Research Article

www.molinf.com

molecular informatics

doi.org/10.1002/minf.202200034

Towards an Enrichment Optimization Algorithm (EOA)-based Target Specific Docking Functions for Virtual Screening

Jacob Spiegel[a] and Hanoch Senderowitz*[a]

Abstract: Docking-based virtual screening (VS) is a common starting point in many drug discovery projects. While ligand-based approaches may sometimes provide better results, the advantage of docking lies in its ability to provide reliable ligand binding modes and approximated binding free energies, two factors that are important for hit selection and optimization. Most docking programs were developed to be as general as possible and consequently their performances on specific targets may be sub-optimal. With this in mind, in this work we present a method for the development of target-specific scoring functions using our recently reported Enrichment Optimization Algorithm (EOA). EOA derives QSAR models in the form of multiple linear regression (MLR) equations by optimizing an enrichment-like metric. Since EOA requires target-specific active and inactive (or decoy) compounds, we retrieved such data for six targets from the DUD-E database, and used them to re-derive the weights associated with the components that make up GOLD's *ChemPLP* scoring function yielding target-specific, modified functions. We then used the original *ChemPLP* function in small-scale VS experiments on the six targets and subsequently rescored the resulting poses with the modified functions. In addition, we used the modified functions for compounds re-docking. We found that in many although not all cases, either rescoring the original *ChemPLP* poses or repeating the entire docking process with the modified functions, yielded better results in terms of AUC and $EF_{1\%}$, two metrics, common for the evaluation of VS performances. While work on additional datasets and docking tools is clearly required, we propose that the results obtained thus far hint to the potential benefits in using EOA-based optimization for the derivation of target-specific functions in the context of virtual screening. To this end, we discuss the downsides of the methods and how it could be improved.

Keywords: enrichment optimization algorithm · EOA · docking · virtual screening · QSAR · GOLD · target specific scoring functions

1 Introduction

Virtual screening (VS) is a common starting point in many drug discovery projects and is performed using a variety of computational techniques such as docking,[1–3] pharmacophore modeling,[4–6] similarity and substructure searches,[7–13] and QSAR equations.[14–16] These techniques are typically classified as either structure-based (i.e., techniques that utilize information on the structure of the bio-target) or ligand-based (i.e., techniques that utilize information on the ligands). An alternative classification scheme could be based on the amount of ligand-related information required to conduct each type of VS. Thus, VS using QSAR equations (or any other type of machine learning-based models) requires a reasonably large dataset of both active and inactive ligands to properly derive and validate the model. Pharmacophore models could be constructed from much smaller datasets and moreover, the availability of data on inactive compounds is not mandatory. Yet such models greatly benefit from accurate knowledge of the ligands' bioactive conformations. Finally, both substructure searches and similarity searches could in principle be initiated from a single active compound.

In contrast with the above, docking-based VS does not require any information on ligands and in principle could be conducted on any target provided its structure is known or could be modeled. Indeed, most docking tools were developed to be as general and as target-independent as possible.[17–21] While this undoubtedly increases the domain of applicability of these tools, it may also compromise their performances on specific targets. Thus, when information on active (and inactive) ligands for a specific target is available, it is common practice to use this information

[a] J. Spiegel, H. Senderowitz
Department of Chemistry, Bar-Ilan University, Ramat-Gan 5290002, Israel
E-mail: hsenderowitz@gmail.com

Supporting information for this article is available on the WWW under https://doi.org/10.1002/minf.202200034

either to select the best docking tool/protocol from among the many available options or to calibrate a specific protocol to the problem at hand. This is typically done by modifying certain parameters (e.g., parameters pertaining to the search algorithm, to the scoring function or to both) in a target-specific manner. The success of this selection/calibration process is often evaluated by VS-aware metrics such as enrichment, area under the ROC curve (AUC) and the percentage of active compounds found within the first 1% of the screened library ($EF_{1\%}$). However, these metrics are not typically used to direct the search for the optimal docking protocol but rather to evaluate different protocols through what is mostly a trial and error process.

One of the decisive factors for the success of any docking-based VS experiment is the quality of the scoring function.[22] Many scoring functions (so-called knowledge based scoring functions[23–27]) could be regarded as QSAR equations. In these equations, the descriptors are calculated from ligand-protein poses and describe ligand-protein key interactions (e.g., H-bonds, salt bridges, hydrophobic interactions, aromatic interactions, VdW interactions) and the weights associated with each type of interaction are selected so as to reproduce the experimental binding free energies of a large and diverse set of ligand-protein complexes.

We reasoned that in order to turn a general scoring function into a target-specific one, it might suffice to re-derive the weights associated with the different terms that make up the scoring function, based on information available on ligands binding to this target. Furthermore, in order for the modified function to be useful in the context of VS, we propose that the re-derivation process should be based on the optimization of a VS-aware metric.

Recently, we have introduced the enrichment optimization algorithm (EOA), that derives QSAR models in the form of multiple linear regression (MLR) equations by optimizing an enrichment-like metric, and demonstrated its superiority in small-scale VS experiments over QSAR equations derived by optimizing a "classical" metric[28] and over three docking tools commonly used for VS.[4] Given the encouraging performances of EOA, in this work we apply it for the re-derivation of the GOLD scoring function.

