e+00)	7.32e+00 (7.7e-01)
1HXW	3.71e-01 (9.6e-02)	2.39e-01 (4.7e-01)	2.96e-01 (4.5e-01)	1.78e+00 (2.2e+00)
1IZH	1.05e-01 (3.2e-02)	6.32e+00 (1.1e-02)	4.92e-01 (3.5e+00)	6.22e+00 (3.1e+00)
1IZI	2.64e-01 (5.0e-02)	2.12e+00 (6.5e-02)	2.82e-01 (2.0e+00)	3.75e+00 (1.9e+00)
1JLD	3.02e-01 (7.5e-02)	2.03e+00 (5.4e+00)	2.43e-01 (9.9e-01)	5.58e+00 (5.1e+00)
1K6C	2.33e-01 (8.7e-02)	4.28e+00 (8.8e-01)	2.06e+00 (2.8e+00)	3.67e+00 (1.6e+00)
1K6P	4.92e-01 (1.5e-01)	1.85e+00 (1.7e+00)	2.73e+00 (2.1e+00)	3.57e+00 (1.7e+00)
1K6T	4.24e-01 (5.6e-02)	1.94e+00 (1.7e+00)	2.02e+00 (1.7e+00)	2.54e+00 (1.3e+00)
1K6V	2.30e-01 (7.9e-02)	4.19e+00 (8.9e-01)	2.86e+00 (1.5e+00)	4.37e+00 (8.9e-01)
1MTR	3.08e-01 (1.5e-01)	4.73e+00 (3.7e-01)	1.73e+00 (2.6e+00)	4.65e+00 (1.4e+00)
1MUI	1.90e-01 (8.8e-02)	7.41e+00 (2.5e+00)	3.02e+00 (3.5e+00)	7.62e+00 (3.3e+00)
2BPX	1.13e-01 (1.2e-01)	3.98e+00 (3.4e+00)	3.29e+00 (2.3e+00)	4.98e+00 (2.3e+00)
4HVP	3.02e-01 (2.0e-01)	4.15e+00 (8.1e-02)	3.02e-01 (2.1e+00)	4.31e+00 (6.4e-01)
4PHV	3.13e-01 (1.4e-01)	1.00e+01 (2.2e+00)	4.06e+00 (6.3e+00)	1.03e+01 (3.3e+00)
5HVP	1.92e-01 (7.7e-02)	5.58e+00 (6.5e-01)	2.06e+00 (2.2e-01)	6.67e+00 (9.4e-01)
1A94	2.68e-01 (7.1e-02)	7.53e+00 (6.8e-03)	3.88e+00 (7.0e+00)	9.19e+00 (2.4e+00)
1HIV	2.72e-01 (1.4e-01)	2.21e+00 (5.1e-01)	1.33e+00 (3.1e+00)	3.72e+00 (1.3e+00)
1HOS	5.59e-01 (5.7e-01)	1.01e+00 (1.9e-01)	4.70e-02 (9.3e-02)	1.16e+00 (6.0e-01)
1HTG	4.42e-01 (5.6e-02)	1.05e+01 (2.0e-02)	8.43e+00 (7.0e+00)	9.98e+00 (2.6e+00)
1HVI	2.18e-01 (8.4e-02)	7.41e+00 (1.5e-01)	5.08e-01 (1.0e+00)	8.25e+00 (1.2e+00)
1HVJ	2.74e-01 (9.6e-02)	5.71e+00 (3.4e-01)	5.58e-01 (1.4e+00)	5.90e+00 (9.2e-01)
1HVK	1.90e-01 (3.0e-02)	7.59e+00 (3.7e+00)	5.21e-01 (3.8e-01)	8.02e+00 (7.6e-01)
1HVL	2.69e-01 (3.6e-02)	8.55e+00 (7.5e-02)	6.06e-01 (1.2e+00)	9.14e+00 (9.9e-01)
1HVS	3.84e-01 (1.1e-01)	1.05e+01 (1.0e-01)	7.10e-01 (7.3e-01)	1.17e+01 (1.5e+00)
1HWR	1.03e-01 (3.3e-02)	6.23e-02 (1.5e+00)	1.59e-01 (3.5e-01)	1.68e+00 (5.5e+00)
1ODY	3.08e-01 (1.5e-01)	1.04e+01 (1.4e+00)	8.27e-01 (8.1e+00)	1.06e+01 (2.2e+00)
1VIJ	3.17e-01 (3.7e-02)	4.61e+00 (2.7e-02)	3.08e-01 (4.3e+00)	5.02e+00 (8.5e-01)
1VIK	3.40e-01 (5.1e-02)	6.74e+00 (1.4e-01)	4.79e+00 (6.2e+00)	6.90e+00 (1.0e+00)
3TLH	4.52e-01 (4.0e-02)	2.18e-01 (3.2e-02)	3.39e-01 (1.7e-01)	6.49e-01 (7.0e-01)
7HVP	3.13e-01 (2.5e-02)	1.89e-01 (1.7e-01)	1.97e-01 (1.8e-01)	9.33e-01 (4.2e+00)
9HVP	6.74e-02 (2.9e-02)	8.42e+00 (1.1e+00)	3.55e-01 (4.1e+00)	9.53e+00 (2.3e+00)
1BV7	4.80e-01 (3.6e-02)	4.56e+00 (8.1e-01)	4.17e+00 (4.7e+00)	5.89e+00 (1.6e+00)
1BV9	5.10e-01 (7.0e-02)	5.21e+00 (2.4e+00)	6.89e-01 (6.8e+00)	7.75e+00 (2.4e+00)
1BWA	4.32e-01 (8.9e-02)	3.49e+00 (1.3e+00)	3.22e+00 (4.4e+00)	5.24e+00 (1.1e+00)
1BWB	1.76e-01 (7.0e-02)	2.92e+00 (2.5e+00)	4.49e-01 (5.1e+00)	6.32e+00 (2.5e+00)
1DMP	1.03e-01 (2.8e-02)	4.41e+00 (4.9e+00)	2.57e+00 (4.0e+00)	7.94e+00 (4.4e+00)
1G35	2.13e-01 (1.1e-01)	5.05e+00 (8.2e-01)	4.63e+00 (3.7e+00)	5.25e+00 (1.2e+00)
1HPO	1.67e-01 (1.1e-01)	1.10e+00 (1.3e+00)	1.82e+00 (2.3e+00)	2.77e+00 (1.4e+00)
1MES	5.71e-01 (1.3e-01)	5.05e+00 (4.0e+00)	7.24e-01 (5.1e+00)	8.30e+00 (4.1e+00)
1MEU	3.38e-01 (8.5e-02)	3.83e-01 (3.1e-01)	3.81e-01 (4.2e-01)	6.33e+00 (3.0e+00)
1PRO	4.86e-01 (7.0e-02)	7.86e+00 (3.6e+00)	9.60e-01 (6.4e+00)	8.57e+00 (3.5e+00)
1QBR	2.69e-01 (9.7e-02)	3.57e+00 (8.4e-01)	3.50e-01 (3.3e+00)	4.87e+00 (1.3e+00)
1QBT	5.87e-01 (1.1e-01)	5.14e+00 (6.2e-01)	5.84e+00 (6.9e+00)	6.85e+00 (1.4e+00)
1QBU	7.07e-02 (1.5e-02)	5.15e-02 (1.9e+00)	1.11e-01 (1.9e+00)	4.08e+00 (2.2e+00)
7UPJ	1.66e-01 (9.5e-02)	2.43e+00 (6.6e-01)	1.49e+00 (1.9e+00)	2.37e+00 (5.6e-01)

