# A New Multi-objective Approach for Molecular Docking Based on RMSD and Binding Energy

Esteban López-Camacho[(⊠)], María Jesús García-Godoy, José García-Nieto,
Antonio J. Nebro, and José F. Aldana-Montes

Khaos Research Group, Department of Computer Science, University of Málaga,
ETSI Informática, Campus de Teatinos, 29071 Málaga, Spain
{esteban,mjgarciag,jnieto,antonio,jfam}@lcc.uma.es

**Abstract.** Ligand-protein docking is an optimization problem based
on predicting the position of a ligand with the lowest binding energy
in the active site of the receptor. Molecular docking problems are tra-
ditionally tackled with single-objective, as well as with multi-objective
approaches, to minimize the binding energy. In this paper, we propose a
novel multi-objective formulation that considers: the Root Mean Square
Deviation (RMSD) difference in the coordinates of ligands and the bind-
ing (intermolecular) energy, as two objectives to evaluate the quality of
the ligand-protein interactions. To determine the kind of Pareto front
approximations that can be obtained, we have selected a set of represen-
tative multi-objective algorithms such as NSGA-II, SMPSO, GDE3, and
MOEA/D. Their performances have been assessed by applying two main
quality indicators intended to measure convergence and diversity of the
fronts. In addition, a comparison with LGA, a reference single-objective
evolutionary algorithm for molecular docking (AutoDock) is carried out.
In general, SMPSO shows the best overall results in terms of energy and
RMSD (value lower than 2Å for successful docking results). This new
multi-objective approach shows an improvement over the ligand-protein
docking predictions that could be promising in *in silico* docking stud-
ies to select new anticancer compounds for therapeutic targets that are
multidrug resistant.

**Keywords:** Molecular docking · Multi-objective optimization · Nature
inspired metaheuristics · Algorithm comparison

## 1 Introduction

Ligand-protein docking is an optimization problem which aims at predicting the
position of a small molecule (ligand) to a receptor (macromolecule) with the goal
of finding the ligand position to the receptor with a minimum binding energy.
Molecular docking problem has been tackled with single-objective algorithms,
to minimize the binding energy [11], as well as with multi-objective approaches,
to minimize the intermolecular energy $E_{inter}$ (energy interaction between lig-
and and the target) and the intramolecular energy $E_{intra}$ (the internal energy
compound) [4].

In this regard, a number of studies based on the application of multi-objective algorithms to the ligand-protein docking have been proposed. A first attempt was carried out in 2006 by Oduguwa et al. [15], in which three evolutionary multi-objective algorithms (NSGA-II, PAES, and SPEA) were applied to evaluate three objectives such as the $E_{inter}$, $E_{intra}$ and shape complementarities on three molecular complexes. Grosdidier et al. [5] proposed a new hybrid evolutionary algorithm called EADock that optimizes two different energy score functions that evaluate the $E_{inter}$, $E_{intra}$ and the solvation free energy. In 2008, Janson et al. [7] designed a parallel multi-objective algorithm using AutoDock 3.05 energy function, called ClustMPSO, minimizing as objectives the $E_{inter}$ and $E_{intra}$ when dealing with six molecular complexes. In the same year, Boisson et al. [1] implemented a parallel evolutionary bi-objective model based on optimizing two objectives: the sum of $E_{inter}$ and $E_{intra}$ and a surface term for the docking of six instances. Sandoval-Perez et al. [16] used the implementation of NSGA-II provided by the jMetal framework to optimize bound and non-bound energy terms of ligand/receptor as objectives applied to four docking instances. Gu et al. [6] developed a new multi-objective approach based on optimizing the solutions generated by an aggregated scoring function that includes terms from force-field, empirical and knowledge-based scoring functions.

In all these previous publications, a series of different multi-objectives formulations have been proposed that focus on energy scoring function. However, they do not consider guiding the search with a new objective when the co-crystallized ligand is known, which could complement the traditional energy function.