GOLD (Genetic Optimization for Ligand Docking)[29] is one of the most accurate and popular docking programs. GOLD uses a genetic algorithm-based conformational search of the ligand and of key residues in the protein's binding site (thereby accounting, at least partially, for the flexibility of the protein). One of the scoring functions used by GOLD is *ChemPLP*[30], which combines the PLP term that models protein-ligand interactions with additional terms that model the ligands' internal energy, allocating default and constant weights to each term. Using the EOA approach, in this work we present the re-derivation of *ChemPLP*'s weights for six targets for which sufficient information on active and inactive ligands for the derivation of EOA models is available. The modified *ChemPLP* equations are then used in the context of VS for pose rescoring and for ligand re-docking.

Several methods for the derivation of target specific scoring functions in the context of docking were reported in the literature.[26,31–36] These methods mostly applied various machine learning algorithms either to re-derive the weights associated with the energy components that make up specific scoring functions or to develop new functions based on descriptors derived from predicted binding modes. However, except of one case,[37] all of these functions were used for pose rescoring rather than for ligand re-docking and none were based on the optimization of a VS-aware metric as suggested in this work.

2 Materials and Methods

2.1 Datasets

Datasets for six protein targets, namely, Androgen Receptor (ANDR), Cytochrome P450 2C9 (CYP2C9), Glucocorticoid receptor (GCR), Human immunodeficiency virus type 1 reverse transcriptase (HIVRT), Cyclooxygenase-1 (PGH1), and Progesterone receptor (PRGR) were retrieved from the DUD-E site.[38,39] The PDB structures associated with each target in the DUD-E database are: 2am9, 1r9o, 3bqd, 3lan, 2oyu and 3kba respectively. These targets represent four different protein families according to the Pfam classifier:[40] The Nuclear Receptor superfamily is represented by ANDR, GCR and PRGR, the Membrane Associated Proteins in Eicosanoid and Glutathione metabolism (MAPEG) family is represented by PGH1, the Cytochromes P450 superfamily is represented by CYP2C9 and the Rnase H family is represented by HIVRT. As per the DUD-E setup, all datasets contain compounds that were experimentally determined to be active as well as decoy compounds (see Table 1). Importantly, we did not attempt to provide a representative subset of targets but rather to focus on challenging ones. To this end, we selected from DUD-E targets for which poor Area Under the Curve (AUC) and enrichment values are reported, provided the datasets associated with them were large enough (i.e. contained sufficient number of actives) for the derivation of EOA models. The AUC and enrichment values for the selected targets are: ANDR: AUC = 51.06, Enrichment = 5.6; CYP2C9: AUC = 59.85, Enrichment = 2.5; GCR: AUC = 43.92, Enrichment = 8.9; HIVRT: AUC = 64.35, Enrichment = 6.5; PGH1: AUC = 52.97, Enrichment = 4.6; PRGR: AUC = 56.37, Enrichment = 7.8. Although the AUC values of our results are limited to the [0,1] range, the AUC values of DUD-E are measured by percentages on their website and in their paper, therefore, their AUC values are limited to the [0,100] range..[38,39]

The active and decoy compounds, and the proteins considered in this work were prepared by Schrodinger's LigPrep[41] program and the Protein Preparation Wizard program,[42] respectively. Protein preparation consisted of

Table 1. Description of the different datasets used in this work. Four equal subsets were derived for each protein target.

Target	PDB	Total # of Actives	Total # of Decoys	Train		Validation		Test	
				Actives	Decoys	Actives	Decoys	Actives	Decoys
PGH1	2oyu	244	10,286	164	1333	40	332	40	8621
CYP2C9	1r9o	172	6,777	116	1333	28	332	28	5112
ANDR	2am9	513	13,509	343	2000	85	499	85	11,010
PRGR	3kba	439	14,993	294	2000	72	499	73	12,494
GCR	3bqd	522	13,792	349	2000	86	499	87	11,293
HIVRT	3lan	634	14,314	424	2000	105	499	105	11,815

completion of missing side chains/residues, assignment of correct protonation states for ionizable residues and addition of hydrogen atoms. Ligand preparation consisted of assigning correct tautomeric forms and correct protonation states at $pH=7$, and obtaining reliable conformations.

The descriptors used in this work for the derivation of EOA equations (see below) were those that make up GOLD's *ChemPLP* scoring function and include: $S_{PLP,}$ which represents the Piecewise Linear Potential used to model the steric complementarity between protein and ligand, $S_{hbond,}$ which represents the distance-dependent hydrogen bonding, S_{cho} which represents the angle-dependent hydrogen bonding, S_{clash} and S_{tors} which account for the ligand clashes and torsion angles, respectively, and S_{metal} which accounts for metal-ligand interactions. The S_{metal} descriptor that was found to be constant across all targets and compounds was removed. In addition, the S_{cho} descriptor which was found to be constant for all PGH1 and CYP2C9 compounds was also removed from these targets.

For the purpose of developing enrichment optimizer algorithm (EOA) models (see below), four subsets, each with identical numbers of active and decoy compounds, were randomly selected from each protein dataset. Each of these subsets were randomly divided into training and validation sets in an 80%/20% ratio, and the unselected compounds served as test sets. As a result of this selection procedure, the test sets contained much smaller percentages of active compounds (0.5–0.9%). This was done on purpose in order to mimic a real VS scenario. Table 1 provides a description of the different datasets used in this work.