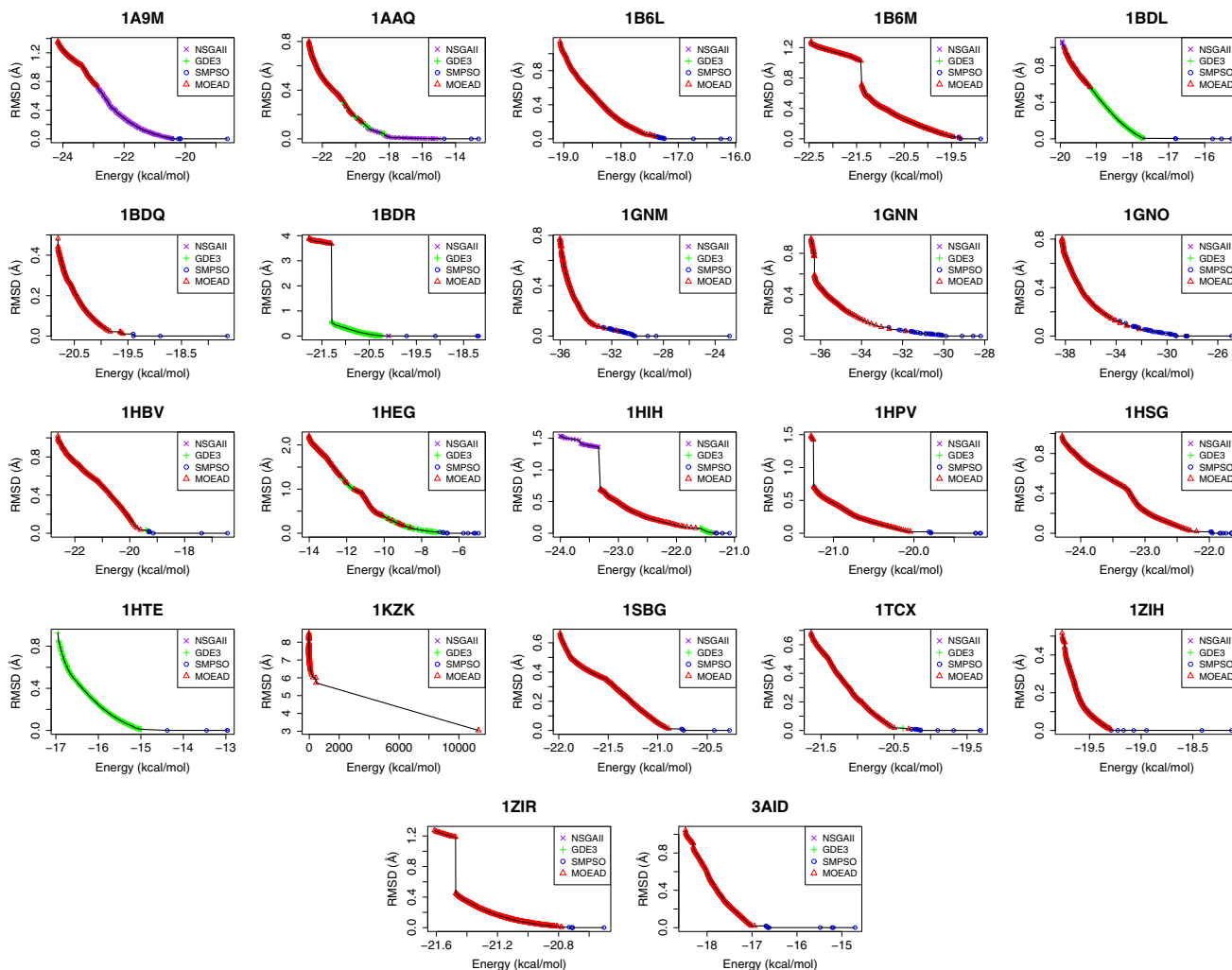
Best and second best median results have dark and light gray backgrounds, respectively



**Table 4** Friedman’s rankings with Holm’s Adjusted  $p$ -values (with confidence level  $\alpha = 0.05$ ) of the techniques compared for the test set of 73 docking instances