With this motivation, we propose in this work a novel multi-objective approach consisting minimizing: (1) the binding energy (the unbound and bound energy terms of the ligand/receptor complex), and (2) the Root-Mean-Square-Deviation (RMSD) score, when the co-crystallized ligand pose is known. These two main objectives have been used to evaluate the quality of the ligand-protein interactions. With this aim, we compare and analyze the performance of four multi-objective metaheuristics when solving 11 flexible ligand-receptor docking complexes taken from the AutoDock 4.2 benchmark [12]. This dataset includes flexible ligands with different sizes and flexible side-chains of HIV-protease receptors for more realistic results. The algorithms used in this study are: Nondominated Sorting Genetic Algorithm II (NSGA-II) [2], Speed Modulation Multi-Objective Particle Swarm Optimization (SMPSO) [13], Third Evolution Step of Generalized Differential Evolution (GDE3) [8], and Multi-Objective Evolutionary Algorithm Based on Decomposition (MOEA/D) [18]. These algorithms constitute a varied set of evolutionary and difference-vector multi-objective techniques representative of the state of the art, performing different learning procedures and inducing different behaviors in terms of convergence and diversity.

This paper is organized as follows: Sect. 2 describes the molecular docking problem from a multi-objective formulation. Studied algorithms are briefly described in Sect. 3. Section 4 reports the experimentation methodology and Sect. 5 analyzes the results obtained. Finally, Sect. 6 contains concluding remarks and future lines of research.

## 2   The Problem: Multi-objective Docking

A multi-objective optimization problem is characterized by two spaces: the decision and the objective spaces. The former involves all the possible feasible solutions, and the latter includes their corresponding objective values.

**Decision space.** 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 the lowest binding energy. The ligand-receptor interaction is evaluated by an energy function calculated through 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. Each problem solution for AutoDock and jMetal (the tools we have used) is encoded by a real-value vector of $7 + n$ variables, in which the first three values correspond to the ligand translation, the next four values correspond to the ligand and/or macromolecule orientation, and the remaining n values are the ligand torsion dihedral angles. Furthermore, in order to allow a rapid evaluation of the energy conformations, a grid-based methodology is implemented. The energy interaction is calculated and assigned to each grid point and is evaluated to obtain the energy of a given ligand pose [12].

**Objective space.** Our bi-objective formulation consists of: the $E_{inter}$ and the RMSD score. The $E_{inter}$ is the energy function as used in Autodock, that is calculated as follows:

$$E_{inter} = Q_{bound}^{R-L} + Q_{unbound}^{R-L} \tag{1}$$

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

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

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. $r_{ij}$ represents the interatomic distance, $A_{ij}$ and $B_{ij}$ in the first term are Lennard-Jones parameters taken from the Amber force field. Similarly, $C_{ij}$ and $D_{ij}$ in the second term are Lennard-Jones parameters for maximum well depth of potential energies between two atoms, and $E(t)$ represents the angle-dependent directionality. The third term uses a Coulomb

approach for electrostatics. Finally, the fourth term is calculated from the volume ($V$) of the atoms that are surrounding a given atom weighted by $S$, and an exponential term which involves atom distances. An extended explanation of all these variables can be found in [12].

The RMSD is a measure of similarity between the real ligand position in the receptor and the computed position of the docking ligand, that takes into account symmetry, partial symmetry (e.g. symmetry within a rotatable branch) and near-symmetry in a simple heuristic way. Ideally, the lower RMSD score the better solution is. A ligand-receptor docking solution with a RMSD score below 2Å is considered as a solution with high docking accuracy. It is worth noting that other docking solutions can be returned with higher RMSD scores and low values of $E_{inter}$, indicating that other possible interaction ligand sites should be considered. The RMSD score for two identical structures $a$ and $b$ is defined as follows:

$$RMSD_{ab} = max(RMSD'_{ab}, RMSD'_{ba}), \ with \ RMSD'_{ab} = \sqrt{\frac{1}{N} \sum_i \min_{\mathbf{j}} r_2^{ij}}$$

(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

We have included in our study four algorithms which are representative of the state-of-the-art in the multi-objective optimization field. A brief description of each one of them is given next:

**NSGA-II:** NSGA-II [2] is a generational genetic algorithm, which uses the typical genetic operators (selection, crossover and mutation) to obtain new individuals from the original population. To promote convergence, a non-dominated sorting procedure based on Pareto ranking is used, while the crowding distance density estimator is applied to foster the diversity of the set of found solutions.