2.2 Molecular Docking

The docking program that was chosen for this study, is GOLD (version 2021.1.0),[29] since the scoring function it uses for docking could be readily modified by the user. First, all compounds from each dataset, actives and decoys, were docked to the active sites of their respective proteins using the *ChemPLP* function. For the purpose of docking, the population size for the genetic algorithm was set to 100 and the maximal number of operations per ligand was set to 100,000. For each compound, the pose with the best (highest) *ChemPLP* value was kept and together with the clashes and torsional terms used for calculating the final score (equations 1 and 2):

$$ChemPLP = -\begin{pmatrix} 1.0 \times S(PLP) + (-3.0 \times S(hbond)) + \\ (-6.0 \times S(metal)) + (-3.0 \times S(cho)) \end{pmatrix} \quad (1)$$

$$Score = ChemPLP - 1.0 \times DE(clash) - 2.0 \times DE(tors) \quad (2)$$

Next, raw (i.e., unweighted) values for all the components that make up the *ChemPLP* function for all ligands were extracted from the output files and those pertaining to the training set ligands were used for the derivation of new weights by EOA as described below. The newly derived weights give rise to modified *ChemPLP* scoring functions (also referred to as EOA-derived scoring function) that were used for both compounds rescoring and compounds re-docking as described below.

2.3 Enrichment Optimizer Algorithm (EOA) Algorithm

In our previous works,[28,43] we presented a novel algorithm for the derivation of multiple linear regression (MLR) equations suitable for usage in virtual screening, based on the optimization of an enrichment-aware objective function. We termed the new algorithm, enrichment optimizer algorithm (EOA). Briefly, EOA accepts as input a set of L active compounds together with a set of O inactive (either known inactive or decoy) compounds characterized by a set of N molecular descriptors. The algorithm then derives an MLR equation by selecting a subset of the molecular descriptors and by assigning to each of them an arbitrary weight. The resulting equation is then used to rank all compounds and the ranked list is scored by a scoring function composed on two components: (1) A primary score consisting of the number of active compounds found within the first L places of the ranked list and ranging between 0 and L. (2) A secondary score which is calculated based on the ranks of the inactive compounds found within the first L places of the ranked list and the ranks of the active compounds found beyond the first L places of the ranked list. The range of this secondary score was purposely limited

to [0,1] range. This combined score is optimized in the space of the descriptors and their weights using a Monte Carlo/simulated annealing (MC/SA) optimizer. The EOA algorithm is detailed in the SI.

In this work EOA models were derived for each of the four subsets selected from each of the six parent datasets (a total of $6\times4=24$ models) using the components that make up the *ChemPLP* scoring function as descriptors. In all cases, models were derived from the training set, evaluated first on the validation set, and then on the test set compounds. Of note, both the validation and test sets are perfectly valid external sets (i.e., sets not used in any way during the model development phase), the only difference between them being the percentage of active compounds. As noted above, a much smaller percent of active compounds was allocated to the test set in order to mimic a real-world VS scenario.

As noted above, all the components that make up GOLD's *ChemPLP* scoring function were used as descriptors except for the *Smetal* descriptor for all datasets and the *Scho* descriptor for the PGH1 and CYP2C9 datasets. In contrast to our previous works,[28,43] in the present study EOA models were derived using all descriptors and only the descriptors weights were optimized. Furthermore, weights were allowed to vary only within a limited range around their original values. This was done since all attempts at unconstrained optimization resulted in unrealistic weights' values. Thus, weights for the S_{plp}, S_{clash} and S_{tor} descriptors were limited to the [0,10] range, and for the S_{hbond} and S_{cho} descriptors the weights were limited to the [0,10] range. A typical MC/SA run for the derivation of an EOA model consisted of 1,000,000 MC steps. Simulated annealing was implemented by means of a saw-tooth procedure whereby repeated annealing cycles were performed. In each cycle the RT term was linearly decreased from 1 to 0.01 in 0.01 intervals, running 400 MC steps per interval. The range of values of the RT term led to an acceptance rate of roughly 2–6%.

The resulting EOA equations (original descriptors and optimized weights) were used in the context of virtual screening both for pose rescoring and for compounds re-docking. For the purpose of pose rescoring, poses obtained with GOLD's original scoring function (equations 1 and 2 above) were rescored using the EOA-derived equations and the evaluation metrics (see below) calculated on the rescored poses. For the purpose of compounds re-docking, the entire docking process was repeated with the EOA-derived equations on each subset for a total of $4\times6=24$ re-docking experiments, and the evaluation metrics calculated on the newly generated poses (best scoring pose for each compound).

2.4 Evaluation Metrics

EOA models were first validated on the validation and test sets by counting the number of active compounds found within the top *L* places of the list ranked according to the EOA equations (Table 2). Next, to test whether the new weights, when used with *ChemPLP*'s components make physical sense in the context of docking, we used the modified *ChemPLP* functions to dock each ligand to its respective binding site and evaluated the results using the RMSD metric (Table 3). Finally, the performances of all virtual screening experiments (using the original GOLD, rescoring the original GOLD poses using the modified, EOA-derived functions, and re-docking the compounds into their respective sites using the modified, EOA-derived functions) were evaluated using two common metrics, namely area under the ROC curve (AUC) and Enrichment at 1% of the library ($EF_{1\%}$) (Tables 4 & 5). AUC informs on the overall performances of the VS procedure whereas $EF_{1\%}$ informs on performances at early stages of the screening process.

