# A Study of Archiving Strategies in Multi-objective PSO for Molecular Docking

José García-Nieto[1(✉)], Esteban López-Camacho[1(✉)], María Jesús García Godoy[1(✉)], Antonio J. Nebro[1(✉)], Juan J. Durillo[2(✉)], and José F. Aldana-Montes[1(✉)]

[1] Khaos Research Group, Department of Computer Sciences, University of Málaga, ETSI Informática, Campus de Teatinos, Málaga, Spain
{jnieto,esteban,mjgarciag,antonio,jfam}@lcc.uma.es
[2] Distributed and Parallel Systems Group, University of Innsbruck, Innsbruck, Austria
juan@dps.uibk.ac.at

**Abstract.** Molecular docking is a complex optimization problem aimed at predicting the position of a ligand molecule in the active site of a receptor with the lowest binding energy. This problem can be formulated as a bi-objective optimization problem by minimizing the binding energy and the Root Mean Square Deviation (RMSD) difference in the coordinates of ligands. In this context, the SMPSO multi-objective swarm-intelligence algorithm has shown a remarkable performance. SMPSO is characterized by having an external archive used to store the non-dominated solutions and also as the basis of the leader selection strategy. In this paper, we analyze several SMPSO variants based on different archiving strategies in the scope of a benchmark of molecular docking instances. Our study reveals that the SMPSOhv, which uses an hypervolume contribution based archive, shows the overall best performance.

**Keywords:** Multi-objective optimization · Particle Swarm Optimization · Molecular docking · Archiving strategies · Algorithm comparison

## 1 Introduction

Molecular docking is a complex optimization problem found in biology, which consists in predicting the position of a small molecule (ligand) in the active site of a receptor (macromolecule) that registers the minimum binding energy. Molecular docking is traditionally faced by means of metaheuristics [4,8] as a continuous optimization problem, since it requires to adjust position variables corresponding to coordinates of translation and torsion movements of molecules.

In the last decade, a number of studies have centered on the application of single- and multi-objective metaheuristics [4–6,8,15] to the molecular docking problem, showing successful results for a number of molecular compounds. In these previous works, different objective formulations were proposed that focused on energy scoring functions. Recently, a new multi-objective approach has been


M. Dorigo et al. (Eds.): ANTS 2016, LNCS 9882, pp. 40–52, 2016.
DOI: 10.1007/978-3-319-44427-7_4

proposed [9] in which two different objectives are to be minimized: the binding energy (the unbound and bound energy terms of the ligand/receptor complex) and the Root-Mean-Square-Deviation (RMSD) score. The latter objective leads the algorithms to guide the search when the co-crystallized ligand is known, which complements the traditional energy function.

Among these optimization techniques, a multi-objective swarm-intelligence approach, namely SMPSO [11], has emerged as one of the most prominent optimizers for molecular docking [4,9]. This technique performs a limitation mechanism of particle's velocity to avoid the movement of particles in search regions out of the problem ranges. SMPSO uses an external archive to store non-dominated solutions according to the crowding distance [2]. This archive is also used in the leader selection mechanism. Here, our motivation is to go one step beyond by evaluating, in the scope of a benchmark of molecular instances, new versions of SMPSO using different archiving strategies (hypervolume, cosine distance, and aggregation) and, consequently, different strategies for the selection of the leaders.

With this aim, we compare and analyze the proposed versions of SMPSO when solving 11 flexible ligand-receptor docking complexes taken from the AutoDock 4.2 benchmark [10]. This dataset includes flexible ligands with different sizes and flexible side-chains of HIV-protease receptors. The performance of the algorithms has been assessed by applying two main quality indicators intended to measure convergence and diversity of the computed Pareto front approximations.

The remainder of this article is organized as follows: Sect. 2 describes the molecular docking problem from a multi-objective formulation. Studied algorithms are described in Sect. 3. Section 4 reports the experimentation methodology and Sect. 5 analyzes the obtained results. Finally, Sect. 6 reports conclusions and future lines of research.

## 2 Molecular Docking

From a biological point of view, the main objective in the molecular docking problem is to find an optimized conformation between the ligand ($L$) and the receptor ($R$) that results in a minimum binding energy. The interaction between $L$ and $R$ can be described by an energy function calculated from three components representing degrees of freedom: (1) the translation of the ligand molecule, involving the three axis values $(x, y, z)$ in cartesian coordinate space; (2) the ligand orientation, modeled as a four variables quaternion including the angle slope ($\theta$); and (3) the flexibilities, represented by the free rotation of torsion (dihedral angles) of the ligand and sidechains of the receptor.

***- Solution Encoding:*** Each problem solution is then encoded by a real-value vector of $7 + n$ variables (as illustrated in Fig. 1), in which the first three values correspond to the ligand translation, the next four values correspond to the ligand and/or receptor orientation, and the remaining $n$ values are the ligand torsion dihedral angles.

