Journal of Computer-Aided Molecular Design manuscript No.
(will be inserted by the editor)

Robust ligand-protein docking using an advanced genetic algorithm

David C. Williams · Eric Chang · David B. Kita · Douglas Pahel · Qiang Wang

Received: date / Accepted: date

Abstract Computational prediction of ligand-protein binding is an essential part of overcoming the synthesis bottleneck for rapid development of small-molecule drugs. Nevertheless, an accurate prediction of binding free energy is challenging without first establishing a ligand binding pose. Reported in this paper is an advanced optimization algorithm that is capable of providing multiple, high-fidelity solutions to a ligand-protein pose even for complicated systems with large numbers of degrees of freedom. This algorithm achieves high performance by incorporating several important features, such as niching and force-field annealing. Results from several challenging use cases are presented and discussed.

Keywords protein-ligand interactions · molecular docking · optimization · genetic algorithm

1 Introduction

Hit finding remains one of the most important rate limiting steps in small-molecule drug development [1–3]. In those cases where the target protein has been identified and its structure established, computational methods hold tremendous promise, allowing potential candidate compounds to be identified without investing precious time and resources into synthesis [4]. A computational approach is particularly suitable for protein targets that have resisted conventional high-throughput screening because it facilitates the exploration of novel chemical space [5,6]. On the other hand,

David C. Williams
Verseon Corporation
Fremont, CA 94538
E-mail: dwilliams@verseon.com

Eric Chang · David B. Kita · Douglas Pahel · Qiang Wang
Verseon Corporation
Fremont, CA 94538

the primary goals of hit finding, namely potency and selectivity, impose constraints on the accuracy of computational methods that are challenging to achieve [7,8].

In a typical drug discovery program, the goal of computational methods during hit finding is to replicate a primary assay, such as enzyme inhibition, through a prediction of ligand-protein binding affinity [3,6]. There are several important contributions to binding affinity that involve the ligand-protein complex, but most can be considered a function of the configuration space of the relative atom coordinates of the ligand and protein (and coordinated water molecules). As such, an accurate prediction of the docked ligand pose is an important aspect in such calculations. A docked pose is especially important for methods, such as free-energy perturbation [9], that use first-principle calculations of binding free energy. Thus, a robust prediction of the ligand-protein binding pose is usually desirable.

The traditional approach to ligand-protein docking is to search the configuration space of the ligand for a configuration that represents the minimum Gibbs free energy, or the minimum of some reasonable proxy for the Gibbs free energy. The difference in Gibbs free energy ΔG between two states (bound and unbound ligand) can be represented by a combination of enthalpy (ΔE) and entropy (ΔS) terms:

$$\Delta G = \Delta E - T\Delta S\,, \tag{1}$$

where T is the fixed temperature of the system. The ΔG for a ligand-protein complex is a macroscopic property and a characteristic of a thermodynamic ensemble that cannot be fully represented by a single ligand pose [10]. Nevertheless, there are many techniques for calculating approximate values for ΔE from a single pose. A major complication is the influence of the solvent environment, particularly for electrostatic contributions. To demonstrate the algorithm reported in this paper, a ΔE calculation is employed using a sum of pairwise energies among the ligand and protein atoms, with a distance-dependent dielectric scheme to account indirectly for the influence of solvent. Details are discussed in the following section.

An approximate parameterization of ΔS is more theoretically challenging [11]. For this reason, a common approach when establishing a docked ligand pose is to ignore the entropy contribution to ΔG (except, perhaps, for entropic contributions of the solvent, depending on methodology [12]), or assume it is proportional in magnitude to ΔE. This is the approach used in this paper, with the understanding that the accuracy of such assumptions is questionable [13].

Finding the ligand-protein configuration that minimizes an approximate representation of enthalpy is a practical approach to simplifying the process of identifying the ligand-protein docked pose and reducing associated computational requirements. Once a ligand-protein pose is established, more defensible calculations of ΔG can be employed to calculate the corresponding binding free-energy. This approach is commonly referred to as "scoring" a pose [14]. The drawback is that it can be difficult to justify why the pose that minimizes ΔE is the same as or even related to the pose that minimizes ΔG.

A simple solution to mitigate the risks that the minimization of ΔE does not produce the correct ΔG minimum pose is to provide multiple solutions, each corresponding to a local minimum in ΔE. The presentation of multiple and valid pose solutions, an option often overlooked in many approaches to ligand-protein docking,

is a natural by-product of the genetic algorithm described in this paper. Each solution can be quantified by the desired scoring methodology, re-minimized if appropriate (using localized methods such as gradient minimization), and filtered accordingly.

An exhaustive search of the entire configuration space of a ligand pose is not considered computational tractable because of its size (d^{3n}, for n ligand atoms, in a box of size d per side). A reasonable compromise is to limit the configuration space to one that preserves bond lengths and angles (because altering either is usually expensive in terms of enthalpy). In the simplest case (*e.g.* neglecting ring conformations), this limits the degrees of freedom to overall displacement, overall rotation, and the torsional angles of an appropriate selection of ligand bonds. Even under this simplification, the configuration space is still enormous for all but the simplest ligands.

To appreciate the size of the configuration space, consider the following example. If the overall displacement of the ligand is limited to 5Å in all directions and is sampled at 0.2Å increments, this produces 25^3 combinations. If overall rotation is incremented in 6° increments, this produces an additional $60 \times 60 \times 30$ combinations. Finally, for a ligand with five available torsional angles, each incremented by 6°, this produces 60^5 combinations. The result is a total configuration space of more than 10^{18} independent combinations (ignoring symmetries). Even in this simplified example, an exhaustive survey of the entire conformational space would be computationally intractable.

The large size of the conformal space of possible pose solutions would not be an issue if the objective function (the approximate enthalpy) was a convex function and the minimum reachable through standard techniques, such as gradient minimization. In contrast, it is intuitively apparent that even the simplest ligand problems introduce multiple minima. A relevant example is water molecules, which are universally identified experimentally at multiple locations both inside and on the surface of proteins. Another example is the large conformal solution space of isolated ligands, even when limited to torsional freedoms.

The solution for multiple minima in non-convex systems is an unsolved problem in optimization [15,16]. For ligand-protein docking, minima are typically separated by large barriers, making iterative searches impractical [17]. In addition, the high-dimensional properties of the solution space is a particular challenge that conventional approaches have failed to fully solve.

The genetic algorithm described in this paper is one of a suite of tools designed and implemented by Verseon for its computational drug discovery platform. It was developed specifically to tackle the challenging problem of identifying multiple, viable minima in the ligand configuration space of a reasonably detailed model of the enthalpy of ligand-protein binding. The algorithm includes advanced techniques that distinguish it from previously reported optimization methods and which provide sufficient capacity to reproduce a good representation of the experimental ligand pose in many nontrivial test cases.

2 Test conditions

For the purposes of demonstrating the effectiveness of the genetic algorithm, the problem of ligand-protein docking is reduced to the following simplified form:

- The docked pose of the ligand is experimentally established and available in a PDB entry.
- The protein structure, as established by the PDB entry, is considered rigid, contains only biologically relevant hetero atoms, and is protonated by established methods.
- The internal conformal freedom of a ligand is represented entirely by the dihedral rotation of any freely rotatable bonds.
- The ligand is assumed to bind in the vicinity of the experimental pose.
- The ionization and tautomer state of the ligand is known in advance and does not change upon binding.

These simplifications are chosen to clarify the presentation of the results in this paper and do not necessarily represent implicit limitations of the algorithm. Additional considerations associated with a practical drug development program are addressed later.

For the purposes of this test, success is defined as a recovery of the experimental pose of the ligand. To avoid any risk of bias, the translation, rotation, and torsional angles of the ligand pose are randomized at input.

For practical reasons, the translational freedom of the ligand must be restricted to a certain region. The configuration space associated with translation is proportional to the volume of this restriction, and, as such, can have a dramatic impact on the performance of an optimization algorithm. As a reasonable compromise between realism and practicality, for the purpose of this test, the translational freedom of the ligand will be restricted to within 10 Å of the experimental pose. There will be no restriction on rotation.

All proteins were prepared using the standard protein preparation tools provided by the Chemical Computing Group [18] using MOE version 2018.1. Default settings ("QuickPrep") were applied for protonation, charge, and termination. The ligand was included during structure relaxation. The selection of protein chains, hetero atoms, and any necessary repairs are described individually for each system later in this paper.

For simplicity of presentation, the objective function of the genetic algorithm will be an approximate representation of ΔE, calculated for the atoms of the ligand, protein, and any relevant protein hetero atoms. This approximation is constructed from the MMFF94 forcefield [19–22], modified to include a distant dependent dielectric. The following is a summary of the relevant aspects of the MMFF94 forcefield, including a description of this modification.

The MMFF94 forcefield separates molecule energies into three major contributions: strain, Van der Waals, and electrostatic. The strain terms (bond stretching, angle bending, stretch bend, out-of-plane bending, and torsion interactions) are used unaltered in this paper (of these, only the torsion interactions are relevant under the stated

conditions). The Van der Waals contribution E_V is the sum of pairwise terms $E_{V,ij}$:

$$E_V = \sum_{i>j} E_{V,ij} \ , \tag{2}$$

where the sum is over all atoms that are either in separate molecules or are separated by at least three covalent bonds within the same molecule. The pairwise term $E_{V,ij}$ has the following parameterization:

$$E_{V,ij} = \varepsilon_{ij} U_V(r_{ij}/\sigma_{ij}) \ , \tag{3}$$

where ε_{ij} is the Van der Waals well depth, σ_{ij} the distance of the minimum, and U_V is the scaled bare Van der Waals potential function:

$$U_V(s) = \left(\frac{1.07}{s+0.07}\right)^p \left(\frac{1.12}{s^p+0.12} - 2\right) \ . \tag{4}$$

Here p is a characteristic exponential power, nominally set to the value of 7. Both ε_{ij} and σ_{ij} are parameterized according to explicit rules defined by the MMFF94 atom typing scheme [19,20].

The modified electrostatic contribution E_Q presented in this paper is also the sum of pairwise contributions:

$$E_Q = \sum_{i>j} \alpha_{ij} E_{Q,ij} \tag{5}$$

where the sum is over all atoms that are either in separate molecules or are separated by at least three covalent bonds within the same molecule. In the special case that the pair of atoms are within the same molecule but separated by no less than four covalent bonds, α_{ij} is set to 0.75, as normally specified by the MMFF94 scheme. Otherwise, $\alpha_{ij} = 1$. The pairwise terms $E_{Q,ij}$ has the following parameterization:

$$E_{Q,ij} = \frac{Q_0}{\kappa(r_{ij})} \frac{q_i q_j}{r_{ij}} \ , \tag{6}$$

where the constant $Q_0 = 332.0716$ is to express the result in the desired units (kcal/mol), r_{ij} is the distance between two atoms in Å, q_i a parameterization of the atom partial charges based on the MMFF94 atom typing [19,20], and $\kappa(r)$ is a distance-dependent effective dielectric intended to compensate for solvation effects [23]. The function $\kappa(r)$ has the form:

$$\kappa(r) = A + \frac{B}{1 + k\exp(-\lambda B r)} \ , \tag{7}$$

where $A = -8.5525$, $\lambda = 0.003627$, $k = 7.7839$, $B = \kappa_0 - A$, and such that κ_0, the dielectric constant at infinite distance, is equal to 78.4, to match the dielectric constant of water [23].

Note that the expression for $E_{Q,ij}$, outside of the distance-dependent dielectric $\kappa(r)$, consists of a pure $1/r$ Coulomb form and does not contain a buffering constant as normally specified in the MMFF94 scheme. This choice was made for convenience and because the implementation described below includes additional modifications to $E_{Q,ij}$ that eliminates the need to include buffering to hide the singularity at $r = 0$.

3 Parameterization

Under the approximation that the internal freedom of a ligand is limited to torsional degrees of freedom, the configuration space of the ligand pose can be completely described by three sets of parameters:

- Overall translation
- Overall rotation
- The dihedral angles of each rotatable bond

For a simple, robust implementation, it is preferable to keep these three parameters independent. To accomplish this, the overall translation is defined in terms of the position of a selected atom in the ligand, referred to as the "root" atom, and the rotational state defined in terms of a rotation around that atom's position.

To minimize the effect of rotation on overall translation of the ligand, it is preferable to select a root atom near the center of mass, with the understanding that the center can vary depending on the value of the dihedral angles. The selection of the root atom is therefore performed by a simple graph analysis of the ligand, as follows.