3 Results

Table 2 presents the results obtained with the EOA models for all four subsets derived from all six protein datasets considered in this work in terms of the number of active compounds found within the first *L* places of the EOA-ranked list.

These results demonstrate the expected yet small decrease in performances on going from the training to the validation sets with an average percentage of actives within the first *L* places of the ranked list of 23.8% and 21.0% for the training and validation sets, respectively. The Pearson correlation between the two sets of values is high at 0.73 indicating the consistency of the results. A sharp decrease is found when going to the test sets. We attribute this decrease to the much smaller percentage of active compounds in these sets (see Method section) making them more difficult to identify within a large pool of decoys. We have previously observed a similar decrease in EOA performances for five other protein targets.[43] Of note the performances of these EOA models are poorer than the one previously reported by us.[28,43] We attribute this to the much smaller number of descriptors used for the derivation of the EOA models.

Next, to test whether the EOA-derived functions make physical sense, we used them to dock each ligand into its respective binding site. The results are presented in Table 3 and demonstrate that in almost all cases, best energy poses obtained with these functions matched the crystallographic poses, RMSD-wise, better than what was obtained with the original GOLD function.

Next, table 4 presents a comparing of the virtual screening results performed on all four test sets for each protein target using the original GOLD scoring function, the original

Table 2. EOA results obtained for all four subsets from all six datasets. Results are provided in terms on the number and percentage (based on the total number of actives) of active compounds appearing within the first L places of the list ranked according to the EOA equation. See the Materials and Method section for an explanation how the different subsets were derived, and Table1 for the composition of the subsets.

Target	# Actives = *L*			# Actives among *L* top Places		
	Train	Validation	Test	Train (%)	Validation (%)	Test (%)
PGH1-1	164	40	40	26 (16%)	10 (25%)	0 (0%)
PGH1-2				32 (20%)	6 (15%)	0 (0%)
PGH1-3				30 (18%)	3 (8%)	0 (0%)
PGH1-4				37 (23%)	4 (10%)	0 (0%)
CYP2C9-1	116	28	28	21 (18%)	3 (11%)	0 (0%)
CYP2C9-2				26 (22%)	6 (21%)	1 (4%)
CYP2C9-3				22 (19%)	5 (18%)	0 (0%)
CYP2C9-4				26 (22%)	4 (14%)	0 (0%)
ANDR-1	343	85	85	71 (21%)	15 (18%)	0 (0%)
ANDR-2				63 (18%)	17 (20%)	0 (0%)
ANDR-3				66 (19%)	18 (21%)	0 (0%)
ANDR-4				67 (20%)	16 (19%)	3 (4%)
PRGR-1	294	72	73	58 (20%)	12 (17%)	0 (0%)
PRGR-2				51 (17%)	8 (11%)	0 (0%)
PRGR-3				51 (17%)	12 (17%)	1 (1%)
PRGR-4				51 (17%)	8 (11%)	0 (0%)
GCR-1	349	86	87	133 (38%)	36 (42%)	7 (8%)
GCR-2				140 (40%)	31 (36%)	4 (5%)
GCR-3				132 (38%)	37 (43%)	3 (3%)
GCR-4				134 (38%)	31 (36%)	4 (5%)
HIVRT-1	424	105	105	123 (29%)	27 (26%)	3 (3%)
HIVRT-2				123 (29%)	19 (18%)	8 (8%)
HIVRT-3				123 (29%)	29 (28%)	8 (8%)
HIVRT-4				124 (29%)	29 (28%)	5 (5%)

Table 3. A comparison between RMSD values obtained by the original GOLD function and by the EOA-derived functions. For each target, the best RMSD is highlighted in orange. The RMSD values in bold are lower (i.e. better) than those obtained with the original GOLD value.

Target	EOA-derived functions					Original GOLD
	Set1	Set2	Set3	Set4	Average	
PGH1	**2.06**	**5.58**	**6.12**	**2.49**	**4.06**	7.52
CYP2C9	**1.88**	**3.01**	**1.75**	**3.15**	**2.45**	6.22
ANDR	**0.82**	**0.81**	**0.80**	**0.80**	**0.81**	0.82
PRGR	**0.85**	**0.92**	**0.84**	**0.67**	**0.82**	0.99
GCR	**0.56**	**0.83**	**0.95**	**0.56**	**0.73**	0.94
HIVRT	0.59	0.60	0.39	0.47	0.51	0.35

GOLD poses rescored by the EOA-derived equations and the EOA equations for compounds re-docking. Table 5 presents the averaged values per target (derived from Table 4).