**GDE3:** The Generalized Differential Evolution (GDE) algorithm [8] is based on NSGA-II, but the genetic mutation and selection operators are replaced by their differential evolution counterparts. Furthermore, GDE3 modifies the crowding distance of NSGA-II as well to generate a better distributed set of solutions.

**SMPSO:** SMPSO [13] is a multi-objective particle swarm optimization algorithm. Its main feature is the limitation of the particle speed to allow new effective particle positions to be produced when the speed becomes too high. SMPSO uses the polynomial mutation as the turbulence factor and an external archive that stores the non-dominated solutions found during the search.

**MOEA/D:** MOEA/D [18] has become the typical representative decomposition-based multi-objective algorithm, where a multi-objective problem is decomposed into a set of single-objective subproblems that then optimized simultaneously. In this study we have used the variant MOEA/D-DE [9], which applies differential evolution as variation operators. This algorithm also applies a polynomial mutation operator to improve its search capability.

In short, we have selected the most widely used algorithm in the field (NSGA-II), a solver based in differential evolution (GDE3), a PSO (SMPSO) and an algorithm based on decomposition (MOEA/D).

## 4   Experimentation

In this section, we include the selected benchmark problems, the experimentation methodology we have followed, and the parameter settings of the algorithms.

### 4.1   Benchmark Problems

In this study, we have selected a benchmark composed of 11 complexes having receptor and ligand flexibility. The selection of these complexes has been motivated as they are actually difficult docking problems containing a wide range of ligand sizes (from small to large inhibitors). The receptors of these complexes have a tunnel-shaped active site that wraps around a peptidomimetic inhibitor [12]. The receptor is a dimer whose subunits are bridged by an arginine-aspartate salt bridge at the end of the tunnel. The docking studies performed with these instances in [12] to test the energy function of AutoDock 4.2 demonstrated that the most difficult problems are those which involve smaller ligands. This is due to the flexibility added to the receptor side-chains (ARG-8) that increases the space of ligand interaction. These instances have been taken from the PDB database[1] and they have been properly prepared for the docking simulations.'

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 (flexibility) for ligands and macromolecules are 10 and 6, respectively, selecting those torsions that allow the fewest number of atoms to move around the ligand core. Therefore, the total number of solution variables ($n$) is 23 (3 for translation, 4 for rotation quaternion, and 16 for torsional degrees).

### 4.2   Methodology

For this work, we have carried out a thorough experimentation consisting in performing 30 independent runs for each combination of algorithm and molecular instance. From these executions, we have calculated the median and interquartile range (IQR) as measures of central tendency and statistical dispersion,

---

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

**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 |

respectively. We have considered two quality indicators to assess the algorithm performance: Hypervolume ($I_{HV}$) and Unary Additive Epsilon Indicator ($I_{\epsilon+}$) [3]. The first indicator takes into account both convergence and diversity, whereas the second one ($I_{\epsilon+}$) gives a measure of the convergence degree of the obtained Pareto front approximations. In this sense, it is worth noting that we are dealing with a real-world optimization problem, and therefore the true Pareto 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.

As mentioned, we have used the implementation of the four algorithms studied provided in the jMetalCpp framework [10] in combination with AutoDock 4.2 to evaluate the new generated solutions. To cope with the high computational requirements needed by carry out our experiments, we have used the Condor[2] system, a middleware platform managing close to 400 cores that acts as a distributed task scheduler (each task dealing with one independent run).

### 4.3   Parameter Setup

The selected algorithms have been configured with a population size of 150 individuals (particles in the case of SMPSO). The stopping condition has been set to compute a number of 1,500,000 function evaluations. These values were chosen because they are the default settings used by AutoDock and they have been used in other studies [14].