Hypervolume ( $I_{HV}$ )			Epsilon ( $I_{\epsilon+}$ )		
Algorithm	Fri <sub>Rank</sub>	Holm <sub>Ap</sub>	Algorithm	Fri <sub>Rank</sub>	Holm <sub>Ap</sub>
*SMPSO	1.20	–	*SMPSO	1.30	–
MOEA/D	2.52	7.54E–10	MOEA/D	1.99	1.34E–03
GDE3	2.93	1.01E–15	GDE3	2.83	1.76E–12
NSGA-II	3.33	6.28E–23	NSGA-II	3.86	1.81E–32

Symbol \* indicates the control algorithm, and column at right contains the overall ranking of positions with regard to indicators  $I_{HV}$  and  $I_{\epsilon+}$



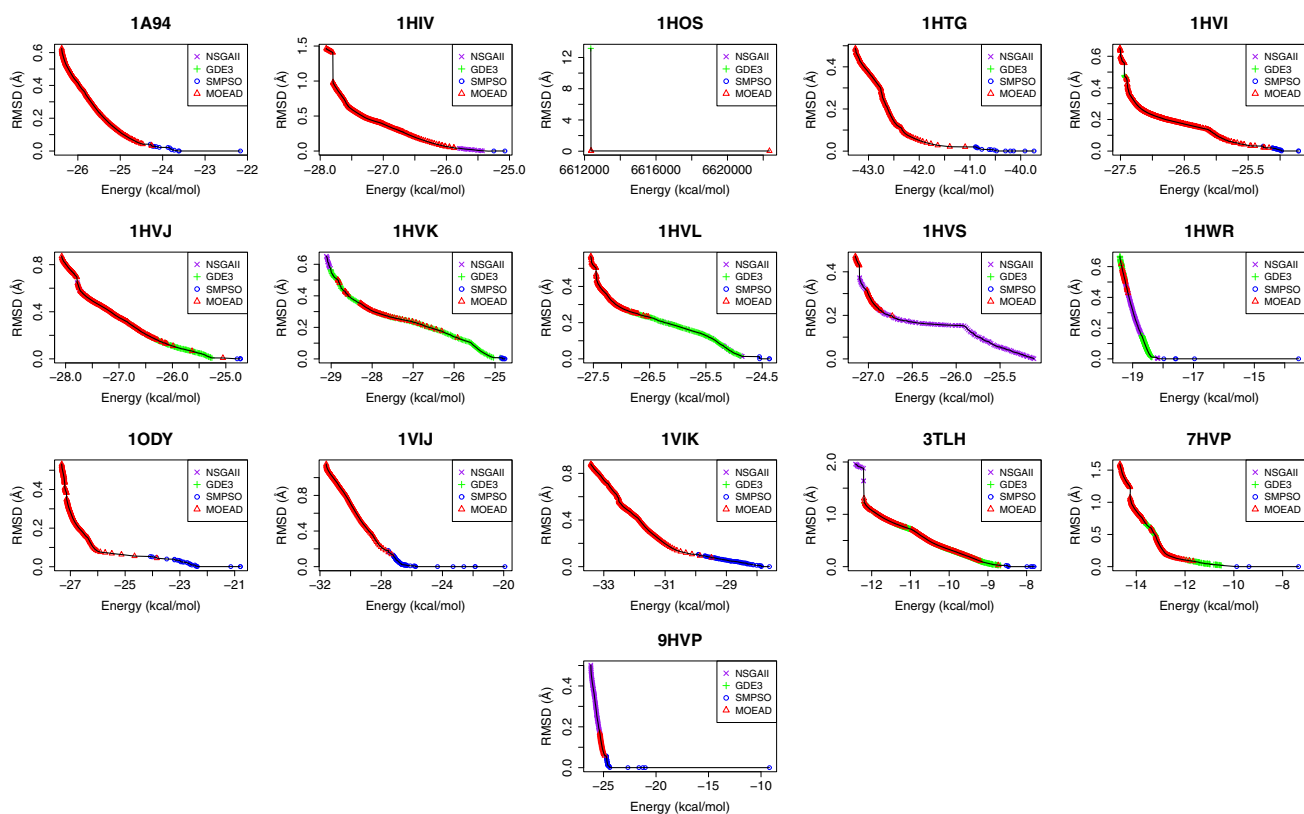
**Fig. 2** Reference fronts from all executions (in continuous line) with regard to the Pareto fronts of the studied algorithms with best hypervolume (in dotted lines) for small-size molecular instances: 1A9M, 1AAQ,

1B6L, 1B6M, 1BDL, 1BDQ, 1BDR, 1GNM, 1GNN, 1GNO, 1HBV, 1HEG, 1HIH, 1HPV, 1HSG, 1HTE, 1KZK, 1SBG, 1TCX, 1ZIH, 1ZIR, 3AID

respond to instances from the benchmark used in this study. The 3EKX and 3WSJ consist in interesting instances that have HIV-proteases as targets, and their inhibitors are being tested in resistant HIV-proteases.

The 1BDR corresponds to the complex HIV-1 protease with inhibitor SB203386, a tripeptide analogue inhibitor. The

chimeric HIV-protease crystal is formed by replacing amino acids in residues in the HIV type 1 with the corresponding residues from the HIV type 2 (31-37) and in the active site. It contains the mutations T31s, L33V, E34A, E35G, M36I and S37E. Some studies (Swairjo et al. 1998) have reported that these mutations can affect the dimensions of