Looking at the results from the individual test sets (Table 4), the original GOLD docking provides the best AUC in one case (PRGR) and the best $EF_{1\%}$ value in one case (ANDR), rescoring provides the best AUC in three cases (ANDR, GCR, and HIVRT) and the best $EF_{1\%}$ values in three cases (PRGR, GCR, and HIVRT), and re-docking provides the best AUC in two cases (PGH1, and CYP2C9) and the best $EF_{1\%}$ values in two cases (PGH1, and CYP2C9). Interestingly, when original GOLD provides the best performances, these are closely matched by either of the other two methods (e.g., original GOLD AUC for PRGR is 0.61 whereas re-docking AUC for PRGR is 0.60 and original GOLD $EF_{1\%}$ for ANDR is 4.51 whereas rescoring $EF_{1\%}$ for ANDR is 3.53). However, when the best results are obtained either by rescoring or by re-docking, in most cases they clearly outperform original GOLD. This is particularly evident for $EF_{1\%}$ values. These trends are also reflected in the averaged values (Table 5) although to a lesser extent. Figure 1 provides the ROC curves with the best AUC across all test sets for each target.

Table 4. A comparison of AUC and $EF_{1\%}$ values for all test sets and all targets between original Gold, EOA-based rescoring of GOLD poses and EOA-based re-docking. Blue and orange shading mark the highest AUC score and highest $EF_{1\%}$ value for each target, respectively.

Target	Original GOLD		Rescoring		Re-docking	
	AUC	$EF_{1\%}$	AUC	$EF_{1\%}$	AUC	$EF_{1\%}$
PGH1-1	0.61	5.38	0.57	0.00	0.60	0.00
PGH1-2	0.58	6.90	0.51	2.50	0.66	8.62
PGH1-3	0.52	6.90	0.53	0.00	0.65	3.45
PGH1-4	0.58	3.23	0.47	0.00	0.72	11.29
CYP2C9-1	0.69	6.98	0.72	3.57	0.68	4.65
CYP2C9-2	0.59	0.00	0.63	3.57	0.72	12.50
CYP2C9-3	0.61	6.12	0.70	3.57	0.71	18.37
CYP2C9-4	0.57	0.00	0.58	3.57	0.77	3.45
ANDR-1	0.46	1.84	0.51	0.00	0.44	1.84
ANDR-2	0.44	2.30	0.53	1.18	0.42	0.46
ANDR-3	0.47	4.29	0.50	0.00	0.45	3.00
ANDR-4	0.46	4.51	0.56	3.53	0.41	0.00
PRGR-1	0.61	2.44	0.38	1.37	0.60	2.44
PRGR-2	0.50	2.46	0.49	2.74	0.43	0.00
PRGR-3	0.56	4.10	0.44	4.11	0.57	4.10
PRGR-4	0.56	2.59	0.48	2.74	0.50	0.00
GCR-1	0.35	6.23	0.63	8.05	0.34	6.23
GCR-2	0.34	5.05	0.63	4.60	0.35	5.05
GCR-3	0.36	4.00	0.63	5.75	0.37	3.11
GCR-4	0.34	3.45	0.61	4.60	0.33	4.43
HIVRT-1	0.41	4.58	0.48	3.81	0.37	4.20
HIVRT-2	0.38	1.82	0.53	7.62	0.32	1.09
HIVRT-3	0.38	3.44	0.49	7.62	0.34	2.06
HIVRT-4	0.41	5.52	0.45	4.76	0.37	5.20

Table 5. A comparison of averaged AUC and $EF_{1\%}$ values for all targets between original Gold, EOA-based rescoring of GOLD poses and EOA-based re-docking. Blue and orange shading mark the highest AUC score and highest $EF_{1\%}$ value for each target, respectively.

Target	Original GOLD		Rescoring		Re-docking	
	AUC	$EF_{1\%}$	AUC	$EF_{1\%}$	AUC	$EF_{1\%}$
PGH1	0.57	5.60	0.52	0.63	0.66	5.84
CYP2C9	0.61	3.27	0.66	3.57	0.72	9.74
ANDR	0.46	2.81	0.51	0.39	0.43	1.77
PRGR	0.56	2.90	0.45	2.74	0.53	1.63
GCR	0.35	4.68	0.63	5.75	0.35	4.70
HIVRT	0.39	3.84	0.49	5.95	0.35	3.14

4 Discussion and Conclusions

Docking-based virtual screening (VS) is a common starting point in many drug discovery projects. While ligand-based approaches may sometimes provide better results,[44–46] the advantage of docking is that it can provide reliable ligand binding modes and approximated binding free energies, two factors that are important for hit selection and optimization.

As noted in the introduction, VS-aware metrics could be used to select, from within multiple docking protocols/scoring functions, that which is most suitable for a particular target. However, these metrics do not actively direct the search for the best protocol. In this work, we therefore take another, more direct approach, for generating target-specific scoring functions for docking-based virtual screening. Using our recently developed Enrichment Optimization Algorithm (EOA) we optimize (rather than just use) a VS-aware metric (see methods section and the description of the EOA in the SI. The optimized function is defined in step (11) as the "total score") in the space of the weights associated with the components that make up GOLD's *ChemPLP* scoring function, thereby deriving target-specific modified *ChemPLP* functions. The clear advantage of this approach over the above-described, basically trial and error docking protocol selection/calibration procedures, lies in its generality and unbiased nature. Given a set of active and inactive or decoy compounds docked to a specific target, EOA would identify the weights that would afford the best value of the optimized function. Importantly, the method is not limited to GOLD's *ChemPLP*. Rather, pose rescoring is applicable to any docking tool that outputs the