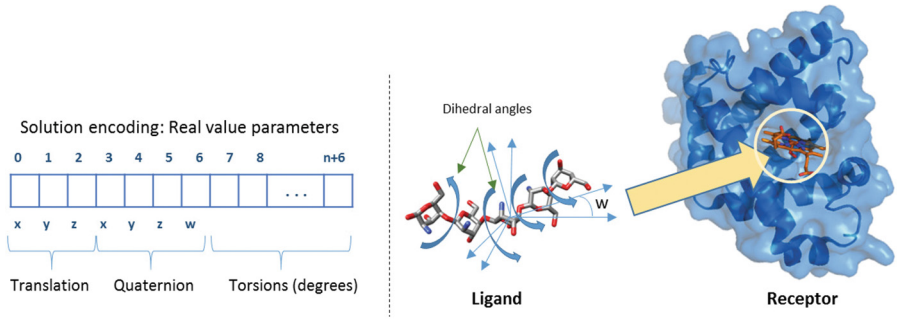


**Fig. 1.** Solution encoding. The first three values (translation) are the coordinates of the center of rotation of the ligand. The next four values correspond to quaternion and ($\theta$). The rest of the values hold the torsion angles in degrees.

The range of translation variables $(x, y, z)$ is $[0 \cdots 120]$, which has been delimited between the limits of the coordinates of a grid space previously set for each problem. Orientation (quaternion) and torsion variables are measured in radians and encoded in the range of $[-\pi, \pi]$.

***- Fitness Functions:*** the bi-objective formulation used here consists of: the $E_{binding}$ and the RMSD score. The $E_{binding}$ is the energy function as used in Autodock, which is calculated as follows:

$$E_{binding} = Q^{R-L}_{bound} + Q^{R-L}_{unbound} \tag{1}$$

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

$Q^{R-L}_{bound}$ and $Q^{R-L}_{bound}$ are the states of bound and unbound of the ligand-receptor complex, respectively. Each pair of energetic evaluation terms includes evaluations ($Q$) of dispersion/repulsion ($vdw$), hydrogen bonds ($hbond$), electrostatics ($elec$) and desolvation ($sol$). Weights $W_{vdw}$, $W_{hbond}$, $W_{conf}$, $W_{elec}$, and $W_{sol}$ of Eq. 2 are constants for Van der Waals, hydrogen bonds, torsional forces, electrostatic interactions and desolvation, respectively. An extended explanation of all these variables can be found in [10].

The RMSD is a measure of similarity between the real ligand position in the receptor and the computed position of the docking ligand. The lower RMSD score the better the solution is. A ligand-receptor docking solution with an RMSD score below 2Å is considered as a solution with high docking accuracy.

The RMSD score for two identical structures $a$ and $b$ is defined as follows:

$$RMSD_{ab} = max(RMSD'_{ab}, RMSD'_{ba}),\ with\ RMSD'_{ab} = \sqrt{\frac{1}{N}\sum_{i} \underset{\mathbf{j}}{min}\, r_2^{ij}} \tag{3}$$

The sum is over all $N$ heavy atoms in structure $a$, the minimum is over all atoms in structure $a$ with the same element type as atom $i$ in structure $b$.

## 3 Algorithms

In this section, we describe the SMPSO variants we are going to study. We start with the original algorithm and then we give details of the considered variants.

SMPSO is a Multi-Objective Particle Swarm Optimization (MOPSO) characterized by two features: a velocity constraint mechanism and an external bounded archive to store the non-dominated solutions found during the search [11]. A perturbation, implemented as a mutation operator, is also incorporated. Its pseudocode is included in Algorithm 1. The archive contains the current Pareto front approximation found by the algorithm, and it applies the crowding distance density estimator [2] to decide which particle to remove when it is full. The archive is also used in the leader strategy selection, consisting on binary tournament based on randomly selecting two solutions from it and taking the one with the highest crowding distance value (i.e., the one located in less crowded region of the front composed by all archived solutions). The local best position of a particle $i$ is obtained by applying a dominance test with the rest of particles in the swarm, in such a way that the current best particle (which initially is particle $i$) is updated when it is dominated by another one.

In [12], a study of different leader selection mechanisms on SMPSO was conducted. In that work, the most salient variant consisted in replacing the crowding distance by the degree of contribution of the solutions in the external archive according to the hypervolume indicator [18]. This way, the leader selection is based on a binary tournament that chooses the particle having the largest hypervolume contribution value. This version was named as SMPSO$_{hv}$ and it is the second selected algorithm to be compared in our study.

We introduce in this paper a new variant of SMPSO. The cosine similarity is a measure of similarity between two vectors that measures the cosine of the angle between them. This way, two vectors in the same direction have a cosine similarity value equals to zero, while two perpendicular vectors have a cosine similarity value of 1. As all the solutions in an external archive are non-dominated, we can define a density estimator by fixing a reference point and computing the cosine similarity among the vectors conformed by the archive solutions with regards to that reference point. The studied problem in this paper has two objectives, so we can sort the solutions in the archive by the first objective and compute, for each solution, a density value by summing up the cosine similarity of each point to their previous and next points; extreme points have a similarity distance equals to 0. This way, points having a largest cosine density value are in

**Algorithm 1.** Pseudocode of SMPSO

```
initializeSwarm()
initializeLeadersArchive()
generation = 0
while generation < maxGenerations do
    computeSpeed()
    updatePosition()
    mutation() // perturbation
    evaluation()
    updateLeadersArchive()
    updateParticlesMemory()
    generation ++
end while
returnLeadersArchive()
```

the most densely populated region. The resulting algorithm is called SMPSOC. An important issue in this technique is to select the proper reference point. Our previous study [9] indicated that the fronts have a convex shape, so we choose an approximation to the nadir point by taking the highest objective values of the solutions in the archive.