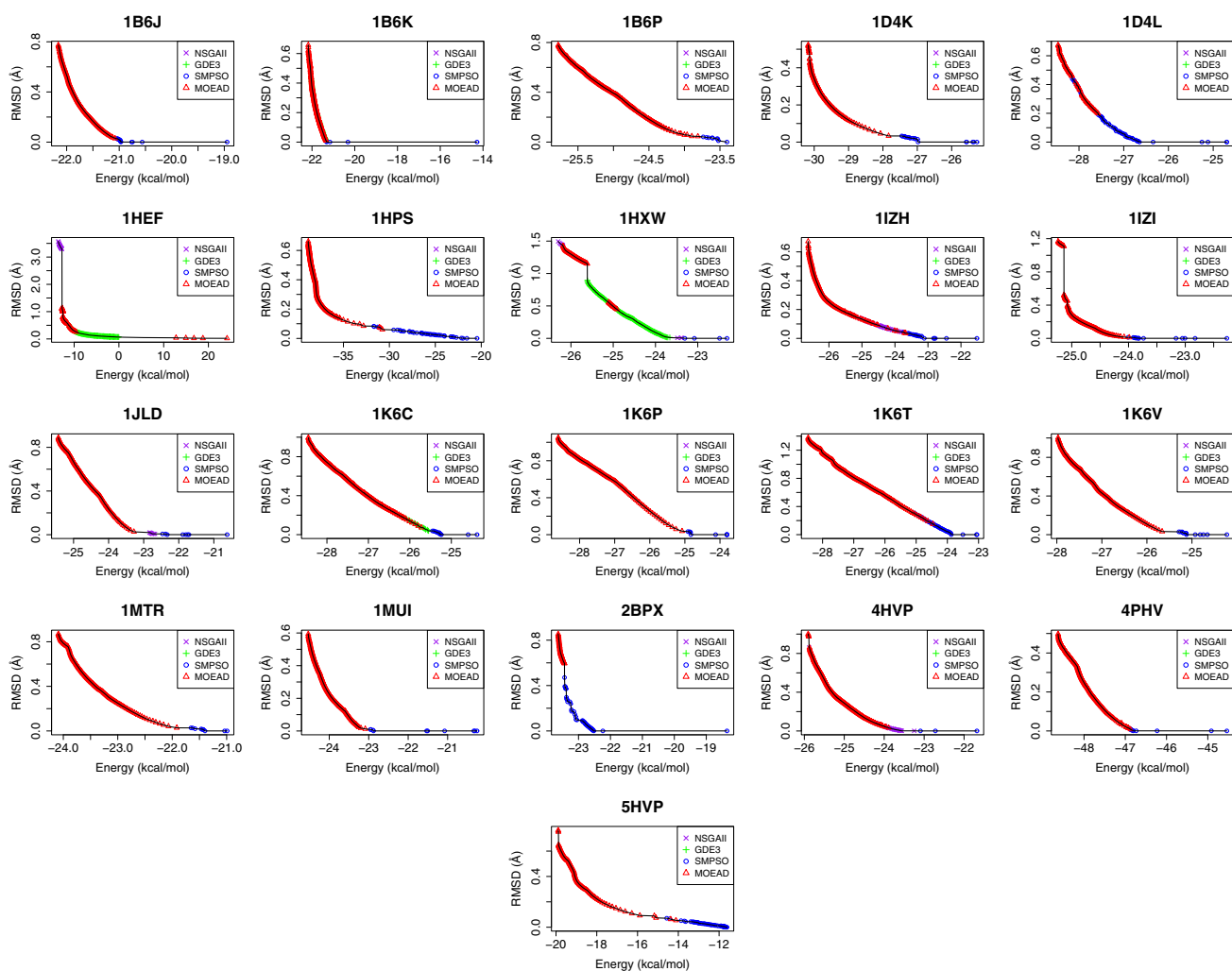
**Fig. 3** Reference fronts from all executions (in continuous line) with regard to the Pareto fronts of the studied algorithms with best hypervolume (in dotted lines) for large-size molecular instances: 1A94, 1HIV, 1HOS, 1HTG, 1HVI, 1HVJ, 1HVK, 1HVL, 1HVS, 1HWR, 1ODY, 1VIJ, 1VIK, 3TLH, 7HVP, 9HVP

the HIV-protease site active and hence the interaction HIV-protease-inhibitors. Due to the accumulation of mutations in this therapeutic target and the difference of response of HIV patients to the treatment with new HIV-protease inhibitors, the application of studies *in silico* with accurate molecular docking techniques to observe how these mutations affect the inhibitor affinity is necessary.

An 1BDR solution was chosen from the extreme of SMPSO non-dominated solutions front where the RMSD scores are lower and the  $E_{inter}$  values are higher (see 1BDR in Fig. 2). The RMSD and the  $E_{inter}$  values of the solution are equal to 0.00044 Å and -19.7 kcal/mol, respectively. In Fig. 6, Image A shows the tridimensional structure of HIV-protease and the inhibitor SB203386. The reference and the predicted ligands are represented with green and pink sticks. The two flexible ARG8 are also represented in both sides of the HIV-protease (sticks in pink). Both structures are overlapped and located in the HIV-protease active site, which is a shaped tunnel. Image B shows the molecular interactions between the ligand and receptor. The H-bonds are represented with green spheres. GLY27 (monomer B of HIV-protease) amide (-CO, carbonyl group) interacts with -NH of atoms N36 and N30. ASP25 side chain (monomer A of HIV-protease) forms a H-bond with -OH (atom O18). GLY27

(monomer A of HIV-protease) interacts with -NH (atom 8) through a H-bond. The results agree with those reported in Swairjo et al. (1998) and also in the ligand pose's view of the PDB entry 1BDR. This means that the predicted ligand's pose is very accurate, and therefore, the proposed strategy can be useful in those cases where HIV-proteases contain mutations that make them drug-resistant.

Another 1BDR solution was chosen from the other extreme of MOEA/D non-dominated solutions front where the RMSD scores are higher and the  $E_{inter}$  values are lower (see 1BDR in Fig. 2). The RMSD and the  $E_{inter}$  values of the solution are equal to 3.8 Å and -21.76 kcal/mol, respectively. This means that, in comparison with the other molecular docking solution presented above, the ligand-receptor is more stable in terms of energy and also a higher RMSD score shows that another interesting binding site can be found. Image A in Fig. 7 shows the reference and computed ligands represented in green and pink sticks, respectively. In this case, both structures are not overlapped. This is explained because the obtained RMSD score is higher than the previous result. Image B in Fig. 7 shows the molecular interactions between the inhibitor SB203386 and the HIV-protease. ASP29 amide (HIV-protease monomer A) forms a H-bond with -NH of atom O1 of the ligand.