Each algorithm has been configured using the parameter setup recommended in the research study where it was proposed, and these parameters are used as

---

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

default in the jMetal framework. In particular, SBX crossover and polynomial mutation are the variation operators used in NSGA-II. The distribution indexes for both operators are $\eta_c = 20$ for crossover, and $\eta_m = 20$ for mutation. The crossover probability is $p_c = 0.9$ and the mutation probability is $p_m = 1/n$, being $n$ the number of decision variables of the tackled problem. NSGA-II applies binary tournament selection. In the case of GDE3 (variant *rand/1/bin*), the two DE control parameters $\mu$ and $C_r$ take a value of 0.5, whereas in MOEA/D $\mu$ is set to 0.5 and $C_r$ is set to 1.0. Both MOEA/D and SMPSO use the polynomial mutation with the same settings applied in NSGA-II. In SMPSO, the acceleration coefficients $\varphi_1$ and $\varphi_2$ are set to 1.5, the inertia weight is $W = 0.9$, and the polynomial mutation is applied to one sixth of the particles in the swarm.

## 5   Results

This section is devoted to presenting and analyzing the results obtained in our study. We start by assessing the performance of the algorithms and then they are compared with the values of a single-objective approach.

### 5.1   Performance Comparisons

We start our analysis by discussing the results yielded by applying the $I_{HV}$ indicator. Let us remind that this indicator is the sum of the contributed volume of each point of a front in respect to a reference point, and the higher the convergence and diversity degrees of a front, the higher its $I_{HV}$ value. Table 2 shows the median and interquartile range of the computed solutions for $I_{HV}$ quality indicators for the set of 11 docking instances and the four algorithms being compared. According to these results, SMPSO achieves the best $I_{HV}$ values in all the eleven considered problems and MOEA/D is the second best performing technique. We have to note that many cells have a $I_{HV}$ value equal to zero; this happens when all the points of the produced fronts are beyond the limits of the reference point. This happens in most of the problems in all the algorithms excepting SMPSO, which indicates we are facing a very hard optimization problem.

**Table 2.** 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.

|  | NSGAII | SMPSO | GDE3 | MOEAD |
|---|---|---|---|---|
| 1AJV | $0.00e+00_{0.0e+00}$ | $3.51e-01_{4.0e-02}$ | $0.00e+00_{0.0e+00}$ | $0.00e+00_{2.9e-01}$ |
| 1AJX | $0.00e+00_{0.0e+00}$ | $5.52e-01_{2.0e-02}$ | $0.00e+00_{0.0e+00}$ | $7.47e-03_{6.8e-01}$ |
| 1D4K | $0.00e+00_{0.0e+00}$ | $4.93e-01_{1.3e-01}$ | $0.00e+00_{0.0e+00}$ | $0.00e+00_{0.0e+00}$ |
| 1G2K | $0.00e+00_{0.0e+00}$ | $3.32e-01_{3.3e-02}$ | $0.00e+00_{0.0e+00}$ | $0.00e+00_{4.1e-01}$ |
| 1HIV | $0.00e+00_{0.0e+00}$ | $5.96e-01_{1.3e-01}$ | $0.00e+00_{0.0e+00}$ | $0.00e+00_{0.0e+00}$ |
| 1HPX | $0.00e+00_{0.0e+00}$ | $2.04e-01_{1.8e-01}$ | $1.27e-01_{6.5e-01}$ | $0.00e+00_{1.1e-01}$ |
| 1HTF | $0.00e+00_{0.0e+00}$ | $5.26e-02_{1.3e-01}$ | $0.00e+00_{4.6e-03}$ | $2.78e-02_{3.3e-01}$ |
| 1HTG | $0.00e+00_{0.0e+00}$ | $3.51e-02_{5.6e-02}$ | $0.00e+00_{0.0e+00}$ | $0.00e+00_{1.9e-01}$ |
| 1HVH | $0.00e+00_{0.0e+00}$ | $7.67e-01_{3.7e-02}$ | $0.00e+00_{0.0e+00}$ | $5.31e-01_{7.7e-01}$ |
| 1VB9 | $0.00e+00_{0.0e+00}$ | $7.34e-01_{6.5e-02}$ | $0.00e+00_{0.0e+00}$ | $0.00e+00_{1.4e-01}$ |
| 2UPJ | $0.00e+00_{0.0e+00}$ | $5.86e-01_{9.8e-02}$ | $0.00e+00_{0.0e+00}$ | $1.90e-01_{5.8e-01}$ |