The fourth SMPSO version in our study, also presented in this paper for the first time, is an archive-less approach and it is called SMPSOD. To leave out the archive, we take the strategy of designing an aggregative version of SMPSO inspired by MOEA/D [17], where a multi-objective problem can be decomposed into a set of single-objective problems that can be optimized at the same time. This way, a set of evenly spread weight vectors $\lambda^1, \lambda^2, \ldots, \lambda^N$ are defined, being $N$ the size of the swarm. Then, each particle $i$ has associated the vector $\lambda^i$ and a neighborhood defined as a set of its several closest weight vectors in $\lambda^1, \lambda^2, \ldots, \lambda^N$. The scalarizing strategy follows the Tchebycheff scheme. The strategy for getting the local best of a particle $i$ is the same procedure as used by MOEA/D to update a neighborhood. The leader updating strategy consists in finding the best solution in the neighborhood by considering the scalar values of the particles taking into account their weight vectors.

SMPSO was inspired in the OMOPSO algorithm proposed by Reyes and Coello in [14], so we decided to also include it in our comparisons as a reference multi-objective particle swarm optimizer in the state of the art.

In summary, in our study we include OMOPSO and four SMPSO variants with different archiving strategies: crowding distance based (original SMPSO), hypervolume contribution based ($\text{SMPSO}_{hv}$), cosine distance based (SMPSOC), and archive-less (SMPSOD), being the last two ones proposed in this paper.

## 4 Experimentation

For the experiments, we have considered a benchmark of 11 molecular instances with receptor and ligand flexibility. These complexes are actually difficult dock-

**Table 1.** The accession codes, the X-ray crystal structure and resolution taken from PDB database are presented.

| PDB code | Protein-ligand complexes | Resolution (Å) |
|---|---|---|
| 1AJV | HIV-1 protease/AHA006 | 2.00 |
| 1AJX | HIV-1 protease/AHA001 | 2.00 |
| 1BV9 | HIV-1 protease/$\alpha$-D-glucose | 2.20 |
| 1D4K | HIV-1 protease/Macrocyclic peptidomimetic inhibitor 8 | 1.85 |
| 1G2K | HIV-1 protease/AHA047 | 1.95 |
| 1HIV | HIV-1 protease/U75875 | 2.00 |
| 1HPX | HIV-1 protease/KNI-272 | 2.00 |
| 1HTF | HIV-1 protease/GR126045 | 2.20 |
| 1HTG | HIV-1 protease/GR137615 | 2.00 |
| 1HVH | HIV-1 protease/Q8261 | 1.80 |
| 2UPJ | HIV-1 protease/U100313 | 3.00 |

ing problems containing a wide range of ligand sizes (from small to large inhibitors). The docking studies performed with these instances in [10] to test the energy function of AutoDock 4.2 demonstrated that the most difficult problems are those involving smaller ligands. This is due to the flexibility added to the receptor side-chains (ARG-8) that increases the space of ligand interactions. These instances have been taken from the PDB database[1].

Table 1 summarizes the set of problems selected showing the PDB accession code, the X-ray crystal structures names and the structure resolution (Å). For all instances, the torsional degrees of freedom for ligands and receptors are 10 and 6, respectively, selecting those torsions that allow the fewest number of atoms to move around the ligand core. Therefore, the solution vector contains: 3 variables for translation, 4 variables for rotation quaternion, and 16 variables for torsional degrees, summing up a total number ($n$) of 23 variables.

### 4.1 Methodology

The followed methodology consists in running each combination of algorithm and molecular instance 30 independent times. From these executions, we have calculated the median and interquartile range (IQR) as measures of central tendency and statistical dispersion, respectively. We have considered two quality indicators to assess the algorithm performance: Hypervolume ($I_{HV}$) [18] and Unary Additive Epsilon Indicator ($I_{\epsilon+}$) [19]. The former takes into account both convergence and diversity, whereas the later gives a measure of the convergence degree of the obtained Pareto front approximations. It is worth noting that we are dealing with a real-world optimization problem, and therefore the Pareto

[1] In URL: http://www.rcsb.org/pdb/home/home.do.