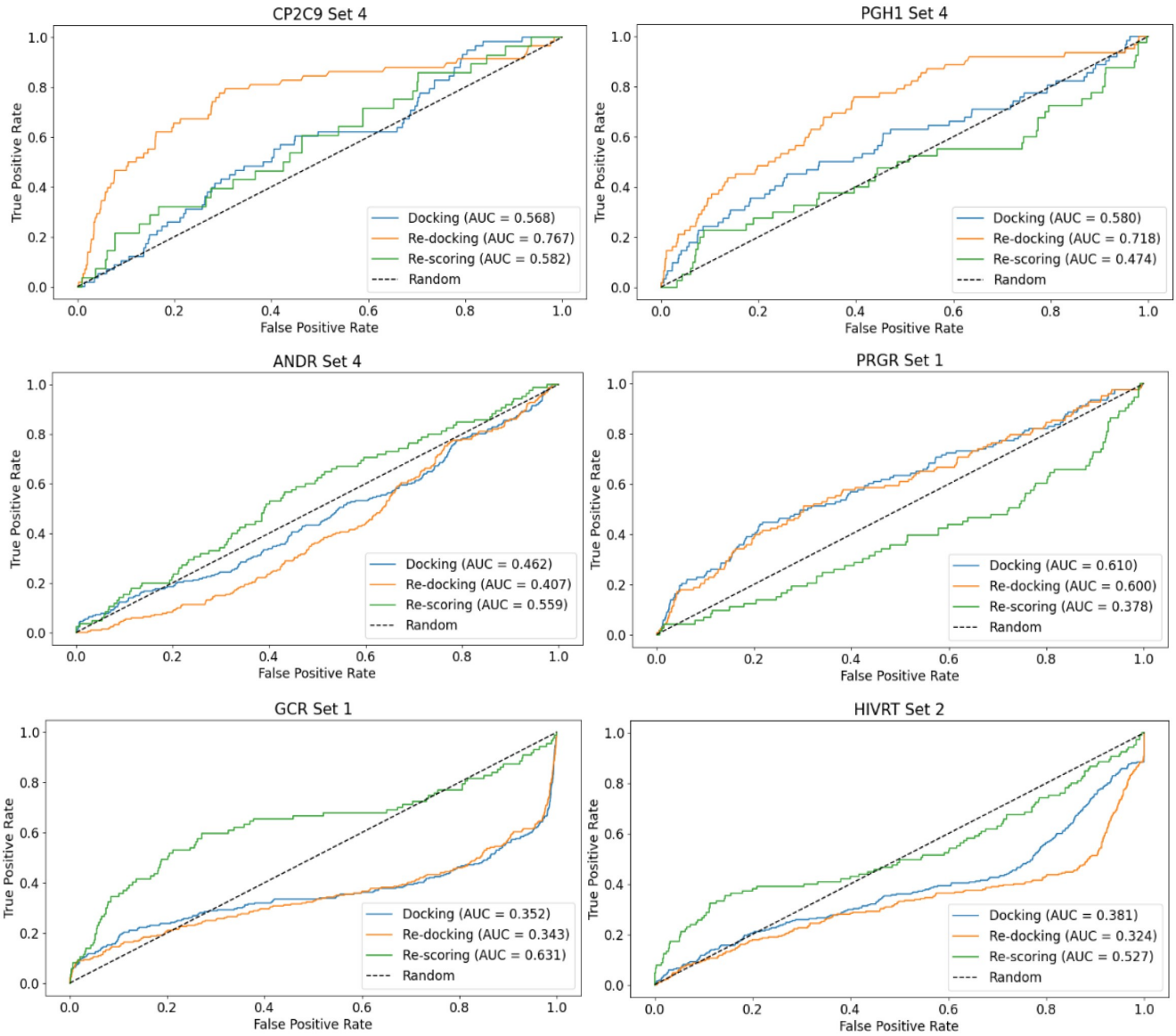

Figure 1. The ROC curve with the best AUC across all test sets for each target.

components that make up its scoring function while compound re-docking is applicable to any docking tool that allows for the modification of its docking function.

In this work, modified *ChemPLP* functions were derived for six targets covering four protein families (Pfam classification) selected from the DUD-E database. We deliberately focused on targets for which the poorest AUC and $EF_{1\%}$ metrics were reported in DUD-E under the assumption that these present the best opportunities for improvements. The modified functions were derived in a manner consistent with common best practices.[15,47,48] Thus, derivation was based on a training set and the resulting functions validated on independent validation and test sets. All divisions into training/validation/test sets were performed at random and repeated four times. Due to the small number of descriptors, the resulting functions are unlikely to be over-fitted or chance correlated.

The performances of the EOA models on training, validation and test sets in terms of the number of retrieved active compounds are reported in Table 2. These are lower than those previously reported by us on other datasets.[28,43] We attribute this to the much smaller number of descriptors used in the derivation of the EOA equations (4–5 descriptors) in comparison with our previous works, as well as to our focusing on "difficult" targets (see above). This second argument can also account for the lower than previously reported performances of the EOA models (as well as of the original GOLD) in the context of VS (either for pose

rescoring or for ligand re-docking) in terms of the AUC and $EF_{1\%}$ metrics (Tables 4 and 5).