**Table 3.** 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.

| | NSGAII | SMPSO | GDE3 | MOEAD |
|---|---|---|---|---|
| 1AJV | $5.23e+00_{1.2e+00}$ | $5.60e-01_{9.8e-02}$ | $5.00e+00_{1.0e+00}$ | $3.87e+00_{4.4e+00}$ |
| 1AJX | $3.43e+00_{2.4e+00}$ | $2.61e-01_{7.2e-02}$ | $1.49e+00_{3.3e-01}$ | $1.01e+00_{2.0e+00}$ |
| 1D4K | $8.06e+00_{2.7e+00}$ | $4.56e-01_{1.4e-01}$ | $8.56e+00_{5.7e-01}$ | $4.65e+00_{2.8e+00}$ |
| 1G2K | $4.28e+00_{1.4e+00}$ | $5.71e-01_{1.2e-01}$ | $3.93e+00_{1.3e+00}$ | $2.69e+00_{3.6e+00}$ |
| 1HIV | $5.12e+00_{1.2e+00}$ | $2.63e-01_{2.1e-01}$ | $4.69e+00_{1.4e+00}$ | $4.07e+00_{1.6e+00}$ |
| 1HPX | $1.42e+01_{3.6e+00}$ | $6.32e-01_{2.8e-01}$ | $6.71e-01_{1.1e+01}$ | $1.03e+01_{1.3e+01}$ |
| 1HTF | $1.76e+00_{5.5e-01}$ | $9.30e-01_{3.0e-01}$ | $1.13e+00_{8.0e-01}$ | $7.94e-01_{9.2e-01}$ |
| 1HTG | $7.48e+00_{7.1e-01}$ | $9.63e-01_{6.6e-02}$ | $6.82e+00_{8.7e-01}$ | $5.03e+00_{6.4e+00}$ |
| 1HVH | $5.94e+00_{1.5e+00}$ | $1.34e-01_{2.7e-02}$ | $4.93e+00_{1.7e+00}$ | $4.16e-01_{2.1e+00}$ |
| 1VB9 | $8.59e+00_{2.4e+00}$ | $1.33e-01_{5.6e-02}$ | $7.85e+00_{1.3e+00}$ | $7.04e+00_{4.9e+00}$ |
| 2UPJ | $3.42e+00_{2.4e+00}$ | $3.03e-01_{6.6e-02}$ | $3.56e+00_{1.1e+00}$ | $7.64e-01_{2.7e+00}$ |

**Table 4.** Average Friedman's rankings with Holm's Adjusted $p$-values ($\alpha = 0.05$) of compared algorithms (SMPSO, GDE3, MOEA/D, and NSGA-II) 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 (HV) | | | Epsilon ($I_{\epsilon+}$) | | |
|---|---|---|---|---|---|
| Algorithm | $Fri_{Rank}$ | $Holm_{Ap}$ | Algorithm | $Fri_{Rank}$ | $Holm_{Ap}$ |
| **\*SMPSO** | **1.02** | - | **\*SMPSO** | **1.09** | - |
| MOEA/D | 2.68 | 2.24e-03 | MOEA/D | 2.00 | 9.87e-02 |
| GDE3 | 3.09 | 1.45e-04 | GDE3 | 3.09 | 2.79e-04 |
| NSGA-II | 3.22 | 5.21e-05 | NSGA-II | 3.81 | 7.25e-07 |

A similar behavior can be observed in Table 3 with regards to $I_{\epsilon+}$, which is a convergence measure. According to these results, SMPSO obtains the best $I_{\epsilon+}$ values in ten out of the eleven problems, while MOEA/D gets the best value in one problem (1HTF) and the second best in all of them but 1HPX.