**Table 2.** Parameter settings.

| Common parameters | |
|---|---|
| Swarm size | 150 Particles |
| Iterations | 10,000 |
| SMPSO [3] & SMPSO$_{hv}$ & SMPSOD & SMPSOC | |
| Archive size | 100 |
| $C_1$, $C_2$ | 1.5 |
| $w$ | 0.9 |
| Mutation | polynomial mutation |
| Mutation probability | 1.66 |
| Mutation distribution index $\eta_m$ | 20 |
| Selection method | Rounds |
| OMOPSO [1] | |
| Archive size | 100 |
| $C_1$, $C_2$ | $rand(1.5, 2.0)$ |
| $w$ | $rand(0.1, 0.5)$ |
| Mutation | uniform + non-uniform + no mutation |
| Mutation probability | Each mutation is applied to 1/3 of the swarm |

fronts to calculate these two metrics are not known. To cope with this issue, we have generated a reference Pareto front for each instance by combining all the non-dominated solutions computed in all the executions of all the algorithms.

We have used the implementation of the five studied algorithms provided in the jMetalCpp framework [7], in combination with AutoDock 4.2 to evaluate the new generated solutions. To cope with the high computational requirements needed to carry out all the experiments, we have used the Condor[2] system, a middleware platform acting a distributed task scheduler of up to 400 cores.

The parameter settings are summarized in Table 2. We set a common subset of parameters which are the same for all the evaluated algorithms. The size of the swarm is 150 and the stopping condition is reached when 1,500,000 function evaluations are performed. These values were chosen as they are the default settings in AutoDock and they have been used in previous studies [13]. The archive size, when applicable, is set to 100.

All SMPSO versions use the polynomial mutation with distribution index $\eta_m = 20$, which is applied to one sixth of the particles in the swarm. The acceleration coefficients $C_1$ and $C_2$ are set to 1.5 and the inertia weight is $w = 0.9$. With these parameters setting, our approach has been to use common settings in order to make a fair comparison, keeping the rest of the parameters of SMPSO and OMOPSO according to the papers where they were originally described.

[2] In URL: http://research.cs.wisc.edu/htcondor/.

**Table 3.** Median and interquartile range of $I_{HV}$ for each algorithm and instance. Best and second best median results have dark and light gray backgrounds, respectively.

| | SMPSO | SMPSO$_{hv}$ | SMPSOD | SMPSOC | OMOPSO |
|---|---|---|---|---|---|
| 1AJV | $3.65e-01_{5.1e-02}$ | $4.33e-01_{4.0e-02}$ | $3.63e-01_{4.6e-02}$ | $3.55e-01_{4.8e-02}$ | $0.00e+00_{0.0e+00}$ |
| 1AJX | $4.31e-01_{2.5e-02}$ | $5.06e-01_{2.7e-02}$ | $4.74e-01_{3.7e-02}$ | $4.43e-01_{3.6e-02}$ | $0.00e+00_{0.0e+00}$ |
| 1D4K | $6.67e-01_{8.1e-02}$ | $8.48e-01_{1.1e-01}$ | $7.11e-01_{9.4e-02}$ | $7.35e-01_{9.2e-02}$ | $0.00e+00_{0.0e+00}$ |
| 1G2K | $3.84e-01_{5.3e-02}$ | $4.58e-01_{5.9e-02}$ | $3.82e-01_{4.1e-02}$ | $3.52e-01_{5.2e-02}$ | $0.00e+00_{0.0e+00}$ |
| 1HIV | $4.86e-01_{2.0e-01}$ | $6.74e-01_{2.9e-02}$ | $5.87e-01_{7.1e-02}$ | $4.66e-01_{2.4e-01}$ | $0.00e+00_{0.0e+00}$ |
| 1HPX | $3.60e-01_{1.8e-01}$ | $6.30e-01_{9.7e-02}$ | $4.77e-01_{1.0e-01}$ | $4.63e-01_{1.4e-01}$ | $0.00e+00_{0.0e+00}$ |
| 1HTF | $2.61e-01_{3.3e-01}$ | $4.17e-01_{2.4e-01}$ | $3.96e-01_{7.9e-02}$ | $2.77e-01_{3.1e-01}$ | $0.00e+00_{0.0e+00}$ |
| 1HTG | $8.33e-02_{1.3e-01}$ | $1.46e-01_{9.6e-02}$ | $1.03e-01_{8.2e-02}$ | $7.13e-02_{1.3e-01}$ | $0.00e+00_{0.0e+00}$ |
| 1HVH | $7.78e-01_{4.7e-02}$ | $8.69e-01_{9.3e-03}$ | $7.70e-01_{2.4e-02}$ | $7.85e-01_{2.9e-02}$ | $0.00e+00_{0.0e+00}$ |
| 1VB9 | $4.10e-01_{1.2e-01}$ | $5.09e-01_{5.6e-02}$ | $4.12e-01_{1.1e-01}$ | $4.38e-01_{9.1e-02}$ | $0.00e+00_{0.0e+00}$ |
| 2UPJ | $5.82e-01_{9.6e-02}$ | $6.96e-01_{5.1e-02}$ | $6.27e-01_{7.4e-02}$ | $6.20e-01_{6.8e-02}$ | $1.99e-01_{6.4e-01}$ |

**Table 4.** Median and interquartile range of $I_{\epsilon+}$ for each algorithm and instance. Best and second best median results have dark and light gray backgrounds, respectively.