**Fig. 4** Reference fronts from all executions (in continuous line) with regard to the Pareto fronts of the studied algorithms with best hyper-volume (in dotted lines) for medium-size molecular instances: 1B6J,

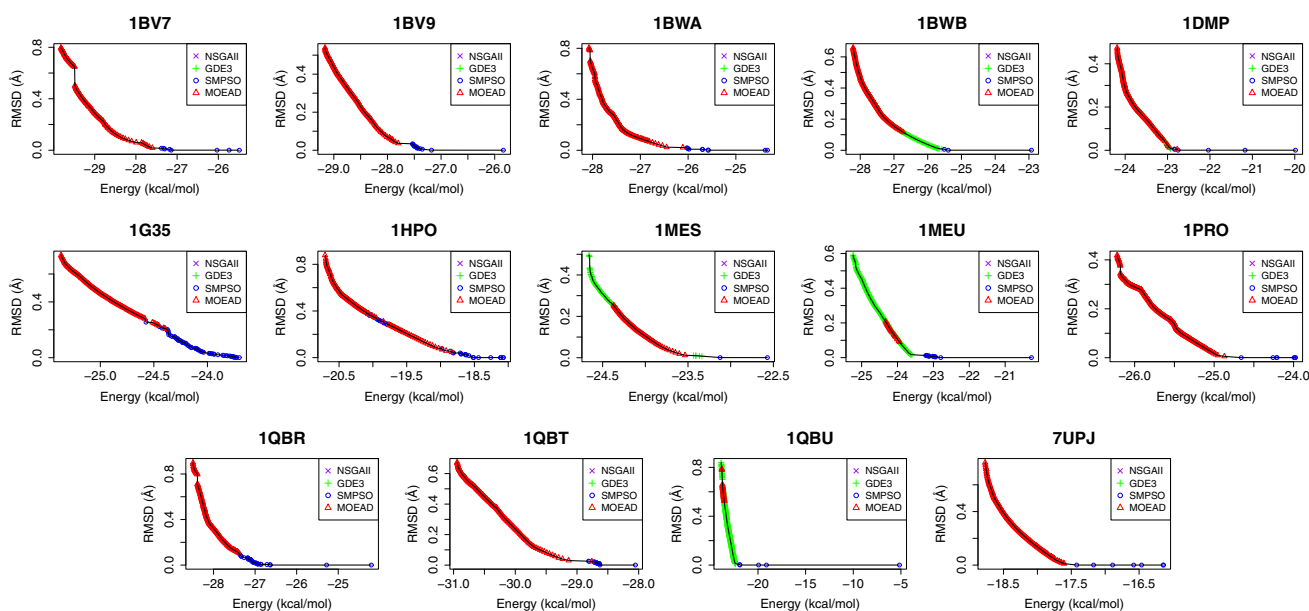
1B6K, 1B6P, 1D4K, 1D4L, 1HEF, 1HPS, 1HXW, 1IZH, 1IZI, 1JLD, 1K6C, 1K6P, 1K6T, 1K6V, 1MTR, 1MUI, 2BPX, 4HVP, 4PHV, 5HVP

Although the HIV-protease receptor’s catalytic site is the characteristic ASP25-THR26-GLY27 sequence (Brick and Wong 2003) and any of these amino acids are involved in the ligand–receptor interactions of the solution, the H-bond predicted is stable in terms of energy. Furthermore, the inhibitor SB203386 binds to the “flaps,” which is a pair of glycine,  $\beta$ -hairpins loops, one from each monomeric HIV-protease, that controls the access and the positioning of the substrate in the active site such as Ung et al. (2014) reports.

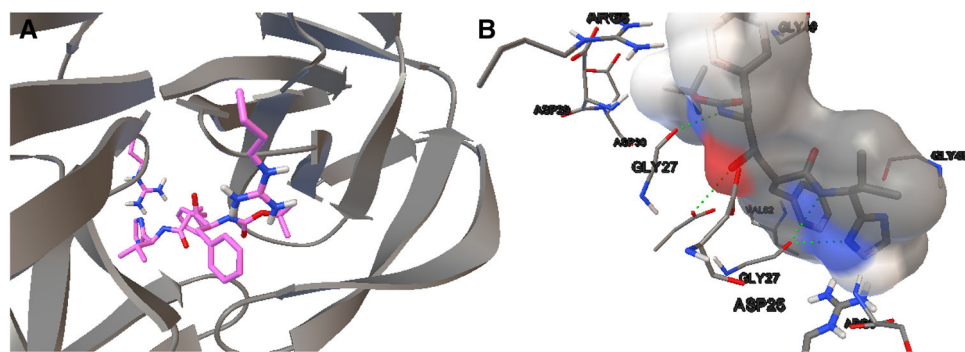
According to the obtained results, it is worth noting that the 1BDR corresponds to an instance that contains a small HIV-protease inhibitor. The results obtained by Morris et al. (2009) show that the LGA algorithm fails in finding a solution energetically stable in those docking problems where the inhibitors’ sizes are small (like 1BDR) and flexibility is applied to ARG8 side chain. The authors argued that these

results returned by LGA are due to the increase in docking space by the application of flexibility to the ARG8 side chains and also the small size of the ligand. Using the proposed multi-objective strategy, the solution returned by MOEA/D shows that the predicted docked ligand–receptor complex is energetically stable. In this case, MOEA/D was able to find a stable ligand’s conformation in terms of energy.