For the purposes of establishing the root atom, all hydrogens in the ligand are ignored. For each heavy atom, all independent outward projecting atom graph traversals are identified. For each of these paths a sum is calculated where each bond contributes two if rotatable or one otherwise. The maximum sum over all paths is then determined for each atom. The atom with the smallest maximum is assigned as the root atom. Ties are broken by selecting the atom that appears first in the molecule file, which, for the purposes of this paper, can be considered random.

To keep the optimization problem tractable and consistent with the test conditions outlined in the previous section, the translation of the root atom is limited to within 10 Å of the experimental value (see Fig. 9 for a representative illustration of this constraint). Since these limits depend on an experimental pose, such a restriction cannot be generalized to arbitrary ligand candidates, but rather is used for the sake of simplicity of presentation. Alternatives are discussed later in this paper.

Overall rotation of the ligand is defined in terms of the three Euler angles for a frame of reference centered on the root atom. Since the rotation is centered on the root atom, a change in rotation does not effect the translation parameters.

Each dihedral angle is defined with respect to the input structure. A modification of a dihedral angle corresponds to a rotational transform along the axis of the bond to be applied to all atoms reachable from the side of the bond opposite the root atom. In this fashion, a change in dihedral angle does not affect the position of the root atom and thus leaves the overall translation and rotation parameters unaltered.

The objective function of the genetic algorithm is based on the parameterized enthalpy of binding ΔE as defined by the modified MMFF94 force field, as described in the previous section. Because the protein is treated as rigid, the pairwise potential between any two atoms in the protein is ignored. Concerning the calculation of pairwise potential between the protein and ligand, for computational efficiency, protein atoms are divided into two categories: those that are included in an explicit pairwise calculation, and those whose electrostatic contribution are approximated on a grid.

To identify the protein atoms to include in an explicit pairwise calculation with the ligand, the 3D structure of the ligand is analyzed to determine its maximum extent with respect to the root atom for all possible dihedral values. This extent is added to the ± 10 Å cube defining the translational space of the root atom in order to establish a grid space. All protein atoms that are within 5 Å of this grid space are assigned to the pairwise category.

By construction, protein atoms assigned to the grid will remain more than 5 Å from any ligand atom. It is therefore assumed that Van der Waals contributions will be small and can be neglected. The electrostatic potential between these protein atoms and the ligand is approximated using the sum of the electrostatic pair potential of all associated protein atoms pre-calculated at points on a Cartesian grid with a spacing of 0.4 Å along all three dimensions. This pre-calculation is performed by presenting a unit test charge at all grid points. The corresponding pair potential for each ligand atom is calculated as a product of the atom charge and a trilinear interpolation along the grid using the atom position.

The genetic algorithm is performed in 1000 successive iterations. As the algorithm progresses, various parameters used in the algorithm are annealed every tenth iteration according to a predetermined schedule. Among those parameters is one designed to soften the MMFF94 force field potential between the ligand and protein and help convergence during early iterations.

This softening parameter is a threshold value h_V, defined as:

$$h_V = h_V^0 \cdot 10^{i_{10}/100} \,, \tag{8}$$

where i_{10} is incremented every tenth iteration, starting from zero. In this fashion, h_V starts at h_V^0 and rises to 10 times that value by the last iteration. This parameter is used to define a companion parameter s_V, the distance at which the scaled Van der Waals function U_V (Eq. 4) is equal to h_V:

$$h_V = U_V(s_V) \,. \tag{9}$$

The parameters h_V and s_V are used to define the truncated function U_V^*:

$$U_V^*(s) = \begin{cases} h_V\,(2 - s/s_V) & s < s_V \\ U_V(s) & s \geq s_V \end{cases} \tag{10}$$

The function $U_V^*(s)$ behaves like $U_V(s)$ in the nominally physical region above s_V but changes into a simple linear function below. This eliminates large positive values of enthalpy that can confuse an optimization algorithm, while leaving a gentle slope to guide minimization.

For atom pairs between the ligand and protein, the truncated U_V^* potential is employed with $h_V^0 = 8$ kcal/mol and $p = 4$ to calculate the Van dar Waals contribution to the enthalpy. The resulting softened form of the potential is illustrated in Fig. 1. For atom pairs inside the ligand, a less restrictive value of $h_V^0 = 100$ kcal/mol is employed with a nominal value of $p = 7$.

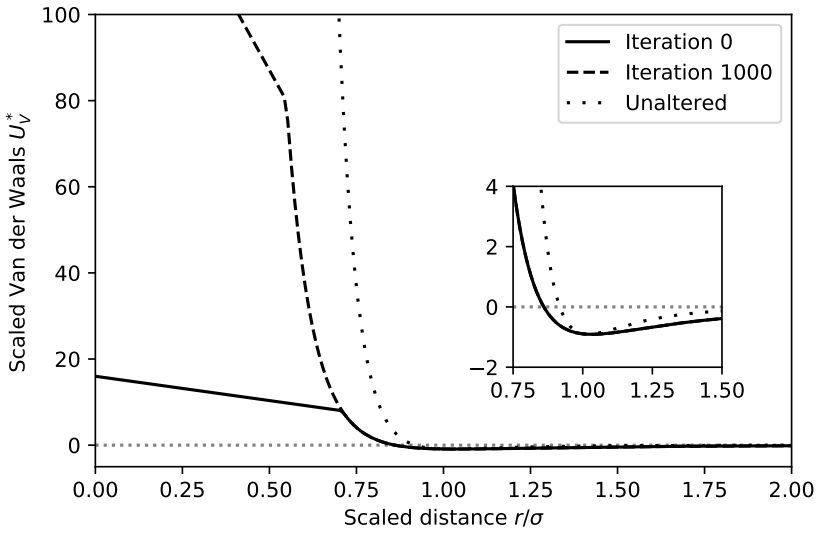


Fig. 1 Softening and truncation of the MMFF94 scaled Van der Waals potential function U_V as used for atom pairs between the ligand and protein.

Some care is required when truncating atom pair potentials. To avoid introducing artificial enthalpy minima, the companion electrostatic pair potential is truncated in a consistent fashion:

$$E^*_{Q,ij}(r_{ij}) = \begin{cases} E_{Q,ij}(s_V \sigma_{ij}) & r_{ij} < s_V \sigma_{ij} \\ E_{Q,ij}(r_{ij}) & r_{ij} \geq s_V \sigma_{ij} \end{cases}, \tag{11}$$

where σ_{ij} is the Van der Waals minimum distance, as introduced in Eq. 3.

To measure the similarity between two poses, a root-mean-square deviation (RMSD) metric is applied to all heavy atoms. To deal with molecule symmetries, all equivalent symmetric permutations are tested, and the minimum used. For the purposes of symmetry, atomic number, connectivity, and formal charge are considered.

4 Description of the algorithm

The genetic algorithm locates enthalpy minima using the following steps:

1. Create an initial set of representative poses by evenly sampling the configuration space in an unbiased fashion
2. Randomly select between mutation and crossover operations and use the selected operation to create an alternate set of representatives
3. Perform a clustering analysis on the combination of original and alternate representatives to select from those two sets those representatives that have good enthalpy and are sufficiently diverse (“niching”)

4. Repeat from step 2 for the specified number of iterations
5. Perform a final clustering analysis after the last iteration to select the set of output poses

During progression of the optimization algorithm, the state of convergence is represented entirely by a set of representative ligand poses. Each pose consists of Cartesian coordinates for each atom in the ligand, a rotational frame consisting of three Euler angles, and a corresponding enthalpy value. Evaluation of the enthalpy is based on the ligand coordinates and varies dependent on the current iteration, counting from 0 to 1000, due to the annealing of the h_V parameters, as discussed in the previous section.

In principle, the rotational frame is redundant, since all relevant information of the ligand pose can be extracted from the atom coordinates. The rotational frame is included to make rotational operations explicit and simpler to apply.

During the progression of the algorithm, the size of the set of representative poses is nominally fixed to a pre-established population size. This population size can be reduced or increased in advance to any even value without changing any other parameter to select a desired balance of accuracy and CPU performance. The algorithm could also be modified to allow the population size to vary depending on iteration. For the purposes of this paper, a fixed population size of 512 is used.

Central to a genetic algorithm is how the parameter space is explored. For this algorithm, the overall translation is represented by the x, y, and z coordinates of the root atom. Translations of the root atom position are applied equally to all atoms of the ligand. For the initial set of poses, root atom coordinates are assigned using a random shuffling of the Sobol sequence [24] in three dimensions.

During a mutation step, the root atom coordinate of each pose has a 30% chance of mutation. To perform a mutation operation, a new root atom position is created by adding a random offset selected within a uniform sphere of 1 Å radius. If the new root atom position would be outside the 10 Å restriction, a different random offset is selected until a suitable position is established. All ligand atoms are then translated accordingly.

During a crossover step, the root atom coordinate of each crossover pairing has a 30% chance of crossover. To perform a crossover operation between two poses, the root atom coordinate of the poses are interchanged. The remaining ligand atom positions are translated accordingly.

Overall rotation with respect to the input coordinates is represented by three Euler angles, which can be divided into two parts: the location of the input z axis on a sphere (α,β) and a rotation around that axis (γ). Operations that are applied to the frame are also applied to all ligand atoms as equivalent rotations about the root atom position. The angles (α, β, and γ) are stored along with the atom coordinates of the pose in order to make rotational transformations numerically explicit.

For the initial set of poses, a uniform distribution on a sphere is generated from the Hammersley distribution [25]. An overall random rotational transform is then applied. A random shuffling of the results is used to define a set of α and β values. A random shuffling of a uniform distribution between 0 and π is then selected for values of γ.

During a mutation step, the rotational frame of each pose has a 30% chance of mutation. To perform a mutation operation, a random point on a sphere is selected in order to define an axis of rotation. A random rotation angle is then selected from a uniform distribution between $\pm\pi/4$.

During a crossover step, the α and β values of each crossover pairing has a 30% chance of crossover, in which cases the corresponding angles of the respective rotational frames are interchanged. In addition, the γ value of each crossover pairing has an independent 30% chance of being swapped.

The torsional freedom of each rotatable bond is defined by an angle between $\pm\pi$ (symmetries are ignored). The value of this angle for a specific pose is unambiguously defined by the dihedral formed from the coordinates of the associated set of neighboring atoms. For the initial set of poses, a random shuffling of a uniform distribution of angles is used combined with a single, random offset for each bond.

During a mutation step, each rotatable bond angle has an independent 20% chance of mutation. To perform a mutation operation, a new dihedral angle is selected from a uniform distribution between $\pm\pi$. During a crossover step, each rotatable bond angle has an independent 20% chance of crossover, in which case the dihedral angles of the two poses are interchanged.

Each iteration of the genetic algorithm begins with a set of representative poses, each with a corresponding enthalpy value. These enthalpy values are either established from the previous iteration or are recalculated as necessary (to account for annealing). A random sampling is applied to select poses for mutation or crossover using a selection probability (or "fitness") based on the rank order in enthalpy. Individual poses may be selected more than once. The relative selection probability is 1.1 for the top ranked pose, 0.9 for the lowest ranked pose, and otherwise linearly interpolated between 1.1 and 0.9 depending on rank order.

For each iteration, a random selection between mutation or crossover is performed, at equal probabilities. If mutation is selected, a list of poses equal in size to the population is created by copying poses selected using the selection criteria described above. A mutation operation is then applied to each of these poses to create a new pose configuration. Since the probability of mutation for individual parameters (translation, rotation, and dihedrals) is less than unity, there is a chance that no change in configuration is generated, in which case the mutation operation is repeatedly applied until a change is generated.

If crossover is selected for an iteration step, a list of pose pairs are created with a size equal to half the size of the population. Each pose in each pair is a copy selected according to the selection criteria described above. If the two poses in a pair are a copy of the same pose, the selection criteria is repeatedly applied until the two poses in each pair are different. The crossover operation is then applied to each pair. Since the probability of crossover for each individual parameter (translation, rotation, and dihedrals) is less than unity, there is a chance that no change is generated, in which case the crossover operation is repeatedly applied until a change is generated.

For each iteration, after the mutation or crossover operations, there exists two sets of poses of equal size. The first set are the poses that existed at the start of the iteration (the "parents"), and the second set are the results of mutation or crossover

(the "children"). The next step of the iteration is to use a niching process to select which poses to use as input to the next iteration from a combination of these two sets.

The niching process uses a two-level clustering scheme. The top level clustering is a type of k-means clustering, where the list of parents and children are divided into k clusters. The number k varies from 1 to 10 depending on the ligand size and iteration number, as explained below. The clustering is performed by first selecting k random poses as initial cluster representatives. Clusters are then formed by assigning each remaining pose to the representative which produces the smallest RMSD. New representatives of each cluster are chosen by identifying the pose with the smallest sum of RMSD to all other cluster members. Clusters are reformed and the process is repeated for at most 20 iterations or until convergence.