| | SMPSO | SMPSO$_{hv}$ | SMPSOD | SMPSOC | OMOPSO |
|---|---|---|---|---|---|
| 1AJV | $5.12e-01_{1.0e-01}$ | $3.94e-01_{6.7e-02}$ | $5.35e-01_{1.0e-01}$ | $5.46e-01_{1.0e-01}$ | $5.31e+00_{2.0e+00}$ |
| 1AJX | $2.31e-01_{1.1e-01}$ | $1.32e-01_{4.3e-02}$ | $1.94e-01_{6.1e-02}$ | $2.57e-01_{9.4e-02}$ | $2.54e+00_{3.2e+00}$ |
| 1D4K | $2.06e-01_{8.6e-02}$ | $4.41e-02_{1.2e-01}$ | $1.54e-01_{7.3e-02}$ | $1.57e-01_{8.1e-02}$ | $8.81e+00_{4.1e+00}$ |
| 1G2K | $4.29e-01_{1.7e-01}$ | $2.81e-01_{2.0e-01}$ | $4.75e-01_{9.7e-02}$ | $5.15e-01_{1.1e-01}$ | $6.01e+00_{2.3e+00}$ |
| 1HIV | $3.95e-01_{3.6e-01}$ | $9.03e-02_{6.4e-02}$ | $2.66e-01_{1.2e-01}$ | $4.36e-01_{3.2e-01}$ | $4.91e+00_{1.1e+00}$ |
| 1HPX | $4.25e-01_{2.8e-01}$ | $1.30e-01_{9.2e-02}$ | $2.95e-01_{1.2e-01}$ | $3.17e-01_{1.7e-01}$ | $1.13e+01_{5.7e+00}$ |
| 1HTF | $6.60e-01_{1.5e+00}$ | $5.46e-01_{3.7e-01}$ | $5.64e-01_{1.1e-01}$ | $6.85e-01_{4.3e-01}$ | $1.49e+00_{6.2e-01}$ |
| 1HTG | $9.07e-01_{1.4e-01}$ | $8.35e-01_{9.3e-02}$ | $8.84e-01_{8.8e-02}$ | $9.23e-01_{2.1e-01}$ | $1.21e+01_{7.7e+00}$ |
| 1HVH | $1.46e-01_{4.4e-02}$ | $6.12e-02_{4.8e-03}$ | $1.47e-01_{3.9e-02}$ | $1.52e-01_{3.1e-02}$ | $5.11e+00_{2.4e+00}$ |
| 1VB9 | $3.34e-01_{2.2e-01}$ | $1.96e-01_{7.7e-02}$ | $3.44e-01_{1.8e-01}$ | $2.97e-01_{1.3e-01}$ | $9.31e+00_{1.6e+00}$ |
| 2UPJ | $2.86e-01_{7.9e-02}$ | $1.76e-01_{9.2e-02}$ | $2.25e-01_{1.4e-01}$ | $2.70e-01_{5.1e-02}$ | $7.74e-01_{4.0e+00}$ |

## 5 Results and Analysis

A first analysis in our experimentation corresponds to the of results in terms of the hypervolume indicator $I_{HV}$. This indicator computes the sum of the contributed volume of each point in the Pareto front (non-dominated solutions) with regards to a reference point. Therefore, the higher the convergence and diversity degree of a front, the higher (better) the resulting $I_{HV}$ value is.

Table 3 shows the median and interquartile range of the computed distributions (out of 30 independent runs) of $I_{HV}$, for the set of 11 docking instances and for the five compared algorithms. As we can observe, SMPSO$_{hv}$ obtains the best median values of $I_{HV}$ for all the molecular instances and SMPSOD is the second best performing technique. We have to mention that some results of OMOPSO have a $I_{HV}$ equal to zero. This happens when all the points of the produced fronts are dominated by the reference point. In contrast, all the SMPSO versions obtained $I_{HV}$ values higher than zero, which indicates that they are all able to produce solutions within the limits of the reference point.

In the case of $I_{\epsilon+}$, a similar observation can be extracted from Table 4. That is, SMPSO$_{hv}$ shows the best results for all instances, followed by SMPSOD and SMPSO (the lower $I_{\epsilon+}$ value, the better the result is). For this indicator, SMPSOC obtains a second best median value only for instance 1VB9.

These results are assessed with statistical confidence (in this study $p$-value = 0.05) by focusing on the entire distribution of each of the two studied metrics. In

**Table 5.** Average Friedman's rankings with Holm's Adjusted $p$-values (0.05) of compared algorithms for the test set of 11 docking instances. Symbol * indicates the control algorithm and column at right contains the overall ranking of positions with regards to $I_{HV}$ and $I_{\epsilon+}$.