The 1QBU solution was chosen from the other extreme of SMPSO non-dominated solutions front where the RMSD scores are lower and the  $E_{inter}$  values are higher (see 1QBU in Fig. 5). This solution has an RMSD score of 0.000036 Å and an  $E_{inter}$  value of  $-19.11$  kcal/mol. This instance corresponds to the HIV-protease with mutation C95A and a cyclic urea amide. Image A in Fig. 8 shows the tridimensional structure of HIV-protease represented in sticks. The reference and predicted ligand structures are represented with



**Fig. 5** Reference fronts from all executions (in continuous line) with regard to the Pareto fronts of the studied algorithms with best hypervolume (in dotted lines) for cyclic-urea molecular instances: 1BV7, 1BV9, 1BWA, 1BWB, 1DMP, 1G35, 1HPO, 1MES, 1MEU, 1PRO, 1QBR, 1QBT, 1QBU, 7UPJ



**Fig. 6** Image A shows the tridimensional structure (its secondary structure is also shown) of the chimeric HIV-protease (1BDR). The reference and predicted ligands are represented with red and green sticks, respectively. Flexible ARG8 of two monomers (A and B) of HIV-protease is also represented with pink sticks. The reference and computed ligands

are overlapped given the low RMSD score of the solution. Image B shows the molecular interactions. The H-bonds are represented with green spheres. The involved and closer amino acids to the receptor-ligand interaction are also labeled

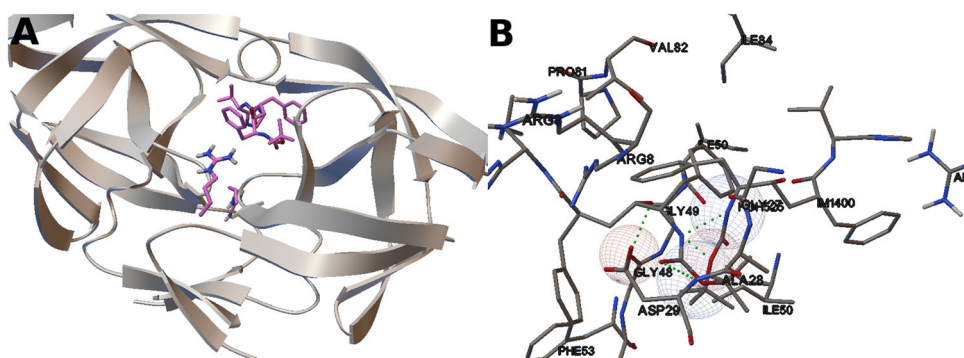
green and pink sticks, respectively. Both structures are overlapped. Figure B shows the molecular interactions between the cyclic urea amide and HIV-protease. The H-bonds are represented with green spheres. The GLY48 amide ( $-NH$ ) of HIV-protease monomer A forms a H-bond with the N27 atom of the ligand. The results agree with those obtained in Jadhav et al. (1997) and the ligand's pose view of 1QBU in the database PDB.

The instances 3EKX and 3WSJ correspond to HIV-proteases bound to novel inhibitors such as Nelfinavir and Indinavir. For these two instances, we performed several docking analyses to determine whether ARG8 side chains showed bad contacts to the inhibitors. ARG8 did not show

bad contacts to the 3EKX and 3WSJ ligands. Therefore, in these two cases, flexibility to the two ARG8 side chains was not applied. The instance 3EKX includes a HIV-protease with mutations Q7K and V64I, which are involved in the emergence of drug-resistant HIV virus. The instance 3WSJ includes the multi-drug-resistant HIV-protease with mutation L40I.

One 3EKX solution was chosen from the extreme of SMPSO non-dominated solutions where the RMSD scores are lower and the  $E_{inter}$  values are higher. The solution presents an RMSD score of 0.44 Å and an  $E_{inter}$  value of  $-19.32$  kcal/mol. Image A in Fig. 9 shows the molecular surface of HIV-protease. Monomers A and B are in green

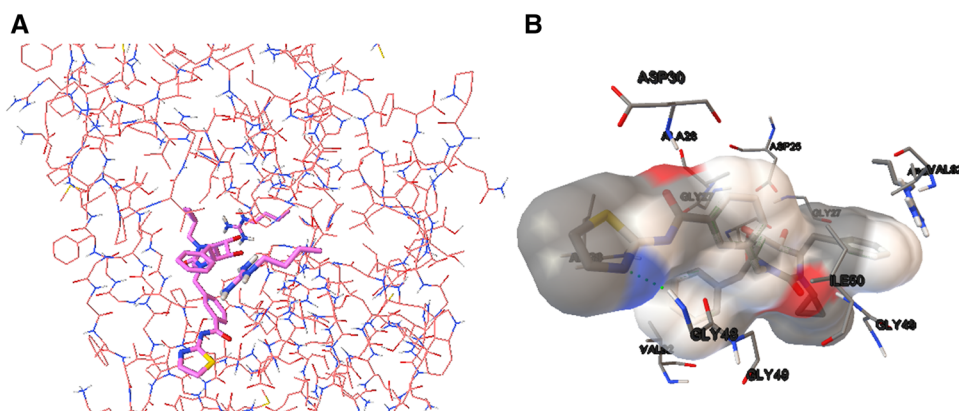




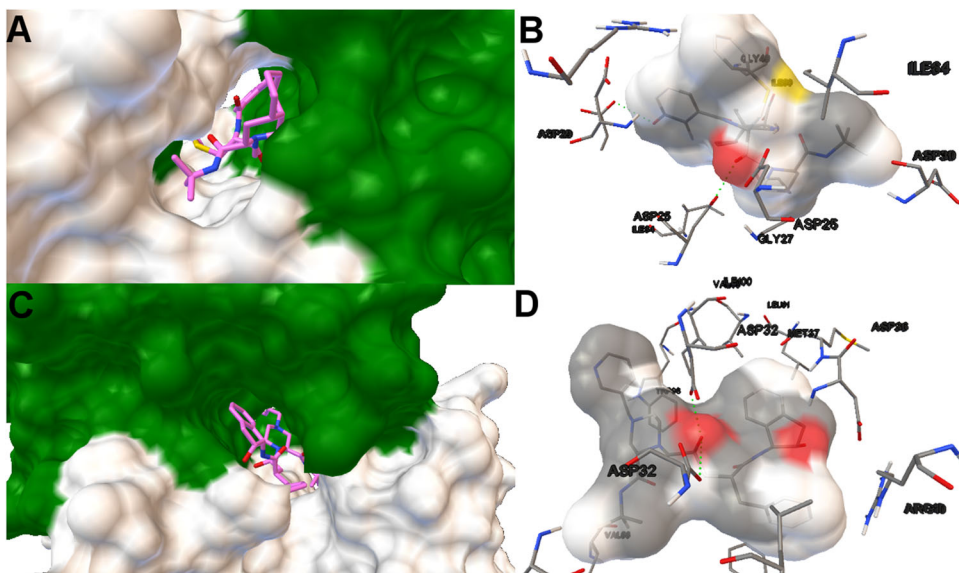
**Fig. 7** Image A shows the tridimensional structure (its secondary structure is also shown) of the chimeric HIV-protease (1BDR). The reference and predicted ligands are represented with green and pink sticks, respectively. The two flexible ARG8 of two HIV-protease monomers (A and B) is also represented with pink sticks. As Image A shows, the ligand