To establish the value for k, each ligand is assigned to one of six categories depending on the number of heavy atoms, as indicated by Table 1. The value of k at the first iteration is assigned the value as shown in this table. At given intervals of later iterations, the value for k is decremented by one. For example, for ligands with between 31 and 40 heavy atoms (category "m"), the value of k starts at 4, decreases to 3 at iteration 400, and then decreases further to 2 at iteration 800.

Table 1 Ligand size categories and associated algorithm parameters.

	xs	s	m	l	xl	xxl
Number heavy atoms	1–20	21–30	31–40	41–50	50–60	> 60
Initial k	1	2	4	6	8	10
Iteration to decrement k	—	500	400	400	300	300
Initial subcluster threshold h_2^0 (Å)	2.0	2.0	3.0	3.0	4.0	4.0

A form of hierarchical clustering is applied independently to the contents of each top level cluster. This second level clustering is performed by considering each subject pose in order of increasing enthalpy. If a pose is within a given threshold h_2 of the RMSD of the representative of any existing cluster, it is added to that cluster. Otherwise it forms the representative of a new cluster. Since the algorithm begins with an empty set, the lowest enthalpy pose becomes the representative of the first cluster. The value of h_2 is an annealed quantity that depends on an initial value h_2^0 and the current iteration:

$$h_2 = h_2^0 \left(\frac{1}{3}\right)^{i_{10}/100}, \tag{12}$$

where i_{10} is incremented every 10th iteration, in the same fashion as used for the pairwise potential (Eq. 8). The value of h_2^0 varies between 2 and 4Å, depending on the ligand size, as show in Table 1.

To select the poses to save for the next iteration, the second level clusters belonging to all top level clusters are collected and sorted by the energies of their representative poses. The clusters are then visited in sorted order and the best enthalpy pose (which is the representative pose) is removed until enough poses are extracted. If additional poses are required, the clusters are visited again in the same order and the next best enthalpy pose (if any remain) is removed from each cluster. This process is repeated as necessary.

After the last iteration of the genetic algorithm, a separate clustering analysis is performed to establish the final output. This clustering is performed on a subset of the final set of poses using another type of k-means clustering. The subset is chosen from either the top 15% of poses based on enthalpy, or the set of poses with enthalpy within 24 kcal/mol of the best enthalpy pose, whichever is larger.

To perform the clustering, 20 cluster seeds are chosen using a greedy algorithm. This greedy algorithm iterates through the subset of poses in order of enthalpy, selecting any pose that is beyond a given RMSD threshold from any previously selected pose. This procedure is repeated using incrementally adjusted RMSD threshold values until the desired number of cluster seeds are generated.

Poses are assigned to the cluster with a representative pose that is closest in RMSD. The representative pose for each cluster is the seed for the initial attempt, and otherwise the member of the cluster with the best enthalpy after all poses have been assigned. This process is repeated for 20 iterations or until convergence. The representatives of the resulting 20 clusters are then selected as the 20 final output poses of the algorithm.

For the purposes of reporting the enthalpy of resulting poses, the quantity ΔE^* is introduced. This quantity is the difference in the modified MMFF94 enthalpy E between a pose in the complex and the minimum enthalpy E_ℓ of the ligand in isolation:

$$\Delta E^* = E - E_\ell \; . \tag{13}$$

The value of E_ℓ is taken from the minimum enthalpy established by running the genetic algorithm without the protein.

As described, this genetic algorithm requires the enthalpy evaluation of approximation 550,000 different ligand states. The cost of the enthalpy evaluation, and thus the execution speed of the algorithm, will depend on the implementation. Molecular dynamic simulations can typically perform similar enthalpy evaluations of a single state involving tens of thousands of atoms in less than a millisecond when provided with appropriate computational resources [26,27]. Although the optimizations available in molecular dynamics are not all available in a genetic algorithm (where successive ligand poses can make large jumps in atom coordinates), nevertheless it should be possible to keep execution times to within a few minutes.

5 Example use cases

Presented in Table 2 are 29 PDBs systems, 28 of which were selected for their diversity and difficulty and one (1stp) selected because it is a well-known example of an easy to solve system. These systems, with ligands that range in molecular weight from 243 to 758, are a subset of the approximately 100 systems used to originally develop the algorithm, selected to be representative of typical protein targets in drug discovery.

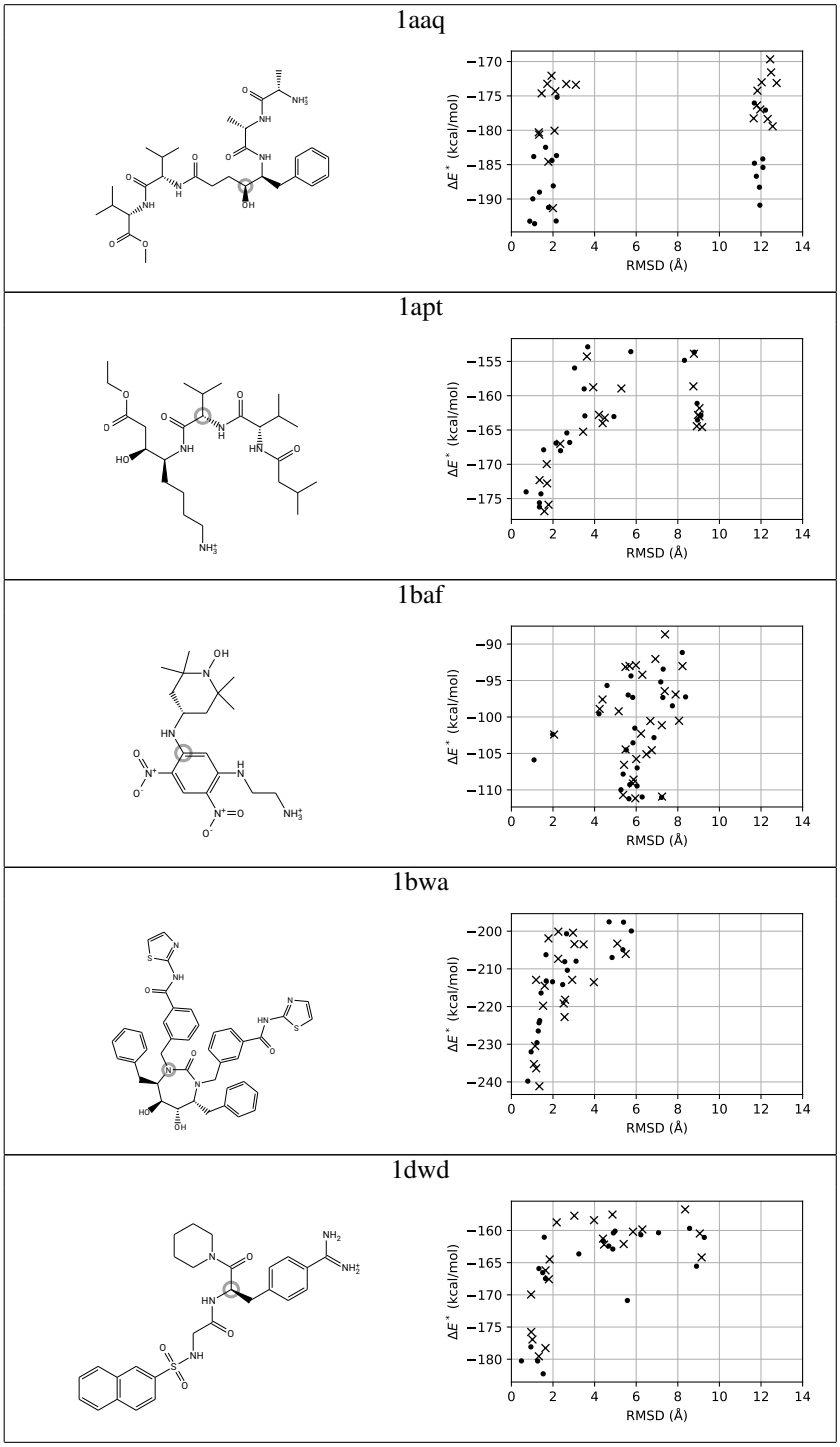
Since the genetic algorithm described in this paper is stochastic, each of the 29 systems were processed a total of five separate times (with different random seeds). The best resulting RMSD from each trial run is included in Table 2. For all systems, a pose with RMSD below 2 Å is reconstructed in a majority of attempts.

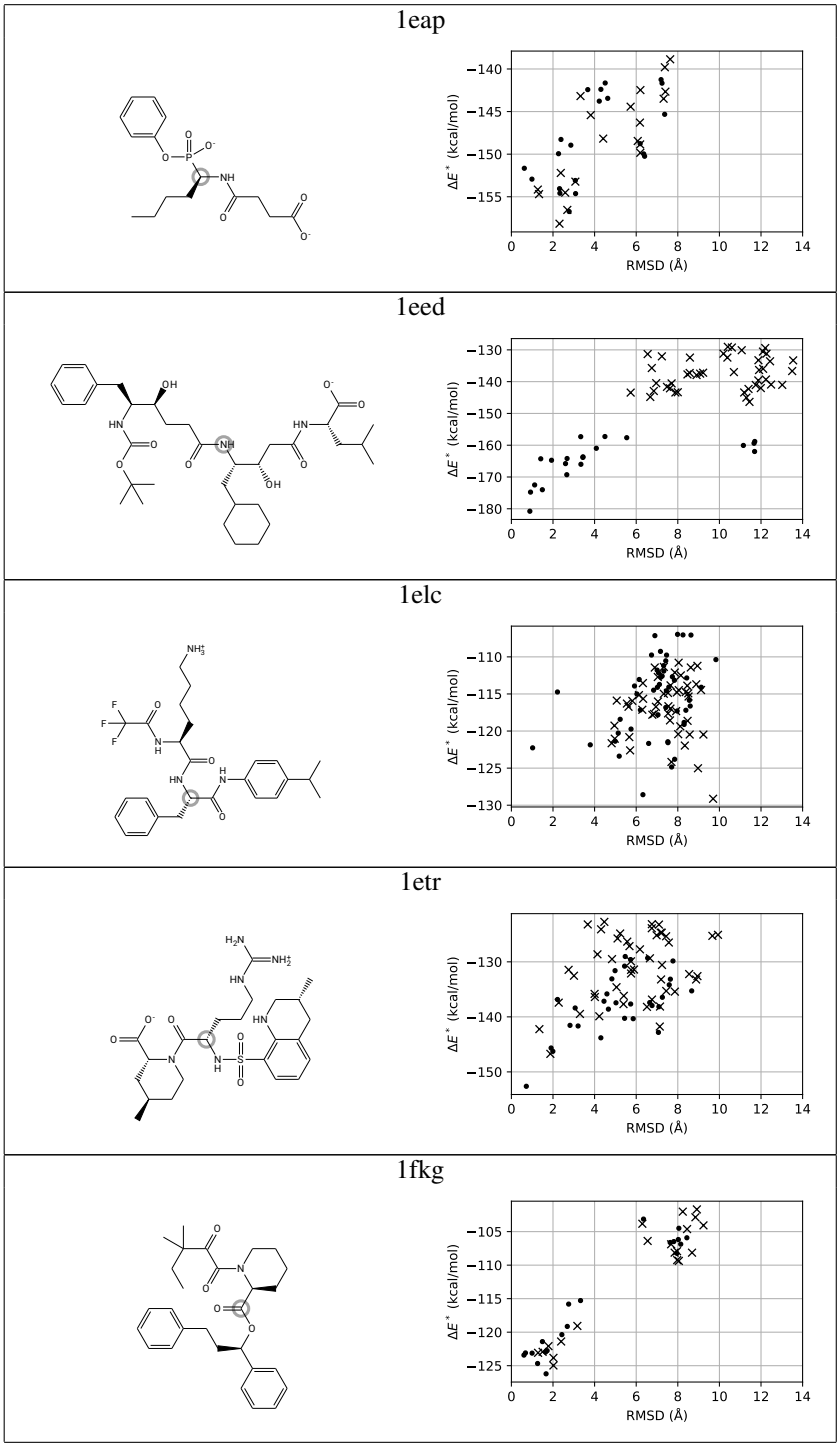
Table 2 Example systems selected from the PDB and the corresponding output from the genetic algorithm. Shown for each system ligand is the molecular weight (mwt), number of rotatable bonds (N_{rot}) and overall charge (Q) for the assumed protonation state (in solution). Also shown is the best RMSD obtained from each of five independent trial runs of the genetic algorithm. RMSD values greater than two are identified with an * symbol.