| Hypervolume ($I_{HV}$) | | | Epsilon ($I_{\epsilon+}$) | | | Overall | |
|---|---|---|---|---|---|---|---|
| Algorithm | $Fri_{Rank}$ | $Holm_{Ap}$ | Algorithm | $Fri_{Rank}$ | $Holm_{Ap}$ | Algorithm | Rank |
| ***SMPSO$_{hv}$** | **1.01** | - | ***SMPSO$_{hv}$** | **1.00** | - | **SMPSO$_{hv}$** | **2** |
| SMPSOD | 2.54 | $2.18e-02$ | SMPSOD | 2.45 | $3.09e-02$ | SMPSOD | 4 |
| SMPSOC | 3.09 | $3.85e-03$ | SMPSO | 2.99 | $6.02e-03$ | SMPSOC | 5 |
| SMPSO | 3.36 | $1.36e-03$ | SMPSOC | 3.54 | $4.79e-04$ | SMPSO | 5 |
| OMOPSO | 4.99 | $1.19e-08$ | OMOPSO | 4.98 | $1.19e-08$ | OMOPSO | 10 |

concrete, we have applied Friedman's ranking and Holm's post-hoc multicompare tests [16] to know which algorithms are statistically worse than the control one (i.e., the one ranking the best).

This way, as shown in Table 5, SMPSO$_{hv}$ is the best ranked variant according to Friedman test for the two indicators ($I_{HV}$ and $I_{\epsilon+}$), and it is followed by SMPSOD. Therefore, SMPSO$_{hv}$ is established as the control algorithm in the post-hoc Holm tests, which is compared with the rest of algorithms. The adjusted $p$-values ($Holm_{Ap}$ in Table 5) resulting from these comparisons are, for the remaining variants (SMPSO, SMPSOD, SMPSO, and OMOPSO), lower than the confidence level (0.05), meaning that SMPSO$_{hv}$ is statistically better than these algorithms. SMPSO and SMPSOC obtained similar overall performances, although showing SMPSOC better ranking than SMPSO in terms of $I_{HV}$.

Figure 2 shows the boxplots of the distributions of results concerning the $I_{HV}$ values, for each compared algorithm and molecular instance. In this figure, we can check that SMPSO$_{hv}$ variant obtains the best distributions for all the instances. An interesting observation can be made regarding OMOPSO, whose distributions denote poor results, although it produces outlier solutions with the best indicator values for some instances: 1AJV, 1AJX, 1HPX, and 1HTF. These outliers lead OMOPSO to contribute with many solutions to the reference Pareto fronts. An example of this can be observed in Fig. 3, where the fronts with best $I_{HV}$ values of all compared algorithms are plotted for instance 1AJX. However, the overall results (in boxplots) of OMOPSO indicate that it behaves irregular (non-robust) for all the molecular instances.

Following with Fig. 3, another interesting observation lies in the ability of SMPSOD to obtain non-dominated solutions in the region of the reference Pareto front with low energy and high RMSD values (top-left in plots of Fig. 3). In contrast with the other compared algorithms, SMPSOD is able to properly cover this area, as well as other areas with low RMSD. Therefore, as suggested in our previous study [9], a hybrid implementation of SMPSO using an aggregative (archive-less) strategy as done in MOEA/D, would cover the reference front with non-dominated solutions in the two objective ends. This assumption is now tested with SMPSOD in this study.