A comparison between the results obtained with the original *ChemPLP* function and the modified *ChemPLP* functions used for pose rescoring and compounds re-docking in terms of both AUC and $EF_{1\%}$ values is encouraging although not conclusive. Using the original *ChemPLP* results as a baseline and focusing first on the less sensitive AUC metric, pose rescoring led to an increase in AUC values in 17 out of the 24 cases (Table 4) with an average increase over these cases of 29.6%. Looking at the other seven cases where the original *ChemPLP* function provided better results, a much small average decrease of only 16.1% is observed. Similarly, compounds re-docking led to an increase in AUC values in nine out of the 24 cases with an average increase of 17.8%. Looking at the remaining 15 compounds where the original *ChemPLP* outperformed the modified functions, a much small decrease of only 7.0% is observed. Providing a percentage-based statistics in the case of $EF_{1\%}$ is complicated by the fact that in several cases (across all methods) zero $EF_{1\%}$ values were obtained. Moreover, since the test sets used in this work were deliberately designed to be highly unbalanced with far fewer active compounds than decoys in order to mimic real-life VS campaigns (see Table 1), small changes in the number of active compounds found within the first 1% of the ordered list may have a large effect on the nominal $EF_{1\%}$ values. Still, looking at absolute (rather than percentage-based) differences we note that pose rescoring outperformed the original *ChemPLP* function in ten out of 24 cases with an average increase of 2.2 $EF_{1\%}$ "units" whereas the average decrease over the other 14 cases where the original *ChemPLP* function gave better results was similar at 2.7 $EF_{1\%}$ units. Compounds re-docking outperformed the original *ChemPLP* function in only six out of 24 cases with an average increase of 6.5 $EF_{1\%}$ "units" whereas the average decrease over the other 13 cases where the original *ChemPLP* function gave better results was much smaller at 2.1 $EF_{1\%}$ "units" (the two methods were tied in 5 cases). Thus, while the original *ChemPLP* function outperforms the other two methods for more cases, the improvements brought about by the modified functions either in the context of rescoring or in the context of re-docking are more pronounced. On average, the original *ChemPLP* function performed the best in three cases (AUC for PRGR and $EF_{1\%}$ for PRGR and ANDR), rescoring performed the best in five cases (AUC for ANDR, GCR, and HIVRT and $EF_{1\%}$ for GCR, and HIVRT) and re-docking performed the best in four cases (AUC and $EF_{1\%}$ for PGH1 and CYP2C9) (Table 5).

Finally, it was gratifying to see that using the modified functions for ligand docking yielded RMSD values for best energy poses with respect to crystallographic poses that were overall better than those obtained with the original function. Thus, the balance between the components that make up the *ChemPLP* function that is crucial for successful docking was clearly maintained and even improved in the modified functions (Table 3).

Our dataset contained three nuclear receptors, ANDR, GCR, and PRGR, two of which (ANDR and PRGR) binding similar ligands in their crystal structures. Yet, the results obtained for these three receptors are somewhat different. Thus, the EOA models were able to retrieve, on average, 19.5%, 38.5% and 17.8% of the actives in the training set for ANDR, GCR, and PRGR, respectively and 19.5%, 33.8%, and 14.0% of the actives in the validation sets, for these three targets (Table 2). The average AUC values obtained for these proteins with the original GOLD scoring function are 0.46, 0.35, and 0.56, upon rescoring these are 0.51, 0.63 and 0.45 and upon re-docking, these are 0.43, 0.35, and 0.53 (Table 5). These differences could be the result of the intermediate sequence identities between the three proteins (3kba vs. 3bqd: 54.37% identity; 3kba vs. 2am9: 55.02% identity; 2am9 vs. 3bqd: 50.61% identity) as well as to the low number of active compounds common to all. Indeed, the two most similar sets (those for PRGR and GCR) share only 19.14% of their compounds. Since these sets formed the basis for the construction of the EOA models, different models were derived which led to different corresponding performances.

Given the modest inconsistent improvements, brought about by using the modified functions with respect to the original GOLD scoring function, a discussion on the advantages and limitations of the EOA method for the derivation of target-specific scoring functions is in order. On the upside, EOA derives models in a target-specific manner whereas GOLD, like most other docking programs was developed to be as global as possible at the expanse of reduced performances in specific cases. This may well explain the instances where EOA models outperformed the original GOLD scoring function. Importantly, we do not consider EOA's reliance on available data as a drawback but rather as a feature. Clearly, there is no harm in using all available information to improve the outcome of virtual screening. Another important advantage of the method is, as noted above, its generality. EOA could be used to re-derive the scoring function of any docking tool provided the components that make up its scoring function are accessible. In fact docking tools such as Glide[19] that have more complex scoring functions will present EOA with more descriptors, perhaps leading to better models.

However, our method clearly has some limitations. First, the MC/SA is not a very efficient optimizer and was selected mainly due to its ease of implementation. Other optimizers, such as Genetic Algorithm,[49,50] Particle Swarm Optimization[51] or the recently developed Grasshopper algorithm[52] could be used and are likely to provide better convergence on the global minimum in the space of the weights. Related to this, in the current implementation of the algorithm we severely limited the search for the optimal values for the weights to the close vicinity of their original values. However, the optimal, per-target weights could

have very different values from the original ones. Following this idea, in an early attempt we performed an unconstrained search in the space of the weights. This search indeed led to EAO models with much better performances across all metrics. Yet, when these weights were used for pose re-docking, the resulting poses were unrealistic, presumably due to allocating an inappropriate weight to the VdW term. This is not surprising since pose "reliability" was not part of the optimized objective function. Yet, it could be easily added within the framework of multi-object optimization (MOOP). Using MOOP, we could even simultaneously optimize for enrichment, the $EF_{1\%}$ values and pose "reliability".