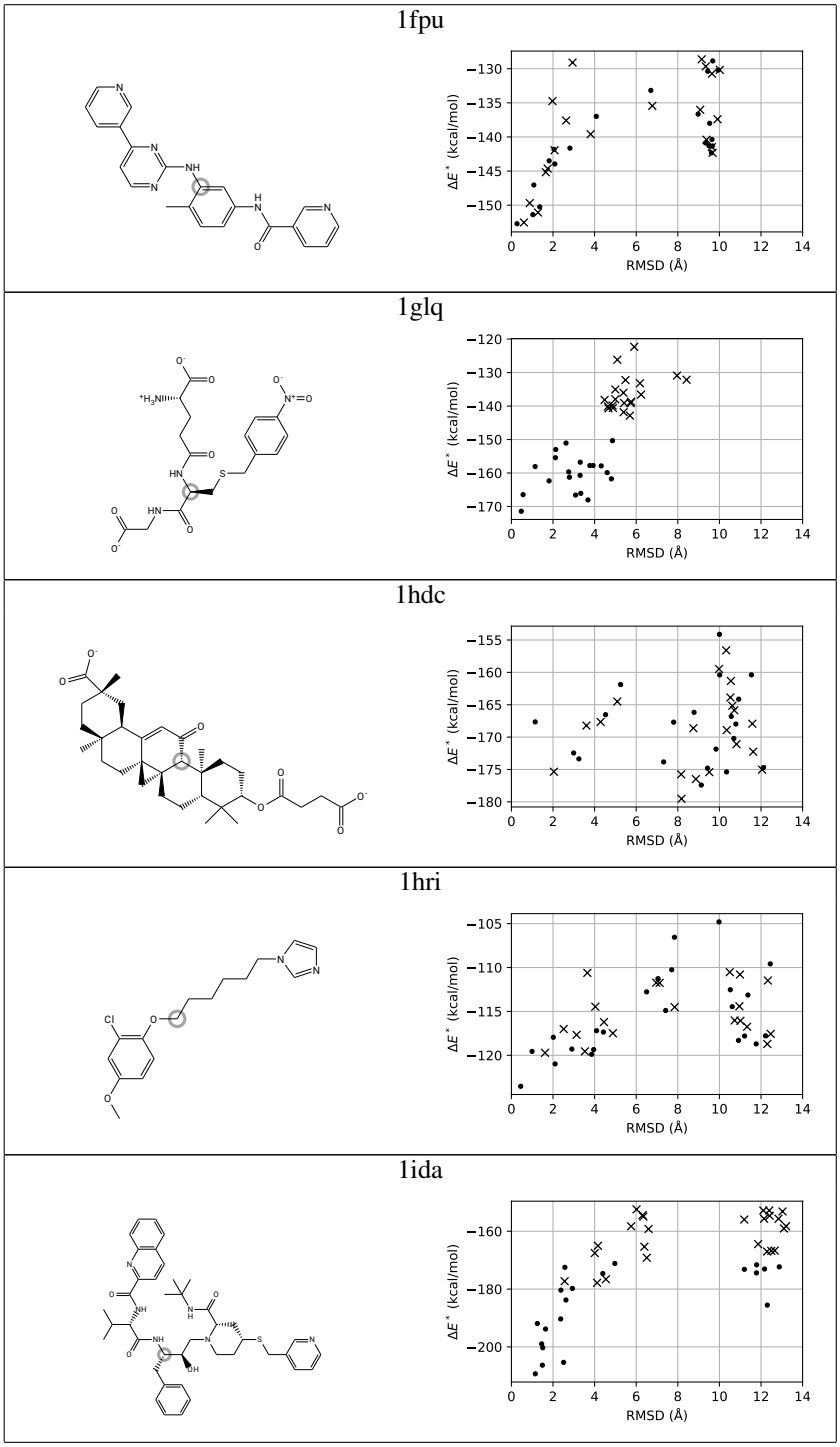
PDB	mwt	N_{rot}	Q	Best RMSD (Å)				
1aaq	579	30	1	1.2	1.0	0.9	1.3	1.2
1apt	502	30	1	1.4	1.2	0.7	0.9	0.9
1baf	397	13	1	1.1	2.1*	1.2	1.2	1.1
1bwa	759	16	0	0.9	1.1	0.9	1.0	0.8
1dwd	523	13	1	0.8	0.9	0.6	0.5	0.8
1eap	341	12	−2	0.8	0.8	0.6	1.3	0.9
1eed	633	29	−1	0.9	1.2	5.7*	1.2	1.1
1elc	508	19	1	1.7	3.9*	1.0	4.8*	1.3
1etr	509	15	0	0.9	0.7	1.3	1.1	1.1
1fkg	450	14	0	0.9	0.6	0.9	1.3	0.9
1fpu	382	7	0	0.3	0.3	0.6	0.6	0.6
1glq	441	16	−1	4.5*	0.6	0.5	0.5	4.1*
1hdc	569	13	−2	1.4	2.0*	1.1	1.3	1.5
1hri	309	10	0	1.3	0.7	1.0	0.4	1.6
1ida	725	24	0	1.1	1.2	1.4	2.6*	1.2
1iep	495	10	1	0.6	0.9	0.7	0.7	0.7
1lic	305	16	−1	1.7	0.9	1.5	2.4*	1.7
1poc	489	26	−1	1.1	1.2	1.2	1.2	1.4
1rne	730	31	0	1.6	1.3	1.3	1.4	0.9
1stp	243	5	−1	0.5	0.4	0.4	0.5	0.4
1udt	475	11	0	0.8	1.1	1.1	1.3	0.8
1vzq	449	7	1	0.5	0.5	0.4	0.6	0.6
2dbl	418	9	−1	0.6	0.7	0.5	0.6	0.7
2r07	349	9	0	0.4	0.4	0.4	1.2	0.3
2uwl	448	9	0	0.8	0.7	0.5	0.6	0.7
4dfr	452	13	−2	2.0	2.2*	1.3	1.3	1.9
4phv	619	17	0	1.2	1.1	1.9	1.6	0.9
5p2p	423	24	−1	0.8	1.0	2.1*	0.9	3.9*
6cpa	476	16	−2	0.5	0.6	0.5	0.7	0.8

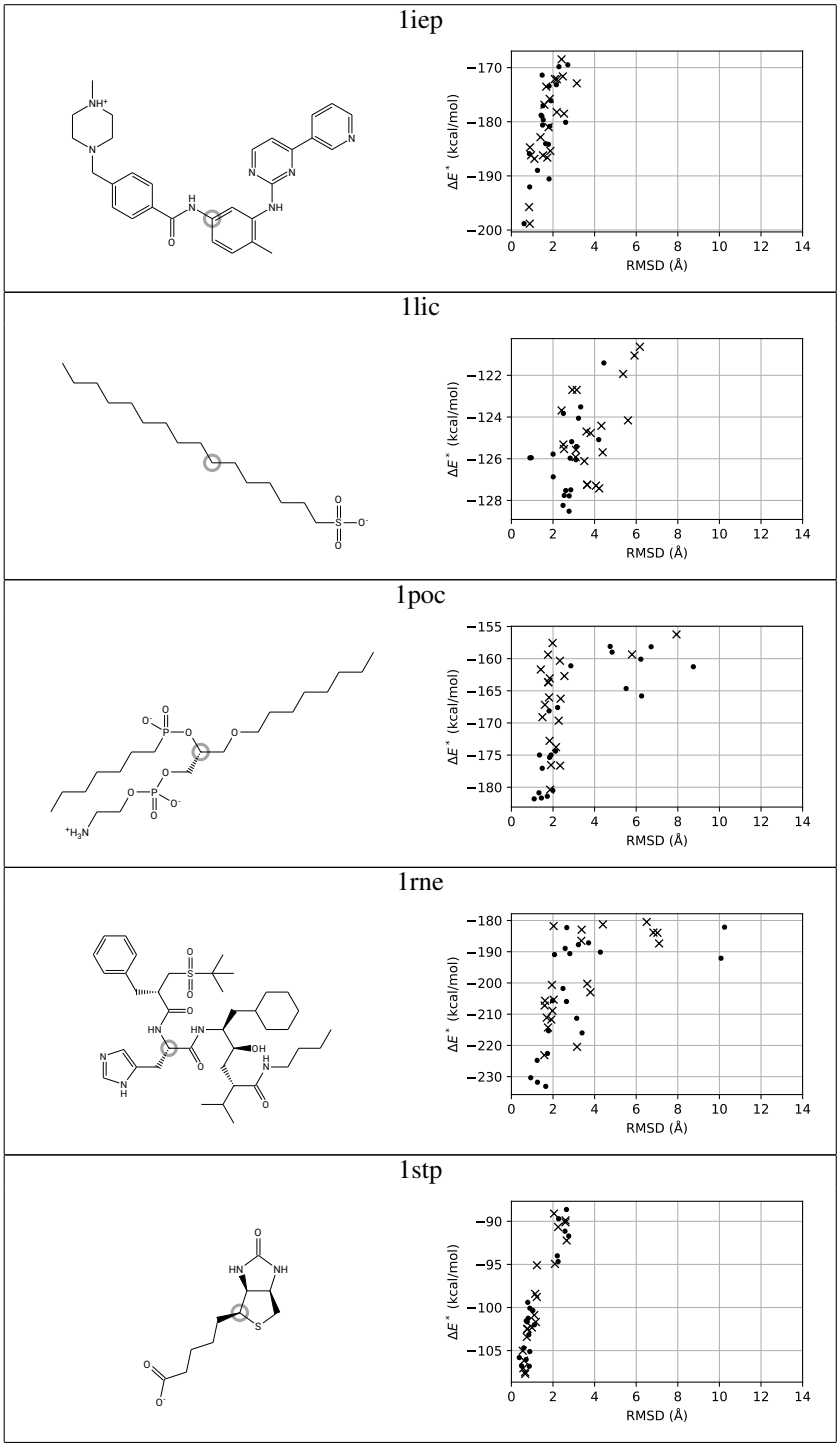
Shown in Table 3 are the ligand structures and plots of RMSD versus ΔE^* for the solution poses from the best (lowest minimum RMSD) and worst (largest minimum RMSD) performing runs. The range of ΔE^* varies depending on the system but is generally quite large, reaching as high as 230 kcal/mol in some cases. The large range of ΔE^* is evidence that the enthalpy calculated from the modified MMFF94 force field is not physical. The enthalpy nevertheless serves as a useful tool for establishing a good approximation of the ligand pose.