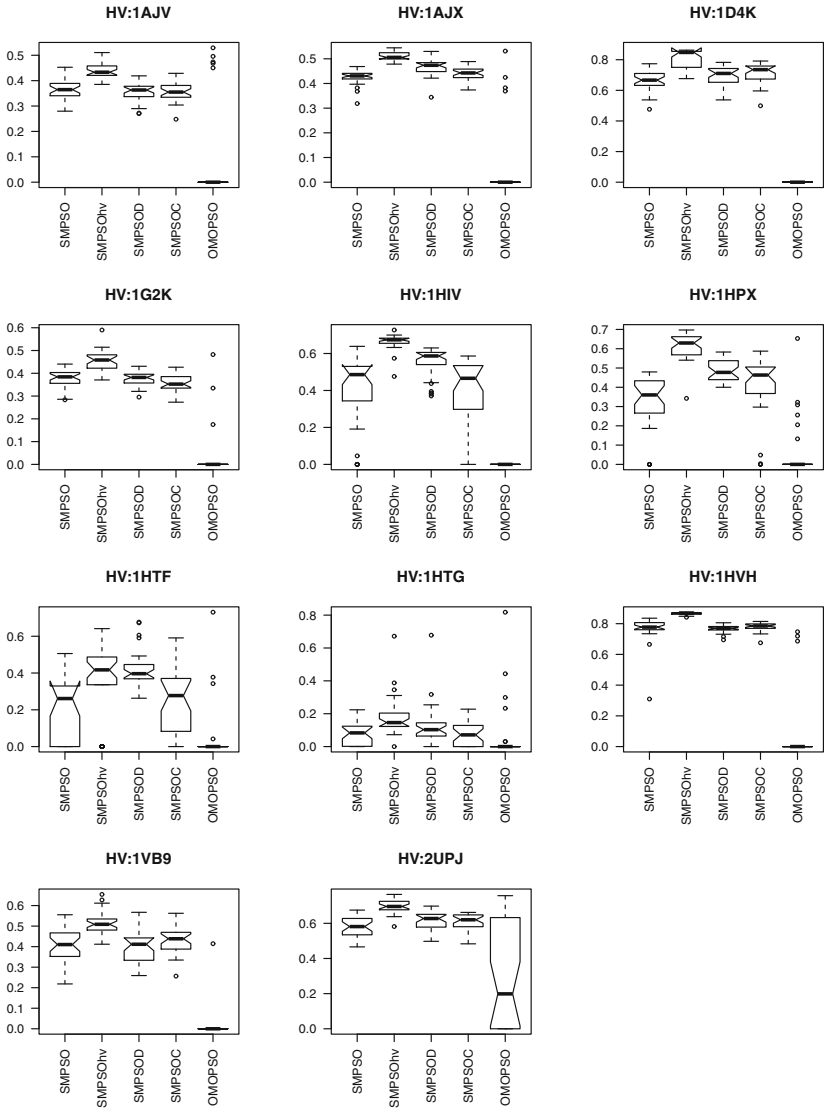


**Fig. 2.** Resulting boxplots of each compared algorithm and instance for $I_{HV}$

In summary, SMPSO$_{hv}$ shows the overall best behaviour followed by SMPSOD. Intuitively, the former obtains the best $I_{HV}$ as it performs a leader selection method of non-dominated solutions (from the external archive) with largest hypervolume contributions. That is, the particles in the swarm are guided by leaders with large hypervolume contributions, which would enable SMPSO$_{hv}$ to obtain, not only high values of $I_{HV}$, but also accurate results in terms of $I_{\epsilon+}$.

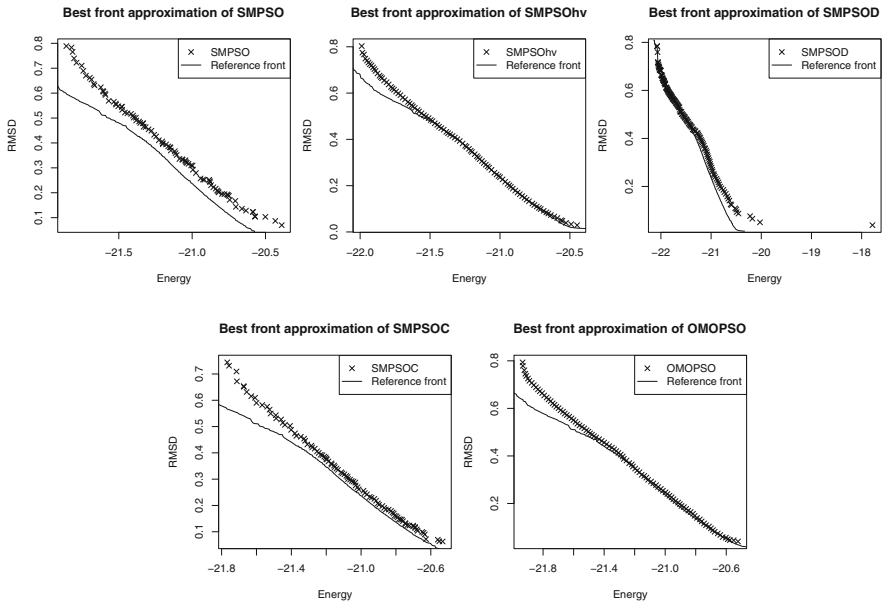


**Fig. 3.** Fronts with best $I_{HV}$ values on problem 1AJX.

## 6 Conclusions

In this paper, we analyze new variants of SMPSO, a multi-objective swarm optimization technique, based on different archiving strategies in the scope of a benchmarking set of molecular docking instances. The problem is formulated as a bi-objective optimization problem, by minimizing the binding energy and the Root Mean Square Deviation (RMSD) difference in the coordinates of ligands.

Our study reveals that SMPSO$_{hv}$ shows the overall best performance, followed by SMPSOD, SMPSOC, and SMPSO. The former variant obtains the best $\boldsymbol{I}_{HV}$ as it performs a leader selection method of those non-dominated solutions (from the external archive) having the largest hypervolume contributions, which seems to be responsible of the best diversity and convergence values in this comparison. OMOPSO shows moderate results, although reaching outperforming outlier solutions for some instances: 1AJV, 1AJX, 1HPX, and 1HTF. Interestingly, SMPSOD variant is able to cover the reference front with non-dominated solutions in the two objective extremes, i.e., with low energy and RMDS values. In this regard, as suggested in our previous study [9], a hybrid implementation of SMPSO using an aggregative (archive-less) strategy as done in MOEA/D, would cover the reference front with non-dominated solutions in the two objective ends. This assumption is now tested with SMPSOD in this study. Ideally, this SMPSO variant would contribute to discover other different (unknown) active sites in the receptor molecule with low energy, but far from the known active site (that is, with low RMSD).

This last open a future line of research for us on the selection and study of interesting solutions to be evaluated from a biological point of view. In addition, a natural extension of this work would be to test these conclusions on a greater number of molecular instances and using other quality indicators.

**Acknowledgments.** This work is partially funded by Grants TIN2011-25840 (Ministerio de Ciencia e Innovación) and P11-TIC-7529 and P12-TIC-1519 (Plan Andaluz I+D+I). This article is based upon work from COST Action CA15140, supported by COST (European Cooperation in Science and Technology).