In order to provide these results with statistical meaning (in this study $\alpha = 0.05$), non-parametric statistical tests have been applied because in several cases the distributions of results did not follow the conditions of normality and homoscedasticity [17]. Therefore, the analyses and comparisons focus on the entire distribution of each of the two metrics studied. Specifically, we have applied Friedman's ranking and Holm's post-hoc multicompare tests [17] to know which algorithms are statistically worse than the control one (with the best ranking).

In this regard, as shown in Table 4, SMPSO is the best ranked technique according to $I_{HV}$ (with a value of 1.02), followed by MOEA/D, GDE3, and NSGA-II. Therefore, SMPSO is established as the control algorithm for $I_{HV}$ in the post-hoc Holm test, which is compared with the remaining algorithms. The adjusted $p$-values ($Holm_{Ap}$ in Table 4) resulting from these comparisons are, for the last three algorithms (MOEA/D, GDE3, and NSGA-II), lower than the confidence level, meaning that SMPSO is statistically better than these algorithms.
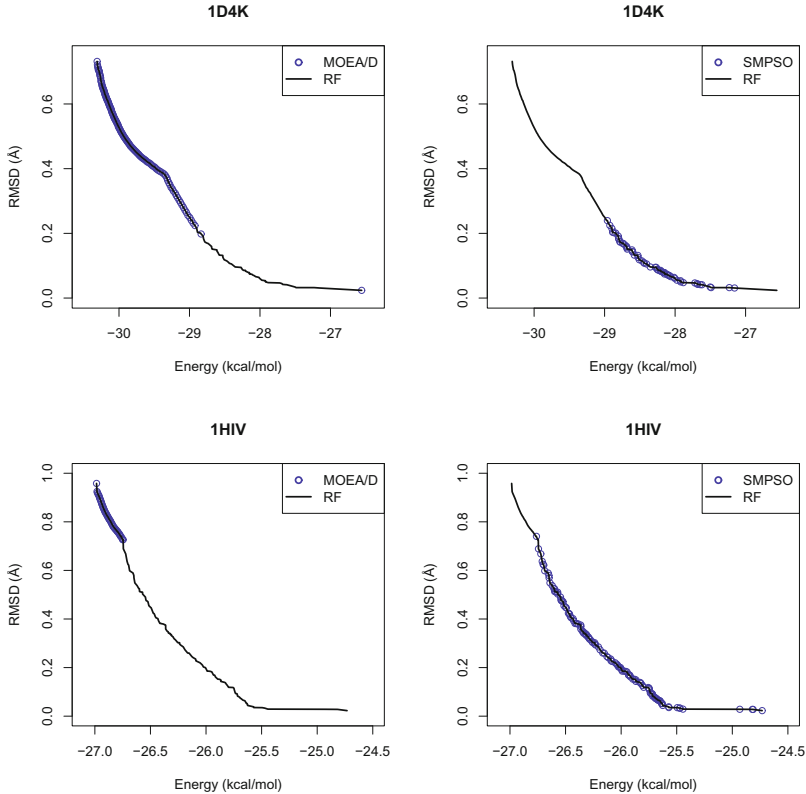
**Fig. 1.** Reference front contributions of docking instances 1D4K and 1HIV. SMPSO and MOEA/D contribute with practically all the solutions of the reference fronts.

In the case of $I_{\epsilon+}$, SMPSO is better ranked than the remaining compared algorithms, although without statistical differences in the case of MOEA/D. SMPSO is statistically better than GDE3 and NSGA-II.

In summary, SMPSO shows the overall best balance for the two quality indicators, followed by MOEA/D. These results are graphically supported by two examples included in Fig. 1, where the reference fronts obtained for two representative instances 1D4K and 1HIV are plotted. In these graphs, the contributions, in terms of solutions, of each algorithm to the global reference front are plotted with different points and colors. As it is easily observable, SMPSO and MOEA/D contribute with almost all solutions taking part to the reference front. Interestingly, SMPSO converges to the region biased towards the RMSD objective, whereas MOEA/D generate non-dominated solutions in a different region to the ones of SMPSO, thereby giving cue to the energy optimization. We can state that the specific learning procedures induced by SMPSO and MOEA/D lead these algorithms to search in different regions of the problem landscape, hence generating solutions in complementary parts of the reference front.