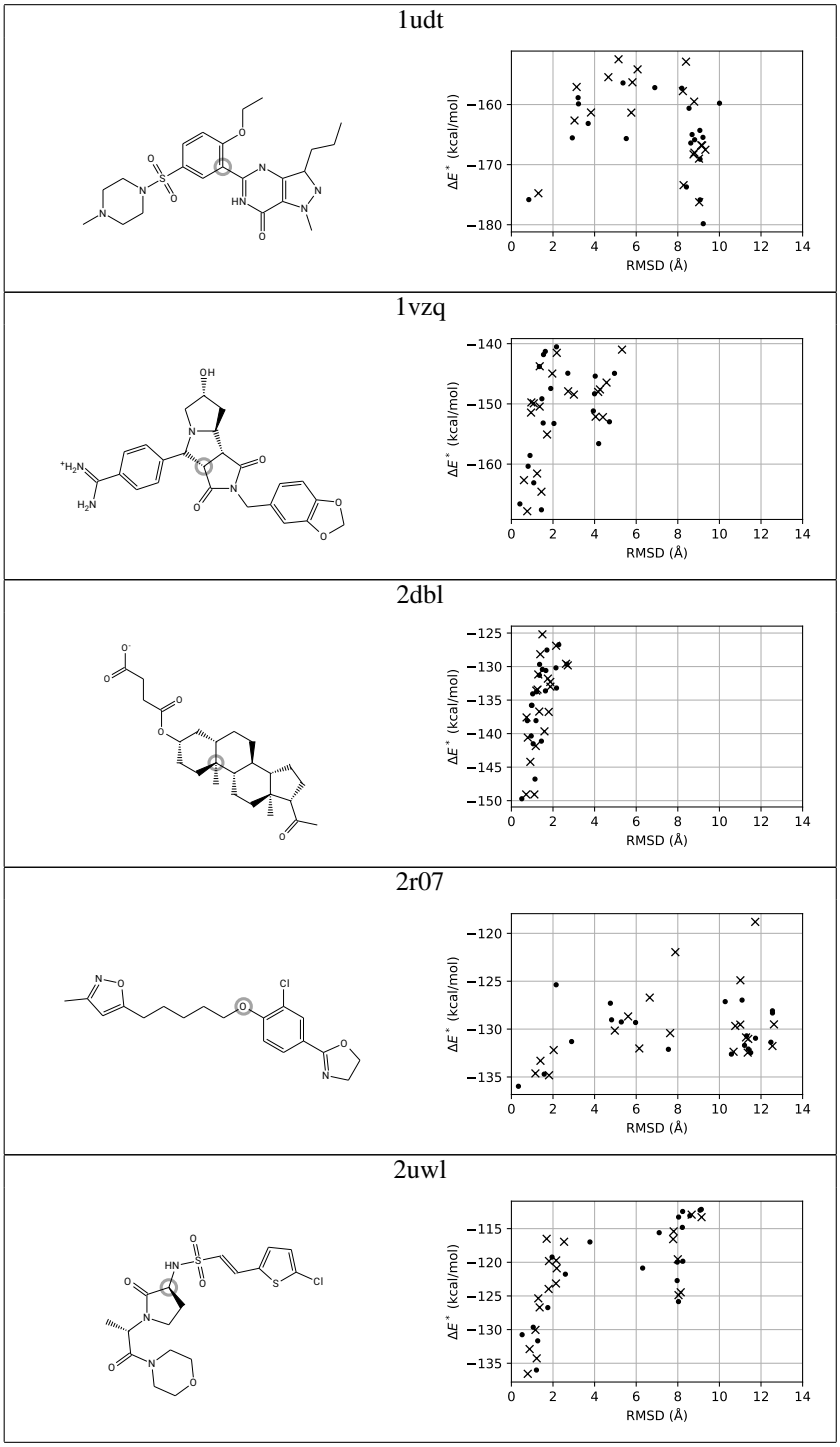
Table 3 Example systems selected from the PDB and the corresponding output from the genetic algorithm. Shown at left are the ligand structures, with the root atom indicated by a grey circle. On the right are plots of RMSD (with respect to the experimental structure) versus ΔE^* for the twenty output solution poses of each system from two trials of the genetic algorithm. The two trials shown are taken from the five generated for this paper and include (dots) the trial that produced the best RMSD value and the trial (crosses) whose best RMSD value was the worse of the five.

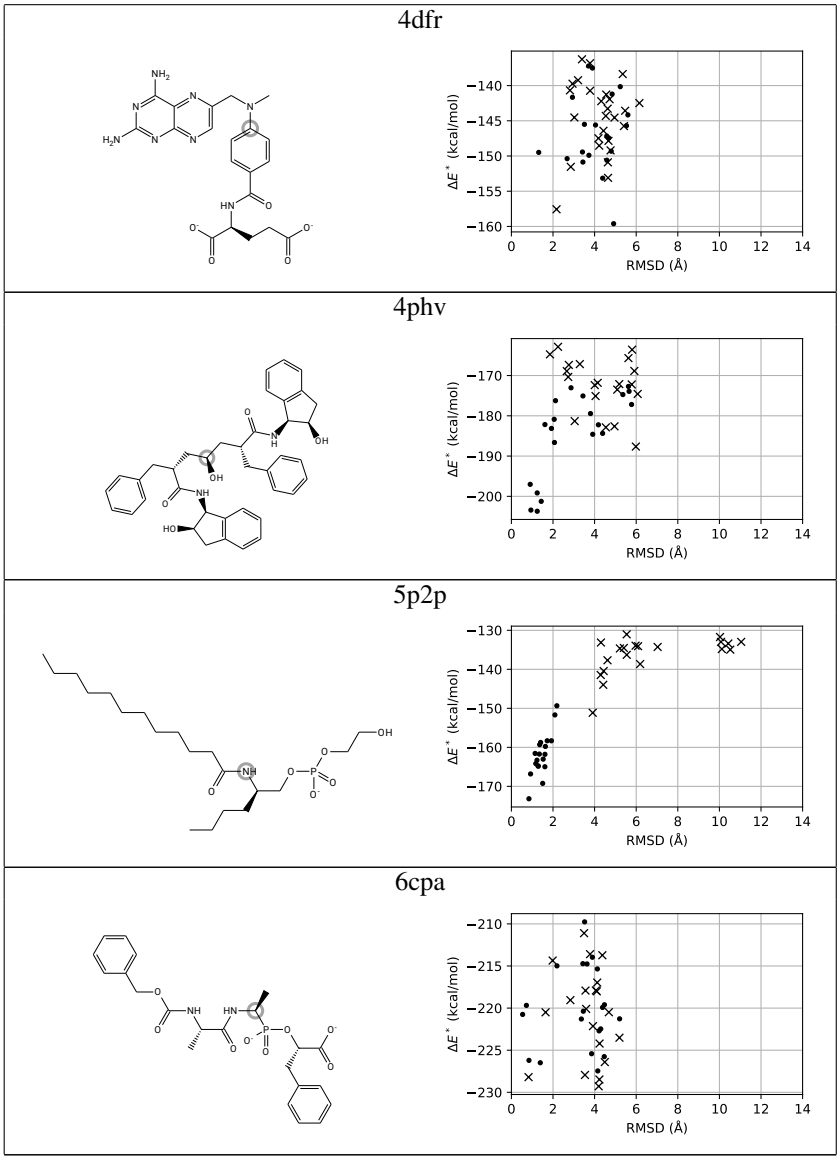

1aaq
1apt
1baf
1bwa
1dwd
ΔE* (kcal/mol)
RMSD (Å)


1eap
ΔE* (kcal/mol)
RMSD (Å)
1eed
ΔE* (kcal/mol)
RMSD (Å)
1elc
ΔE* (kcal/mol)
RMSD (Å)
1etr
ΔE* (kcal/mol)
RMSD (Å)
1fkg
ΔE* (kcal/mol)
RMSD (Å)


1fpu
1glq
1hdc
1hri
1ida
ΔE* (kcal/mol)
RMSD (Å)


1iep
1lic
1poc
1rne
1stp
ΔE* (kcal/mol)
RMSD (Å)


1udt
ΔE* (kcal/mol)
RMSD (Å)
1vzq
ΔE* (kcal/mol)
RMSD (Å)
2dbl
ΔE* (kcal/mol)
RMSD (Å)
2r07
ΔE* (kcal/mol)
RMSD (Å)
2uwl
ΔE* (kcal/mol)
RMSD (Å)



The success of the genetic algorithm is due in part to the identification of a large number of diverse poses. The amount of diversity is demonstrated by the data in Table 4. In the majority of systems, the pose with the best enthalpy is not the pose closest to the experimental result. In such cases, the difference in enthalpy can be as large as 10 kcal/mol.

It should be emphasized that RMSD is not a transitive property. For example, two poses may have similar RMSD with respect to a third pose, but may be even further apart in RMSD between each other. This non-transitive nature is a natural

Table 4 Details from a trial run of the genetic algorithm. The data in this table was extracted for each system from the run in Table 2 that produced the smallest RMSD with respect to experiment. Included is δE, the difference in calculated enthalpy between the pose with the smallest RMSD and the pose with the best enthalpy (in those systems where they are different). One measure of pose diversity is the RMSD between different solution poses (the relative RMSD). Shown is the minimum, median, and maximum relative RMSD between all pairs of the 20 solution poses.

	δE	Relative RMSD (Å)		
PDB	(kcal/mol)	min	median	max
1aaq	0.4	1.1	11.2	12.4
1apt	2.2	1.3	5.3	9.7
1baf	5.3	1.0	5.4	8.7
1bwa	—	1.5	11.0	12.4
1dwd	2.0	1.3	5.1	9.8
1eap	5.1	1.1	5.1	7.4
1eed	—	1.3	4.4	12.2
1elc	6.3	1.7	7.3	10.8
1etr	—	1.3	6.5	9.9
1fkg	2.8	1.1	5.9	9.1
1fpu	—	1.0	8.3	10.0
1glq	—	1.2	3.3	5.6
1hdc	9.8	1.2	10.0	12.5
1hri	—	1.2	9.0	12.6
1ida	—	1.3	5.4	13.5
1iep	—	1.1	2.3	4.1
1lic	2.6	0.9	2.5	4.5
1poc	—	1.1	4.9	10.5
1rne	2.8	1.5	3.4	11.1
1stp	1.0	0.7	1.6	3.2
1udt	4.0	1.7	6.0	10.0
1vzq	1.0	1.0	2.7	5.2
2dbl	—	1.0	1.9	3.6
2r07	—	1.4	9.8	12.8
2uwl	5.3	1.2	7.8	10.2
4dfr	10.1	2.1	4.0	6.1
4phv	6.7	1.4	9.5	11.3
5p2p	—	1.0	1.7	3.2
6cpa	6.7	1.1	4.2	8.5

consequence of projecting the large, high-dimensional configuration space onto a single value. Included in Table 4 is the minimum, median, and maximum relative RMSD between solution poses for each system. Median values range as high as 11 Å.

The distribution of relative RMSD is strongly influenced by the system, as shown in Figure. 2. The system represented by 1aaq, for example, tends to favor two groups of solutions, whereas 1apt favors a more disjointed solution space.

The characteristics of each test system is described individually in more detail in the following.

1aaq. This system is a five-peptide analog bound to the HIV-1 protease [28], a relatively small protein (and the same protein target as 1bwa and 1ida). The ligand tends to bind along a fairly linear binding pocket that extends through the length of the protein. The genetic algorithm tends to group pose solutions into two nearly degenerate groups, each corresponding to the ligand bound in opposite directions

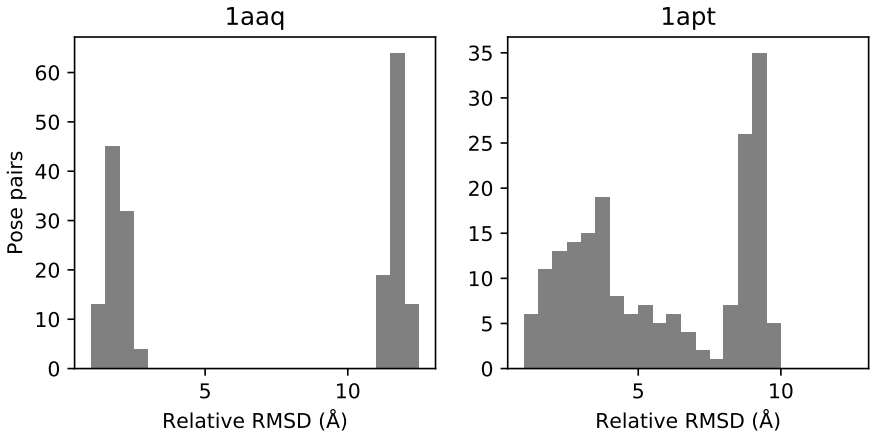


Fig. 2 Distribution of RMSD between all pair combinations of the twenty solution poses from a single trial of the systems 1aaq (left) and 1apt (right).

(Fig. 3). Given that the protein is a symmetric dimer, this degeneracy may not be surprising.

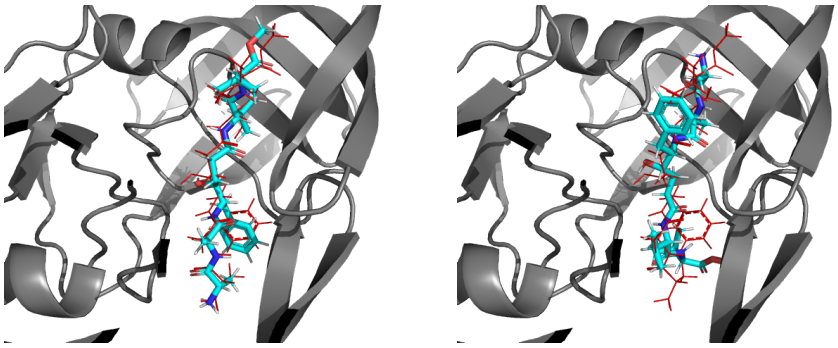


Fig. 3 Degenerate solutions of system 1aaq. The experimental pose is shown in red lines. Solution poses are shown in colored licorice. Shown on left is a solution pose in the correction orientation with an RMSD of 0.73 Å. Shown on right is a typical solution with the ligand reversed.

1apt. This system includes a four-peptide pepstatin analogue bound to penicillopepsin [29]. A single glycosylation included with the PDB structure was removed before protein processing. The binding pocket is a well-defined groove and the ligand is tightly coordinated by up to nine hydrogen bonds. The correct pose is consistently identified.