interaction is close to the “flaps” that play a role in the HIV-protease–substrate binding. In this case, the reference and computed ligands are not overlapped given the higher RMSD score of the solution. Image B shows the H-bonds are represented with green spheres. The involved and closer amino acids to the receptor interaction are also labeled

**Fig. 8** Image A shows HIV-protease represented with pink sticks (1QBU). The reference and computed ligands are represented with green and pink sticks, respectively. Both structures are overlapped given the accuracy of the results obtained for the RMSD score. Image B shows the molecular ligand–receptor interactions. The H-bonds are represented with green spheres. The involved and closer amino acids to the HIV-protease receptor are also labeled



**Fig. 9** Images A and C show the molecular surface of 3EKX and 3WSJ HIV-proteases. The reference and predicted ligands are represented with green and pink sticks, respectively. Monomers A and B are presented in green and white, respectively. In both cases, the reference and computed ligands are overlapped. Image B and D show the molecular interactions between ligand and receptor. The H-bonds are represented with green spheres. The involved and closer amino acids to the HIV-protease receptors are also labeled



and white, respectively. As shown, the inhibitor is bound to the multi-drug-resistant HIV-protease’s active site. The reference and the predicted ligands’ conformations are over-

lapped. Image B shows the ligand–receptor interactions. The predicted two H-bonds are represented with green spheres. ASP29 and ASP25 side chains of HIV-protease monomer A

form two H-bonds with –OH of atoms O38 and O21. In the 3EKX crystal structure, two H-bonds are formed between ASP25 of monomers A and B with atom O21. A third H-bond between ASP30 of monomer A and O38 is also formed (King et al. 2012). Although the prediction ligand's conformation shows an optimal conformation (below 2 Å), some of the predicted H-bonds interactions from the docking simulation are different than the observed in the crystallographic structure. However, a low value of  $E_{\text{inter}}$  indicates a stable ligand–receptor interaction.

For 3WSJ, a solution was chosen from the extreme of SMPSO non-dominated solutions where the RMSD scores are lower and the  $E_{\text{inter}}$  values are higher. The solution shows an RMSD score of 0.038 Å and an  $E_{\text{inter}}$  of – 15.94 kcal/mol. As the example above, Image C in Fig. 9 shows that the ligand is bound to the binding active site of HIV-protease. Image D shows the H-bonds formed between HIV-protease and Indinavir. The H-bonds are formed between ASP32 of monomers A and B and the atom O2 of Indinavir. The predicted molecular interactions are in accordance with Kuhnert et al. (2014) and the ligand's pose view of 3WSJ in the database PDB.

## 6 Conclusion

In this work, we perform a thorough experimentation to evaluate our multi-objective formulation for the molecular docking problem. This formulation involves a bi-objective optimization process, by minimizing the binding energy and the RMSD obtained in ligand–protein conformations. We extend here our previous research in López-Camacho et al. (2016), by assessing the performance of NSGA-II, SMPSO, GDE3 and MOEA/D, when tackling a large benchmark of 73 PDB structures (instances).

The following main conclusions are extracted from the analysis, as follows:

1. Our study confirms that using a multi-objective approach to optimized ligand–protein docking is able to approximate a set of solutions. Now we can provide practitioners with trade-off solutions in terms of RMSD and binding energy.
2. As experimented in our previous work (López-Camacho et al. 2016), SMPSO shows the best overall results in terms of hypervolume and epsilon indicators, in the scope of an extensive benchmark of molecular instances. Statistical procedure in Table 4 ranks MOEA/D as the second best performance technique, followed by GDE3 and NSGA-II.
3. A special behavior is observed for SMPSO, which approximates Pareto fronts of non-dominated solutions with low RMSD values, while MOEA/D converges to the

region focused on low energies. A combination of these two algorithms could lead to a new hybrid technique with prominent performance for ligand–protein docking.

4. In terms of biological validation, the results from the instances 1BDR and IQBU show that this multi-objective strategy allows practitioners to select a specific ligand's conformation from a range of solutions. This can be very interesting from a pharmacological point of view in those cases where the RMSD scores are higher and the  $E_{\text{inter}}$  is lower, specially, in those cases where the ligands bind to a regulation binding site of the macromolecule altering the enzyme–substrate interaction.

As future work, we are designing new algorithmic proposals by hybridizing search mechanisms included in SMPSO and MOEA/D algorithms with the aim of approximation solutions covering the full Pareto front. In this regard, the use of unbounded archives and restarting methods could be beneficial for such experiments in which other alternative active sites in compounds are required to be explored and discovered.

**Acknowledgements** This work has been partially funded by Grants TIN2017-86049-R (Spanish Ministry of Education and Science). Esteban López Camacho and María Jesús García Godoy are supported by Grants P11-TIC-7529 and P12-TIC-1519, respectively. José García Nieto is the recipient of a Post-Doctoral fellowship of “Captación de Talento para la Investigación” Plan Propio at University of Málaga.

## Compliance with ethical standards

**Conflicts of interest** The authors declare that they have no conflict of interest.

**Human and animal rights** This article does not contain any studies with human participants or animals performed by any of the authors.