**Table 5.** Best RMSD scores (Å) calculated from all SMPSO solutions in comparison with the best RMSD values of LGA single-objective solutions.

|       | 1AJV | 1AJX | 1BV9 | 1D4K | 1G2K | 1HIV | 1HPX | 1HTF | 1HTG | 1HVH | 2UPJ |
|-------|------|------|------|------|------|------|------|------|------|------|------|
| SMPSO | 0.02 | 0.03 | 0.02 | 0.03 | 0.02 | 0.02 | 0.03 | 0.03 | 0.00 | 0.11 | 0.03 |
| LGA   | 5.23 | 5.12 | 4.46 | 4.46 | 4.46 | 6.47 | 6.09 | 3.89 | 0.38 | 6.71 | 4.55 |
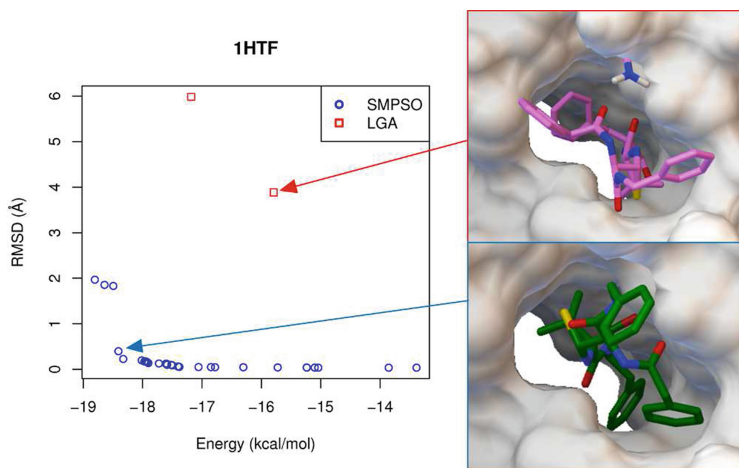


**Fig. 2.** Set of non-dominated solutions obtained by SMPSO with regards to those by LGA, for instance 1HTF. Corresponding ligand conformation structures (captured from AutoDock) of two representative solutions are shown at right side.

## 5.2   Comparison Single Versus Multi-objective

After the performance comparison of multi-objective algorithms, we are now interested in knowing how competitive their solutions are against those yielded by the LGA single-objective technique provided by Autodock 4.2. This way, we will be able to determine whether our bi-objective formulation has a positive effect in the search of solutions with the lowest RMSD score, or not. The values are included in Table 5, in which we can observe that SMPSO outperforms LGA in all the instance problems with large and small ligands.

With the multi-objective approach proposed here, SMPSO is able to return better results (RMSD scores below 2Å) for all the instances, since the optimization procedure is actively guided to compute solutions to the real ligand pose.

The use of RMSD as objective to guide the search procedure could be counter-intuitive, since it would restrict us to work only with molecular structures whose co-crystallized ligand are known beforehand (experimentally determined). Nevertheless, this new focus is useful in those typical cases in which the active site of a given therapeutic target mutates and makes it multidrug resistant. Therefore, new compounds analogous to the reference ligand should be tested to be considered as new pharmacological candidates.

Figure 2 shows the non-dominated solutions obtained by SMPSO, for molecule 1HTF. In addition, the two solutions with the best binding energy and RMSD values obtained by LGA are also plotted. In order to visualize the computed ligand docked to the active site of the HIV-protease receptor, we have selected two solutions with the best RMSD values. The best RMSD solutions for the LGA and the SMPSO are 3.89Å and 0.39Å with binding energies of $-5.99$ kcal/mol and $-18.04$ kcal/mol, respectively. As Fig. 2 shows, the ligand computed by the SMPSO is docked to the active site of the receptor with a better pose than the LGA computed ligand that is partially inside.