1baf. This system is a hapten bound to a monoclonal antibody [30]. The ligand includes a dinitrophenyl core. The protein was prepared in the dimer form, as presented in the PDB. The lack of strong, favorable electrostatic contacts may be one reason that the genetic algorithm struggles to isolate the experimental pose. Another

difficulty is that an alternate pose, flipped 180 degrees from the experimental result (Fig. 4), has more favorable calculated enthalpy.

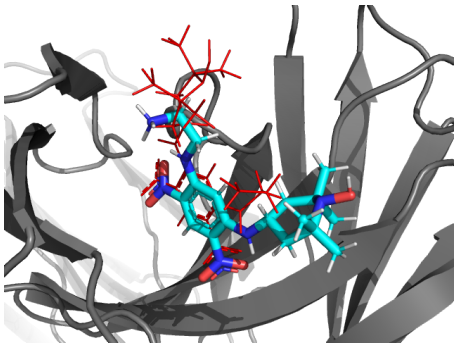


Fig. 4 Dominate solution for system 1baf. The solution pose (colored licorice) that is rotated 180 degrees from experiment (red lines) has more favorable enthalpy and tends to dominate the optimization solution.

1bwa. This system is the inhibitor xv638 from Dupont Pharmaceuticals bound to the HIV-1 protease (the same protein target as 1aaq and 1ida) [31]. This inhibitor is large and symmetric, with a seven member aliphatic central ring. Like 1aaq, the ligand has degenerate poses, but in this case, the degeneracy corresponds to a ligand symmetry and poses no difficulty for the algorithm.

1dwd. This system is NAPAP bound to thrombin (the same protein target as 1etr and 1vzq) [32]. Missing calcium and sodium ions were added based on their location in system 1vzq. This inhibitor contains a benzamidine functional group that forms a hydrogen-bond complex with Asp189. The correct pose is consistently identified.

1eap. This system is a hapten bound to a Cyclophilin-E Antibody [33]. The complex is characterized by strong electrostatic interactions between the central phosphate group of the ligand and Arg96 on chain A and Lys97 on chain B. In additional to the experimental result, the algorithm establishes multiple ligand poses that preserve this interaction (see Fig. 5 for an example).

1eed. This system is a five-peptide analog bound to endothiapepsin [34]. The large ligand has favorable enthalpy in its bound form due to the coordination of multiple hydrogen bonds. Despite this, the size of the ligand and the wide binding pocket of the protein are a particular challenge for the genetic algorithm, and one of the five trials fails to produce the experimental pose, nor any pose within 30 kcal/mol in calculated enthalpy of the correct pose (see Table 3).

1elc. This system is a flexible psuedo-peptide bound to an elastase [35]. The binding pocket in this protein is fairly open to the surface, with multiple channels for binding. Interestingly, this PDB structure was published as part of a study in which the authors describe how similar ligands can bind quite differently to a protein. As such, it is not surprising that the genetic algorithm struggles to locate the experimental pose.

1etr. This system consists of a close chiral partner of argatroban bound to bovine α-thrombin (the same target protein as 1etr and 1vzq) [36]. This ligand contains an

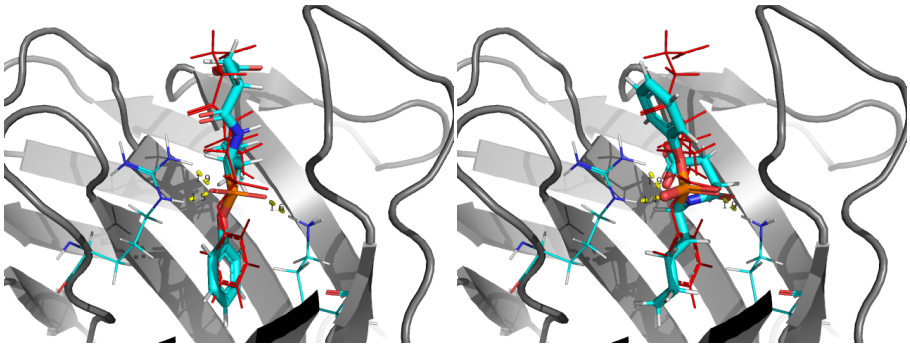


Fig. 5 One of several alternate solutions of 1eap. The experimental pose is shown in red lines. Solution poses are shown in colored licorice. Shown on left is a solution pose at the correction orientation with an RMSD of 0.83 Å. Shown on right is an alternate solution in a reversed ligand orientation that preserves the phosphate charged interactions with the protein.

arginine group that mimics the nature substrate and forms a tight hydrogen-bond complex with Asp189 (similar to system 1dwd). The experimental pose is consistently reproduced.

1fkg. This system is a simple ligand with a piperidine core bound to isomerase FKBP1A [37]. The experimental pose is consistently reproduced with favorable enthalpy. An alternate solution with considerably worse (15 kcal/mol) calculated enthalpy is also consistently predicted.

1fpu. This system is imatinib (Gleevec) bound to Abelson Tyrosine Kinase [38]. Only the first of the pair of solutions contained in the PDB file is used. Two main groups of solutions are resolved by the algorithm. The experimental pose has a favorable enthalpy and is consistently reproduced.

1glq. This system is a glutathione-based ligand bound to glutathione S-transferase P 1 [39]. It is assumed that this protein occurs naturally as the same dimer system resolved in the PDB structure. The first ligand of the two copies found in the PDB structure is used as the basis for the test.

The binding pockets on 1glq are fairly open, making this system particularly challenging. Two of the five trials fails to identify the experimental structure and instead produce a cluster of solutions around a pose of inferior calculated enthalpy. This alternate pose conflicts with two explicit water molecules resolved in the PDB structure (Fig. 6) and would presumably be avoided if either or both of these water molecules was included in the model.

1hdc. This system is a steroidal inhibitor bound to a bacterial hydroxysteroid dehydrogenase [40]. It is assumed that the tetramer structure resolved in the PDB is the biologically relevant tertiary form. The first of the four ligands in the PDB is used for the test. The experimental pose forms inside a U-shaped binding pocket that is exposed at both ends to solution. A lack of strong electrostatic interactions or hydrogen bonds means that the pose is defined primarily by Van dar Waals forces. As such, it is not surprising that several alternate poses are identified (Fig. 7).

1hri. This system is an inhibitor bound to human rhinovirus 14 [41]. The relatively light ligand consists of two aromatic ring systems connected by an alkane

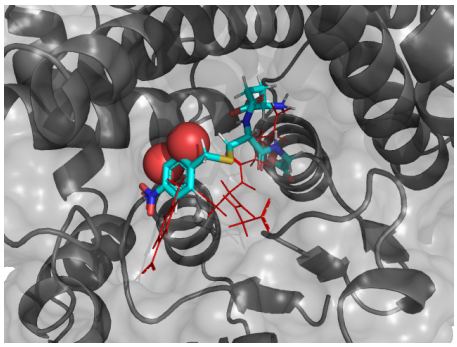


Fig. 6 Alternate solution for system 1glq. This solution pose (colored licorice) is displaced from experiment (red lines) and dominates in two of the five trial runs of the genetic algorithm despite a poorer final calculated enthalpy. This solution pose clashes with two waters resolved in the PDB experimental structure and would likely have been excluded if either of these waters were included in the model.

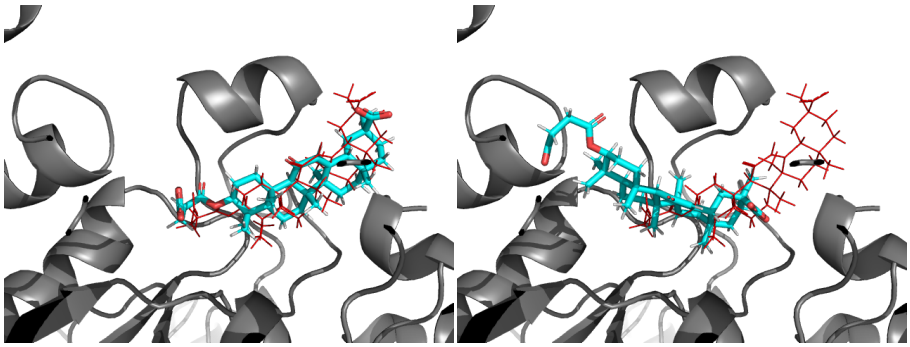


Fig. 7 Solutions of 1hdc. The experimental pose is shown in red lines. Solution poses are shown in colored licorice. Shown on left is a solution pose near the correction orientation with an RMSD of 1.3 Å. Shown on right is an alternate solution shifted along the U-shaped binding pocket.

chain. The long and narrow binding pocket in the PDB structure is nearly inaccessible from solution. The experimental pose is consistently identified.

1ida. This system is a large (725 MW) ligand bound to the HIV-1 protease (the same protein target as 1aaq and 1bwa) [42]. Interestingly enough, despite the symmetric nature of the protein, the experimental orientation of the ligand dominates the solution, which suggests a structural bias in the PDB protein, perhaps from experimental artifacts in crystalization [42].

1iep. This system is imatinib bound to murine Tyrosine-protein kinase ABL1 [43]. The experimental structure includes two copies of the complex. Only the first copy was prepared for this test. The two relevant chlorine hetero atoms were included, which appear close to the bound ligand. The experimental pose is consistently identified with little ambiguity.

1lic. This system is palmitic acid bound to murine fatty acid-binding protein [44]. The oxidized cysteine Osc117, in contact with the ligand, was changed to cysteine for force field compatibility during protein processing. Although this ligand is not drug

like, it serves as an interesting test of optimizing a long alkane chain. An alternate pose (RMSD 2.7 Å, Fig. 8) positions the end of the alkane chain towards a different part of the binding pocket. This alternate pose tends to have more favorable Van dar Waals contacts and dominates the solution.

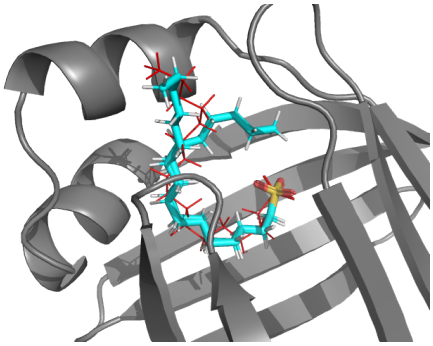


Fig. 8 Example solutions for system 1lic. Two solution poses (colored licorice) are superimposed on the experiment result (red lines).

1poc. This system is a transition-state analog bound to bee-venom phospholipase A2 [45]. The phosphates of the ligand bind in close coordination with a calcium ion, included in the model. In spite of multiple alkane chains, the experimental pose is consistently reproduced.

1rne. This system is a relatively heavy and flexible inhibitor bound to human renin [46]. Protein processing included the repair of a broken loop between Arg157 and Phe161 and the deglycation of Asn67. Neither of these modifications are expected to have any significant affect on ligand binding. The experimental pose is coordinated through multiple hydrogen bonds and is consistently identified.

1stp. This system is an often-cited experimental solution for biotin bound to streptavidin [47]. It is commonly used as an example for docking algorithms [48] and is also presents an interesting example of high ligand-binding efficiency [49]. The genetic algorithm readily identifies the experimental result and only reluctantly permits one class of alternate solution (Fig. 9).

1udt. This system is sildenafil bound to cGMP-binding cGMP-specific phosphodiesterase [50]. During processing, the loop between Tyr664 and Tyr676 was repaired. The zinc and magnesium ions, both about 6 Å from the bound ligand, were included in the model. The experimental pose is consistently identified. Included as an alternate solution is a pose that contains a close contact with the magnesium ion (Fig. 10). Part of the difficulty with this system could be associated with a MMFF94 parameterization that has inadequate support for metals.

1vzq. This system is an inhibitor bound to thrombin (the same protein target as 1dwd and 1etr) [51]. A broken loop between Trp148 to Glu151 was repaired during processing. The calcium and sodium hetero atoms were included in the model. The inhibitor contains a benzamidine functional group that forms a hydrogen-bond complex with Asp189. The correct pose is consistently identified.

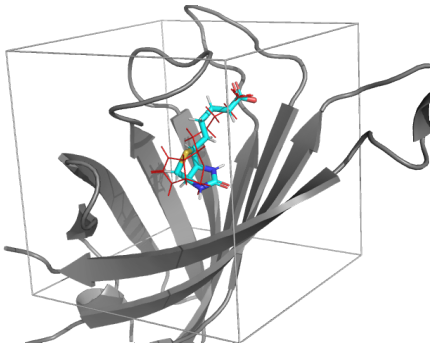


Fig. 9 An alternate solution for system 1stp. The solution pose (colored licorice) is superimposed on the experiment result (red lines). The box is the 10 Å constraint on the root atom position.