## References

1. Coello, C.A., Toscano, G., Lechuga, M.S.: Handling Multiple objectives with Particle Swarm Optimization. IEEE Trans. Evol. Comp. **8**(3), 3 (2004)
2. Deb, K., Pratap, A., Agarwal, S., Meyarivan, T.: A fast and elitist multiobjective genetic algorithm: NSGA-II. IEEE Trans. Evol. Comput. **6**(2), 182–197 (2002)
3. Durillo, J.J., García-Nieto, J., Nebro, A.J., Coello, C.A.C., Luna, F., Alba, E.: Multi-objective particle swarm optimizers: an experimental comparison. In: Ehrgott, M., Fonseca, C.M., Gandibleux, X., Hao, J.-K., Sevaux, M. (eds.) EMO 2009. LNCS, vol. 5467, pp. 495–509. Springer, Heidelberg (2009)
4. García-Godoy, M.J., López-Camacho, E., García Nieto, J., Nebro, A.J., Aldana-Montes, J.F.: Solving molecular docking problems with multi-objective metaheuristics. Molecules **20**(6), 10154–10183 (2015)
5. Gu, J., Yang, X., Kang, L., Wu, J., Wang, X.: MoDock: a multi-objective strategy improves the accuracy for molecular docking. Algs. Mol. Bio. **10**, 8 (2015)
6. Janson, S., Merkle, D., Middendorf, M.: Molecular docking with multi-objective particle swarm optimization. Appl. Soft Comput. **8**(1), 666–675 (2008)
7. López-Camacho, E., García-Godoy, M.J., Nebro, A.J., Aldana-Montes, J.F.: jMetalCpp: optimizing molecular docking problems with a C++ metaheuristic framework. Bioinformatics **30**(3), 437–438 (2014)
8. López-Camacho, E., García-Godoy, M.J., García-Nieto, J., Nebro, A.J., Aldana-Montes, J.F.: Solving molecular flexible docking problems with metaheuristics: a comparative study. Appl. Soft Comput. **28**, 379–393 (2015)
9. López-Camacho, E., García-Godoy, M.J., García-Nieto, J., Nebro, A.J., Aldana-Montes, J.F.: A new multi-objective approach for molecular docking based on RMSD and binding energy. In: 3rd International Conference on Algorithm for Computational Biology (2016, in-Press)
10. Morris, G.M., Huey, R., Lindstrom, W., Sanner, M.F., Belew, R.K., Goodsell, D.S., Olson, A.J.: AutoDock4 and AutoDockTools4: automated docking with selective receptor flexibility. J. Comput. Chem. **30**(16), 2785–2791 (2009)
11. Nebro, A., Durillo, J., Garcia-Nieto, J., Coello Coello, C.A., Luna, F., Alba, E.: SMPSO: a new PSO-based metaheuristic for multi-objective optimization. In: IEEE Symposium on Computational Intelligence in Multi-criteria Decision-Making, pp. 66–73 (2009)
12. Nebro, A., Durillo, J., Coello Coello, C.A.: Analysis of leader selection strategies in a MOPSO. In: Proceedings of IEEE Congress on Evolutionary Computation (CEC), pp. 3153–3160, June 2013
13. Norgan, A.P., Coffman, P.K., Kocher, J.P.A., Katzmann, D.J., Sosa, C.P.: Multilevel parallelization of AutoDock 4.2. J. Cheminform. **3**(1), 12 (2011)

14. Sierra, M.R., Coello Coello, C.A.: Improving PSO-based multi-objective optimization using crowding, mutation and $\epsilon$-dominance. In: Coello Coello, C.A., Hernández Aguirre, A., Zitzler, E. (eds.) EMO 2005. LNCS, vol. 3410, pp. 505–519. Springer, Heidelberg (2005)
15. Sandoval-Perez, A., Becerra, D., Vanegas, D., Restrepo-Montoya, D., Nino, F.: A multi-objective optimization energy approach to predict the ligand conformation in a docking process. In: Krawiec, K., Moraglio, A., Hu, T., Etaner-Uyar, A.Ş., Hu, B. (eds.) EuroGP 2013. LNCS, vol. 7831, pp. 181–192. Springer, Heidelberg (2013)
16. Sheskin, D.J.: Handbook of Parametric and Nonparametric Statistical Procedures. Chapman & Hall/CRC, Boca Raton (2007)
17. Zhang, Q., Li, H.: MOEA/D: a multiobjective evolutionary algorithm based on decomposition. IEEE Trans. Evol. Comp. **11**(6), 712–731 (2007)
18. Zitzler, E., Thiele, L.: Multiobjective evolutionary algorithms: a comparative case study and the strength pareto approach. IEEE Trans. Evol. Comp. **3**(4), 257–271 (1999)
19. Zitzler, E., Thiele, L., Laumanns, M., Fonseca, C.M., da Fonseca, V.G.: Performance assessment of multiobjective optimizers: an analysis and review. IEEE Trans. Evol. Comp. **7**(2), 117–132 (2003)