## 6    Conclusions

In this paper, we propose a novel multi-objective formulation of the molecular docking problem, where the RMSD and binding energy are the goals to optimize. This new approach has been incorporated in four multi-objective algorithms: NSGA-II, SMPSO, GDE3 and MOEA/D. A heterogenous set of 11 protein-ligand complexes with flexible ligands and receptors were selected in order to carry out the experiments. The main conclusions can be outlined as follows:

1. Using a multi-objective approach to solve the molecular docking could lead to a broad set of solutions, which can be selected according to the weight of the RMSD and binding energy, instead of only focusing on energy values.
2. SMPSO provides the best overall performance according to the two quality indicators used, and for the studied molecular instances.
3. For all studied molecular instances, SMPSO converges to the region biased towards the RMSD, whereas MOEA/D generates its fronts of non-dominated solutions in a different region, thereby giving cue to energy optimization.
4. According to the single-objective (AutoDock 4.2) fitness function, SMPSO algorithm find, in most of the cases, better solutions than the ones obtained by LGA. This is a noticeable result since SMPSO is a general purpose optimization technique, while LGA is specifically adapted to deal with the molecular docking problem.
5. The use of RMSD as objective to guide the search is useful in those typical cases in which the active site of a given therapeutic target mutates and makes it multi-drug resistant.

As future work, the most natural extension would be to design a hybrid algorithm combining search procedures from both SMPSO and MOEA/D algorithms in order to get solutions covering the full Pareto front. To test this, a greater number of molecular instances could be used and the solutions obtained could be studied from a more biological point of view.

# References

1. Boisson, J.C., Jourdan, L., Talbi, E., Horvath, D.: Parallel multi-objective algorithms for the molecular docking problem. In: IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology, pp. 187–194, September 2008

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. Deb, K.: Multi-objective Optimization Using Evolutionary Algorithms. Wiley, New York (2001)

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. Grosdidier, A., Zoete, V., Michielin, O.: EADock: docking of small molecules into protein active sites with a multiobjective evolutionary optimization. Proteins **67**(4), 1010–1025 (2007)

6. Gu, J., Yang, X., Kang, L., Wu, J., Wang, X.: MoDock: a multi-objective strategy improves the accuracy for molecular docking. Algorithms Mol. Biol. **10**, 8 (2015)

7. Janson, S., Merkle, D., Middendorf, M.: Molecular docking with multi-objective particle swarm optimization. Appl. Soft Comput. **8**(1), 666–675 (2008)

8. Kukkonen, S., Lampinen, J.: GDE3: the third evolution step of generalized differential evolution. In: The 2005 IEEE Congress on Evolutionary Computation, vol. 1, pp. 443–450 (2005)

9. Li, H., Zhang, Q.: Multiobjective optimization problems with complicated pareto sets, MOEA/D and NSGA-II. IEEE Trans. Evol. Comput. **13**(2), 229–242 (2009)

10. 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)

11. 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)

12. 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)

13. Nebro, A.J., Durillo, J.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, March 2009

14. Norgan, A.P., Coffman, P.K., Kocher, J.P.A., Katzmann, D.J., Sosa, C.P.: Multi-level parallelization of AutoDock 4.2. J. Cheminform. **3**(1), 12 (2011)

15. Oduguwa, A., Tiwari, A., Fiorentino, S., Roy, R.: Multi-objective optimisation of the protein-ligand docking problem in drug discovery. In: Proceedings of the 8th Annual Conference on Genetic and Evolutionary Computation, pp. 1793–1800 (2006)

16. 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)

17. Sheskin, D.J.: Handbook of Parametric and Nonparametric Statistical Procedures. Chapman & Hall/CRC, New York (2007)
18. Zhang, Q., Li, H.: MOEA/D: a multiobjective evolutionary algorithm based on decomposition. IEEE Trans. Evol. Comput. **11**(6), 712–731 (2007)