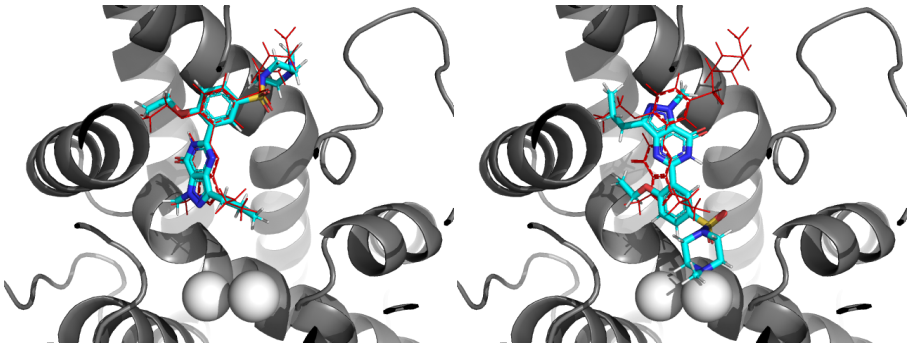


Fig. 10 Solutions of 1udt. The experimental pose is shown in red lines. Solution poses are shown in colored licorice. Shown on left is a solution pose near the correction orientation with an RMSD of 0.8 Å. Shown on right is an alternate solution that includes an unnaturally close contact with a magnesium ion.

2dbl. This system is a steroid bound to an antibody/antigen complex [52]. The experimental pose is consistently identified.

2r07. This system is an antiviral agent bound to human rhinovirus 14 coat protein [53]. This system is notable due to the hydrophobic nature of both the ligand and binding pocket. These properties could explain why the genetic algorithm produces a large variety of solutions (Fig. 11). The experimental pose is reproduced consistently.

2uwl. This system is an inhibitor from GlaxoSmithKline bound to human factor Xa [54]. The experimental pose is consistently identified.

4dfr. This system is methotrexate bound to bacterial dihydrofolate reductase [55]. There are two complexes in the experimental structure. Only the first, along with the associated chlorine hetero atom, is modeled. The experimental pose is associated with several hydrogen bond contacts. The genetic algorithm tends to favor solutions that set the ligand deeper in the binding pocket with more favorable Van dar Waals pair energies.

4phv. This system is a large, psuedo-symmetric ligand bound to HIV Gag-Pol polyprotein [56]. The experimental structure is solved for two closely related isomers

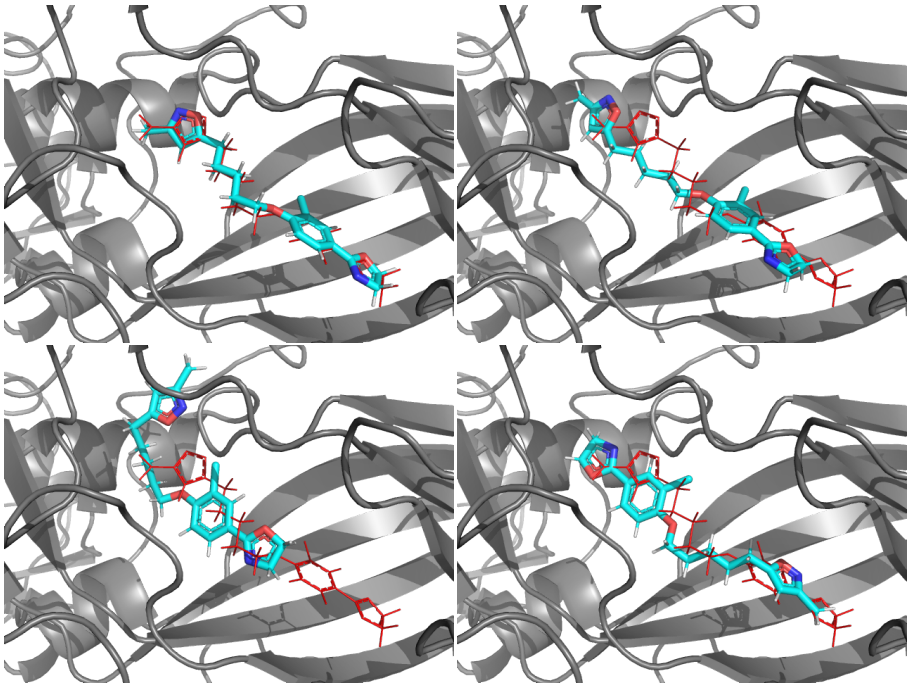


Fig. 11 Solutions of 2r07. The experimental pose is shown in red lines. Solution poses are shown in colored licorice. Shown on top left is a solution pose near the correct configuration with an RMSD of 0.4 Å. Three alternative solutions are also shown, with RMSD varying from 3 to 11 Å. The ligand is rotated 180 degrees in the solution shown in the lower right.

of the ligand. A specific chirality is arbitrarily selected for the purposes of this test. This experimental pose is consistently reproduced.

5p2p. This system is a substrate analog bound to porcine phospholipase A2 [57]. The experimental structure includes two copies of the complex. The first copy is isolated and the associated calcium ion is used in the model. The phosphate makes direct contact with the ion and may be the reason, in combination with the long alkane chain, that this complex is difficult to solve (Fig. 12).

6cpa. This system is a phosphonate ligand bound to bovine carboxypeptidase A1 [58]. The experimental pose is coordinated by a zinc ion, included in the model. One end of the ligand is exposed to solution and introduces some ambiguity. Nevertheless, the experimental pose is consistently reproduced.

6 Limitations and possible extensions

When interpreting the results presented above, it should be recognized that the protein structures used were derived from experimental measurements of the very same ligand-protein complex. As such, the protein structures were already matched to the associated ligands. Protein preparation, which involves relaxation of the ligand-protein complex, further ties the protein state to the ligand.

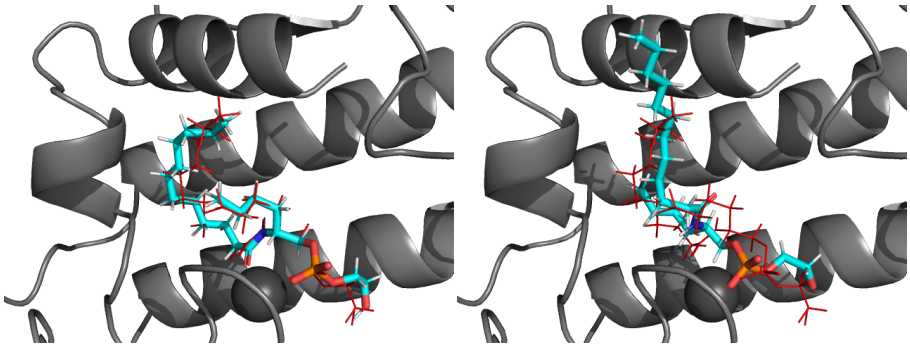


Fig. 12 Solutions of 5p2p. The experimental pose is shown in red lines. Solution poses are shown in colored licorice. Shown on left is a solution pose near the correction orientation with an RMSD of 0.8 Å. The solution on the right has an abnormal ion contact and likely contributes to the failure of the genetic algorithm in two of the five trials.

Clearly, a method for predicting the experimental pose of a ligand-protein complex has limited utility when the structure of the complex has already been established experimentally. In a drug discovery program, it will be necessary to predict the bound pose of novel compounds to a given protein target. As such, perhaps the most detrimental limitation of the genetic algorithm in the form presented here is the assumption of a fixed protein.

As a partial solution to protein flexibility, the genetic algorithm can be applied independently to several variants of a protein structure, at a proportional computational cost. In such a strategy, any scoring model would need to include an appropriate free energy penalty or reward for moving between protein structures. When confronted with multiple independent protein alterations, the cost of this approach grows geometrically.

A more advanced level of protein flexibility could be incorporated directly into the genetic algorithm by including rotameric states of select protein residues in the solution space. Besides substantially increasing the difficulty of the problem, adding protein alterations to the configuration space introduces the additional computational cost associated with calculating pairwise energies within the protein.

A more ambitious form of protein flexibility could incorporate some form of backbone movement in the configuration space, in order to accommodate changes, for example, in exposed protein loops that sometimes play an important role in ligand binding [59]. This would require some way to parameterize the backbone movement. Large scale tertiary changes, such as protein hinges, are also theoretically possible, but are presumably more difficult to parameterize.

In addition to the protein, there are degrees of freedom associated with the ligand that are neglected. The conformal freedom of aliphatic rings is one important example (see, in particular, 1bwa, 1fkg, 1rne, 1vzq and 2uwl). The advantage of using structures based on experimental data is that these conformal freedoms have been established. This is not an option for novel compounds.

There are several techniques that have been developed specifically for ligand conformer generation [60,61]. It is generally recognized that multiple solutions exist for

many aliphatic rings. For completeness, it would be necessary to sample all such solutions within a reasonable enthalpy window. This could be accomplished by multiple executions of the presented genetic algorithm, as long as some method is introduced to apply appropriate penalties in scoring. For ligands with multiple aliphatic rings, the cost of this approach grows geometrically.

Another option is to include distinct ring conformations in the solution space of the optimization algorithm. To do so will require some procedure for switching between ring conformers, with the recognition that some of the other configuration variables, such as root atom position, rotational frame, and torsional angles could be affected. In addition, it would be reasonable to include the complete set of strain terms provided by the MMFF94 force field in order to apply appropriate enthalpy penalties.

In addition to ring conformation, there are other important ligand freedoms that should not be neglected. For example, a bond angle near the center of a ligand can produce significant changes in atom positions at the ends of the ligand. Nitrogen pyramidalization is another important detail that should not be ignored [21].

The compounds in the test systems presented here (Table 3) have a fairly established protonation state. Tautomeric ambiguity, however, from functional groups such as azoles is not unusual in drug-like compounds [62]. Since the effect on enthalpy predictions can be dramatic [63], it can be important to consider all accessible protonation states, both in the complex, and for the ligand in isolation (since protonation can change on binding).

In addition to tautomeric freedoms, protonation of a ligand can vary due to ionization. There are established methods for predicting the ionization state of a compound in solution, but accuracy remains a concern [64]. For microscopic pKa values near biologically relevant pH values, there may be multiple possible ligand states to consider.

Strategies for accounting for multiple protonation states can use the same general approach used for conformational states. For example, sampling various protonation states independently, or applying mechanism for converting between states during optimization. However, it should be recognized that forcefield schemes such as MMFF94 are not designed to account for changes in bond structure [65]. In addition, some mechanism for calculating appropriate penalties for altering ionization states (*i.e.* macroscopic pKa) will need to be considered.

Concerning protonation, it is also important not to neglect tautomeric and ionization states of the protein, which can change during binding [66]. Special attention is required if the protonation state of the protein is closely tied to a particular PDB structure used to define the protein. Methods employed for variable ligand states can also be applied to the protein, with the same limitations noted above.

In addition to the protein protonation state, another important detail to establish when applying this algorithm for novel ligands is the definition of the binding pocket. For the use cases presented in this paper, the experimental ligand pose was used to define the positional constraints on the solution. For novel ligand structures, the constraints defined by a specific reference experimental structure may not be appropriate. Alternately, the protein state could be defined by an apo structure, or the target for

a particular drug campaign could be an unusual location on the protein (*e.g.* for an allosteric inhibitor).

Fortunately, although a typical drug discovery program involves a large number of ligand candidates, the number of different proteins will be limited. There should be no practical reason why a substantial effort could not be invested in preparing each protein target. For example, positional constraints can be manually established if necessary. Computationally expensive methods such as molecular dynamic simulations could also be used as an aid. Alternatively, if experimental structures for multiple ligands are available, some type of consensus algorithm can be applied.

As observed in systems 1poc, 1udt, 5p2p, and 6cpa, catalytic proteins that incorporate metal ions can be challenging. Additional work on forcefield parameterizations that account both for their non-spherical geometry and strength may help these systems converge more accurately [67–69].

Finally, it is important to recognize the limitations of continuum water models of ligand-protein complexes and the importance of water networks in binding [70–72]. The placement of explicit water molecules may be one approach to improving pose prediction accuracy (see, for example, Fig. 6). Including water centers in scoring is also an approach that is gaining some popularity [73]. Dynamic placement of water molecules during pose optimization is a theoretical possibility but will at least require an appropriate enthalpy penalty for water placement and/or removal.

7 Conclusion

Accurate ligand pose prediction is an important aspect of computational approaches to small-molecule drug development and a key element of Verseon's drug-discovery process. Presented in this paper is a genetic algorithm designed to provide multiple pose solutions for a given ligand-protein complex and is capable of reproducing experiment results for systems with large, flexible ligands and high degrees of freedom. The performance of the algorithm has been demonstrated using several challenging use cases.