## References

- Brick A, Wong CH (2003) Hiv-protease: mechanism and drug discovery. *Org Biomol Chem* 1:5–14
- Deb K (2001) Multi-objective optimization using evolutionary algorithms. Wiley, New York
- Deb K, Pratap A, Agarwal S, Meyarivan T (2002) A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE Trans Evol Comput* 6(2):182–197
- Gu J, Yang X, Kang L, Wu J, Wang X (2015) MoDock: a multi-objective strategy improves the accuracy for molecular docking. *Algorithms Mol Biol* 10:8. <https://doi.org/10.1186/s13015-015-0034-8>
- Jadhav PK, Ala P, Woerner FJ, Chang CH, Garber SS, Anton ED, Bacheler LT (1997) Cyclic urea amides: Hiv-1 protease inhibitors with low nanomolar potency against both wild type and protease inhibitor resistant mutants of hiv. *J Med Chem* 40(2):181–191
- Janson S, Merkle D, Middendorf M (2008) Molecular docking with multi-objective particle Swarm optimization. *Appl Soft Comput* 8(1):666–675

- King NM, Prabu-Jeyabalan M, Bandaranayake RM, Nalam MNL, Nali-vaika EA, Özen A, Haliloğlu T, Yılmaz NK, Schiffer CA (2012) Extreme entropy-enthalpy compensation in a drug-resistant variant of hiv-1 protease. *ACS Chem Biol* 7(9):1536–1546
- Kufareva I, Ruben A (2010) Methods of protein structure comparison. *Methods Mol Biol* 857:231–257
- Kuhnert M, Steuber H, Diederich WE (2014) Structural basis for htlv-1 protease inhibition by the hiv-1 protease inhibitor indinavir. *J Med Chem* 57(14):6266–6272
- Kukkonen S, Lampinen J (2005) GDE3: the third evolution step of generalized differential evolution. *IEEE Evolut Comput* 1:443–450. <https://doi.org/10.1109/CEC.2005.1554717>
- Leonhart PF, Spieler E, Ligabue-Braun R, Dorn M (2018) A biased random key genetic algorithm for the protein-ligand docking problem. *Soft Comput*. <https://doi.org/10.1007/s00500-018-3065-5>
- Li H, Zhang Q (2009) Multiobjective optimization problems with complicated pareto sets, MOEA/D and NSGA-II. *IEEE Trans Evolut Comput* 13(2):229–242
- López-Camacho E, García Godoy MJ, Nebro AJ, Aldana-Montes JF (2014) jmetalcpp: optimizing molecular docking problems with a c++ metaheuristic framework. *Bioinformatics* 30(3):437–438
- López-Camacho E, García Godoy MJ, García-Nieto J, Nebro AJ, Aldana-Montes JF (2015) Solving molecular flexible docking problems with metaheuristics: a comparative study. *Appl Soft Comput* 28:379–393. <https://doi.org/10.1016/j.asoc.2014.10.049>
- López-Camacho E, García-Godoy MJ, García-Nieto J, Nebro AJ, Aldana-Montes JF (2016) A new multi-objective approach for molecular docking based on RMSD and binding energy. Springer, Cham, pp 65–77
- Morris GM, Huey R, Lindstrom W, Sanner MF, Belew RK, Goodsell DS, Olson AJ (2009) AutoDock4 and AutoDockTools4: automated docking with selective receptor flexibility. *J Comput Chem* 30(16):2785–2791
- Nebro AJ, Durillo JJ, Garcia-Nieto J, Coello Coello CA, Luna F, Alba E (2009) SMPSO: A new PSO-based metaheuristic for multi-objective optimization. In: *IEEE symposium on computational intelligence in multi-criteria decision-making*, pp 66–73. <https://doi.org/10.1109/MCDM.2009.4938830>
- Norgan AP, Coffman PK, Kocher JPA, Katzmann DJ, Sosa CP (2011) Multilevel parallelization of autodock 4.2. *J Cheminf* 3(1):12
- Peh SCW, Hong JL (2016) Glsdock—drug design using guided local search. In: Gervasi O, Murgante B, Misra S, Rocha AMA, Torre CM, Taniar D, Apduhan BO, Stankova E, Wang S (eds) *Computational science and its applications—ICCSA 2016*. Springer, Cham, pp 11–21
- Ru X, Song C, Lin Z (2016) A genetic algorithm encoded with the structural information of amino acids and dipeptides for efficient conformational searches of oligopeptides. *J Comput Chem* 37(13):1214–1222. <https://doi.org/10.1002/jcc.24311>
- Sandoval-Perez A, Becerra D, Vanegas B, Restrepo-Montoya D, Niño F (2013) A multi-objective optimization energy approach to predict the ligand conformation in a docking process. In: *EuroGP*, pp 181–192
- Sheskin DJ (2007) *Handbook of parametric and nonparametric statistical procedures*. Chapman & Hall/CRC, London
- Swairjo M, Towler E, Debouck C, Abdel-Meguid S (1998) Structural role of the 30's loop in determining the ligand specificity of the human immunodeficiency virus protease. *Biochemistry* 4(10):928–36. <https://doi.org/10.1021/bi980784h>
- Townsend PD, Rodgers TL, Glover LC, Korhonen HJ, Richards SA, Colwell LJ, Pohl E, Wilson MR, Hodgson DRW, McLeish TCB, Cann MJ (2015) The role of protein-ligand contacts in allosteric regulation of the escherichia coli catabolite activator protein. *J Biol Chem* 290(36):22225–22235. <https://doi.org/10.1074/jbc.M115.669267>
- Trott O, Olson AJ (2010) Autodock vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J Comput Chem* 31(2):455–461
- Ung PMU, Dunbar JB, Gestwicki JE, Carlson HA (2014) An allosteric modulator of hiv-1 protease shows equipotent inhibition of wild-type and drug-resistant proteases. *J Med Chem* 57(15):6468–6478. <https://doi.org/10.1021/jm5008352>
- Zhang Q, Li H (2007) MOEA/D: a multiobjective evolutionary algorithm based on decomposition. *IEEE Trans Evol Comput* 11(6):712–731. <https://doi.org/10.1109/TEVC.2007.892759>

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.