The algorithm achieves a high level of performance by incorporating several important features including annealing to soften the enthalpy function at early iterations and clustering to preserve diversity. The natural ability to provide multiple solutions for a ligand-protein complex is an important advantage and provides alternate choices for further analysis.

The test cases presented in this paper are designed to demonstrate the effectiveness of the algorithm but do include some simplifications, such as fixed protein, ligand conformation, and protonation state. To take proper advantage of the algorithm in a drug discovery campaign, these deficiencies would need to be addressed. Examples on how to do so are reviewed, along with other possible extensions.

8 Declarations

8.1 Funding

All original research was funded by Verseon Corporation.

8.2 Conflicts of interest/Competing interests

Not applicable

8.3 Availability of data and material

All example cases were performed on publicly available PDB structures.

8.4 Code availability

The computer code used for this study is proprietary. All details necessary for replication of the algorithm are provided.

8.5 Authors' contributions

All authors contributed to the design and refinement of the algorithm. Material preparation, data collection and analysis were performed by David C. Williams. The first draft of the manuscript was written by David C. Williams. All authors have given approval to the final version of the manuscript.

References

1. G.M. Keserű, G.M. Makara, Drug Discovery Today **11**(15-16), 741 (2006)
2. J.P. Hughes, S. Rees, S.B. Kalindjian, K.L. Philpott, British Journal of Pharmacology **162**(6), 1239 (2011)
3. K.H. Bleicher, H.J. Böhm, K. Müller, A.I. Alanine, Nature Reviews Drug Discovery **2**(5), 369 (2003)
4. S. Kar, K. Roy, Expert Opinion on Drug Discovery **8**(3), 245 (2013)
5. D.E. Clark, Expert Opinion on Drug Discovery **3**(8), 841 (2008)
6. A. Lavecchia, C. Di Giovanni, Current Medicinal Chemistry **20**(23), 2839 (2013)
7. M.K. Gilson, H.X. Zhou, Annu. Rev. Biophys. Biomol. Struct. **36**, 21 (2007)
8. B. Waszkowycz, D.E. Clark, E. Gancia, Wiley Interdisciplinary Reviews: Computational Molecular Science **1**(2), 229 (2011)
9. S.E. Boyce, D.L. Mobley, G.J. Rocklin, A.P. Graves, K.A. Dill, B.K. Shoichet, Journal of Molecular Biology **394**(4), 747 (2009)
10. M.K. Gilson, J.A. Given, B.L. Bush, J.A. McCammon, Biophysical Journal **72**(3), 1047 (1997)
11. N. Singh, A. Warshel, Proteins: Structure, Function, and Bioinformatics **78**(7), 1724 (2010)
12. S.J. Irudayam, R.H. Henchman, The Journal of Physical Chemistry B **113**(17), 5871 (2009)
13. D.A. Winkler, Journal of Chemical Information and Modeling **60**(10), 4421 (2020)
14. I.J. Enyedy, W.J. Egan, Journal of Computer-Aided Molecular Design **22**(3-4), 161 (2008)

15. Y.N. Dauphin, R. Pascanu, C. Gulcehre, K. Cho, S. Ganguli, Y. Bengio, in *Advances in Neural Information Processing Systems* (2014), pp. 2933–2941
16. P.M. Pardalos, D. Shalloway, G. Xue, Journal of Global Optimization **4**(2), 117 (1994)
17. J.P. Doye, D.J. Wales, PhRvL **80**(7), 1357 (1998)
18. Chemical Computing Group ULC, 1010 Sherbooke St. West, Suite #910, Montreal, QC, Canada, H3A 2R7, *Molecular Operating Environment (MOE)*, 2018th edn. (2018)
19. T.A. Halgren, Journal of Computational Chemistry **17**(5-6), 490 (1996)
20. T.A. Halgren, Journal of Computational Chemistry **17**(5-6), 520 (1996)
21. T.A. Halgren, Journal of Computational Chemistry **17**(5-6), 553 (1996)
22. T.A. Halgren, R.B. Nachbar, Journal of Computational Chemistry **17**(5-6), 587 (1996)
23. T. Solmajer, E.L. Mehler, Protein Engineering, Design and Selection **4**(8), 911 (1991)
24. P. Bratley, B.L. Fox, ACM Transactions on Mathematical Software **14**(1), 88 (1988)
25. J.M. Hammersley, The Annals of Mathematical Statistics pp. 447–452 (1950)
26. D.E. Shaw, R.O. Dror, J.K. Salmon, J. Grossman, K.M. Mackenzie, J.A. Bank, C. Young, M.M. Deneroff, B. Batson, K.J. Bowers, et al., in *Proceedings of the Conference on High Performance Computing Networking, Storage and Analysis* (2009), pp. 1–11
27. S. Páll, M.J. Abraham, C. Kutzner, B. Hess, E. Lindahl, in *International Conference on Exascale Applications and Software* (Springer, 2014), pp. 3–27
28. G. Dryer, D. Lambert, T. Meek, T. Carr, T. Tomaszek, A. Fernandez, H. Bartus, E. Cacciavilani, A. Hassell, M. Minnich, et al., Biochemistry **31**, 6646 (1992)
29. M. James, A. Sielecki, J. Moult, V. Hruby, D. Rich, in *Peptides: Structure and Function Proceedings of the of the Eighth American Peptide Symposium* (1983), pp. 521–530
30. A.T. Brünger, D.J. Leahy, T.R. Hynes, R.O. Fox, Journal of Molecular Biology **221**(1), 239 (1991)
31. P.J. Ala, E.E. Huston, R.M. Klabe, P.K. Jadhav, P.Y. Lam, C.H. Chang, Biochemistry **37**(43), 15042 (1998)
32. D.W. Banner, P. Hadvary, Journal of Biological Chemistry **266**(30), 20085 (1991)
33. G.W. Zhou, J. Guo, W. Huang, R.J. Fletterick, T.S. Scanlan, Science **265**(5175), 1059 (1994)
34. J. Cooper, W. Quail, C. Frazao, S. Foundling, T. Blundell, C. Humblet, E. Lunney, W. Lowther, B. Dunn, Biochemistry **31**(35), 8142 (1992)
35. C. Mattos, B. Rasmussen, X. Ding, G.A. Petsko, D. Ringe, Nature Structural & Molecular Biology **1**(1), 55 (1994)
36. H. Brandstetter, D. Turk, H.W. Hoeffken, D. Grosse, J. Stürzebecher, P.D. Martin, B.F. Edwards, W. Bode, Journal of Molecular Biology **226**(4), 1085 (1992)
37. D.A. Holt, J.I. Luengo, D.S. Yamashita, H.J. Oh, A.L. Konialian, H.K. Yen, L.W. Rozamus, M. Brandt, M.J. Bossard, M.A. Levy, et al., Journal of the American Chemical Society **115**(22), 9925 (2002)
38. Y. Liu, N.S. Gray, Nature Chemical Biology **2**(7), 358 (2006)
39. I. García-Sáez, A. Párraga, M.F. Phillips, T.J. Mantle, M. Coll, Journal of Molecular Biology **237**(3), 298 (1994)
40. D. Ghosh, M. Erman, Z. Wawrzak, W.L. Duax, W. Pangborn, Structure **2**(10), 973 (1994)
41. A. Zhang, R.G. Nanni, T. Li, G.F. Arnold, D.A. Oren, A. Jacobo-Molina, R.L. Williams, G. Kamer, D.A. Rubenstein, Y. Li, et al., Journal of Molecular Biology **230**(3), 857 (1993)
42. L. Tong, S. Pav, S. Mui, D. Lamarre, C. Yoakim, P. Beaulieu, P.C. Anderson, Structure **3**(1), 33 (1995)
43. B. Nagar, W.G. Bornmann, P. Pellicena, T. Schindler, D.R. Veach, W.T. Miller, B. Clarkson, J. Kuriyan, Cancer Research **62**(15), 4236 (2002)
44. J.M. LaLonde, D.A. Bernlohr, L.J. Banaszak, Biochemistry **33**(16), 4885 (1994)
45. S.P. White, D.L. Scott, Z. Otwinowski, M.H. Gelb, P.B. Sigler, Science **250**(4987), 1560 (1990)
46. J. Rahuel, J.P. Priestle, M.G. Grütter, Journal of Structural Biology **107**(3), 227 (1991)
47. P.C. Weber, D.H. Ohlendorf, J. Wendoloski, F. Salemme, Science **243**(4887), 85 (1989)
48. R.D. Taylor, P.J. Jewsbury, J.W. Essex, Journal of Computer-Aided Molecular Design **16**(3), 151 (2002)
49. A. Holmberg, A. Blomstergren, O. Nord, M. Lukacs, J. Lundeberg, M. Uhlén, Electrophoresis **26**(3), 501 (2005)
50. B.J. Sung, K.Y. Hwang, Y.H. Jeon, J.I. Lee, Y.S. Heo, J.H. Kim, J. Moon, J.M. Yoon, Y.L. Hyun, E. Kim, et al., Nature **425**(6953), 98 (2003)
51. D.W. Banner, P. Hadvary, Journal of Biological Chemistry **266**(30), 20085 (1991)
52. J.H. Arevalo, M.J. Taussig, I.A. Wilson, Nature **365**(6449), 859 (1993)
53. J. Badger, I. Minor, M.A. Oliveira, T.J. Smith, M.G. Rossmann, Proteins: Structure, Function, and Bioinformatics **6**(1), 1 (1989)

54. R.J. Young, D. Brown, C.L. Burns-Kurtis, C. Chan, M.A. Convery, J.A. Hubbard, H.A. Kelly, A.J. Pateman, A. Patikis, S. Senger, et al., Bioorganic & Medicinal Chemistry Letters **17**(10), 2927 (2007)
55. J.T. Bolin, D.J. Filman, D.A. Matthews, R.C. Hamlin, J. Kraut, Journal of Biological Chemistry **257**(22), 13650 (1982)
56. R. Bone, J.P. Vacca, P.S. Anderson, M.K. Holloway, Journal of the American Chemical Society **113**(24), 9382 (1991)
57. M.M. Thunnissen, A. Eiso, K.H. Kalk, J. Drenth, B.W. Dijkstra, O.P. Kuipers, R. Dijkman, G.H. de Haas, H.M. Verheij, Nature **347**(6294), 689 (1990)
58. H. Kim, W.N. Lipscomb, Biochemistry **29**(23), 5546 (1990)
59. K. Bastard, C. Prévost, M. Zacharias, Proteins: Structure, Function, and Bioinformatics **62**(4), 956 (2006)
60. J.P. Ebejer, G.M. Morris, C.M. Deane, Journal of Chemical Information and Modeling **52**(5), 1146 (2012)
61. A.E. Cleves, A.N. Jain, Journal of Computer-Aided Molecular Design **31**(5), 419 (2017)
62. V.I. Minkin, A.D. Garnovskii, J. Elguero, A.R. Katritzky, O.V. Denisko, Advances in Heterocyclic Chemistry **76**, P157 (2000)
63. P. Pospisil, P. Ballmer, L. Scapozza, G. Folkers, Journal of Receptors and Signal Transduction **23**(4), 361 (2003)
64. K.S. Alongi, G.C. Shields, Annual Reports in Computational Chemistry **6**, 113 (2010)
65. A.C. Van Duin, S. Dasgupta, F. Lorant, W.A. Goddard, The Journal of Physical Chemistry A **105**(41), 9396 (2001)
66. J.E. Nielsen, J.A. McCammon, Protein Science **12**(9), 1894 (2003)
67. M.B. Peters, Y. Yang, B. Wang, L. Fusti-Molnar, M.N. Weaver, K.M. Merz Jr, Journal of Chemical Theory and Computation **6**(9), 2935 (2010)
68. A. Vedani, D.W. Huhta, Journal of the American Chemical Society **112**(12), 4759 (1990)
69. M.A. Addicoat, N. Vankova, I.F. Akter, T. Heine, Journal of Chemical Theory and Computation **10**(2), 880 (2014)
70. M. Levitt, B.H. Park, Structure (London, England) **1**(4), 223 (1993)
71. P.A. Karplus, C. Faerman, Current Opinion in Structural Biology **4**(5), 770 (1994)
72. B. Breiten, M.R. Lockett, W. Sherman, S. Fujita, M. Al-Sayah, H. Lange, C.M. Bowers, A. Heroux, G. Krilov, G.M. Whitesides, Journal of the American Chemical Society **135**(41), 15579 (2013)
73. D. Cappel, W. Sherman, T. Beuming, Current Topics in Medicinal Chemistry **17**(23), 2586 (2017)