Next, the scoring function used for the derivation of the EOA equations can also be improved. At present, the optimization procedure only considers active and inactive compounds putting much more emphasize on the former. Thus, a solution containing five inactive compounds within the first *L* places of the ranked list (*L* being the total number of active compounds; see the EOA algorithm in the SI) will always be preferred over a solution containing six inactive compounds within this range, irrespective of the precise location of the inactive compounds in the ranked list. This strategy could be modified to favor cases where inactive compounds are located towards the end of the active portion of the list even if their number is somewhat higher. Furthermore, a more refined partitioning of the data (e.g., active, semi-active, inactive) could be used. For example, the dataset could be divided into *L* active compounds, *M* semi-active compounds and *O* inactive compounds and the algorithm would try to maximize the number of active compounds within the first *L* places of the rank-ordered list, the number of semi-active compounds within the next *M* places and the number of inactive compounds within the *O* last places.

Yet, despite these limitations, the results presented in this work suggest that EOA is an interesting method for the derivation of target-specific scoring function. Clearly, improvement should be introduced and the method should be tested on many more data sets and docking tools to assess its "applicability domain" and performances. Work along these lines is currently being conducted in our laboratory.

Conflict of Intereset

None declared.

References

[1] D. B. Kitchen, H. Decornez, J. R. Furr, J. Bajorath, *Nature Reviews Drug Discovery*. **2004**, *3*, 935–949, https://doi.org/10.1038/nrd1549.

[2] X. Xu, M. Huang, X. Zou, *Biophys. Reports* **2018**, *4* (1), 1–16, https://doi.org/10.1007/S41048-017-0045-8.

[3] B. J. Bender, S. Gahbauer, A. Luttens, J. Lyu, C. M. Webb, R. M. Stein, E. A. Fink, T. E. Balius, J. Carlsson, J. J. Irwin, B. K. Shoichet, *Nat. Protoc.* **2021**, *16* (10), 4799–4832, https://doi.org/10.1038/s41596-021-00597-z.

[4] T. Lengauer, C. Lemmen, M. Rarey, M. Zimmermann, *Drug Discov. Today* **2004**, *9* (1), 27–34, https://doi.org/10.1016/S1359-6446(04)02939-3.

[5] D. Horvath, Pharmacophore-Based Virtual Screening in *Chemoinformatics and Computational Chemical Biology. Methods in Molecular Biology, vol 672* (Ed.: J. Bajorath), Humana Press, Totowa, NJ, **2010**, pp. 261–298, https://doi.org/10.1007/978-1-60761-839-3_11.

[6] D. Schaller, D. Šribar, T. Noonan, L. Deng, T. N. Nguyen, S. Pach, D. Machalz, M. Bermudez, G. Wolber, *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2020**, *10* (4), e1468, https://doi.org/10.1002/WCMS.1468.

[7] P. Willett, *Drug Discov. Today* **2006**, *11* (23–24), 1046–1053, https://doi.org/10.1016/J.DRUDIS.2006.10.005.

[8] T. G. Kristensen, J. Nielsen, C. N. S. Pedersen, *Comput. Struct. Biotechnol. J.* **2013**, *5* (6), e201302009, https://doi.org/10.5936/CSBJ.201302009.

[9] P. Willett, *Comput. Struct. Biotechnol. J.* **2013**, *5* (6), e201302002, https://doi.org/10.5936/CSBJ.201302002.

[10] X. Yan, C. Liao, Z. Liu, T. A. Hagler, Q. Gu, J. Xu, *Curr. Drug Targets* **2016**, *17* (14), 1580–1585, https://doi.org/10.2174/1389450116666151102095555.

[11] H. Eckert, J. Bajorath, *Drug Discov. Today* **2007**, *12* (5–6), 225–230, https://doi.org/10.1016/j.drudis.2007.01.011.

[12] I. Muegge, discovery, P. Mukherjee, *Expert Opin Drug Discov.* **2016**, *11* (2),137–148, https://doi.org/10.1517/17460441.2016.1117070.

[13] J. Vazquez, A. Deplano, A. Herrero, E. Gibert, E. Herrero, F. J. Luque, *J. Chem. Inf. Model.* **2020**, *60* (9), 4231–4245, https://doi.org/10.1021/ACS.JCIM.9B01191/SUPPL_FILE/CI9B01191_SI_001.PDF.

[14] A. Tropsha, A. Golbraikh, *Curr. Pharm. Des.* **2007**, *13* (34), 3494–3504, https://doi.org/10.2174/138161207782794257.

[15] A. Tropsha, *Mol. Inform.* **2010**, *29* (6–7), 476–488, https://doi.org/10.1002/MINF.201000061.

[16] M. Rudrapal, D. Chetia, *J. Drug Deliv. Ther.* **2020**, *10* (4), 225–233, https://doi.org/10.22270/JDDT.V10I4.4218.

[17] N. Moitessier, P. Englebienne, D. Lee, J. Lawandi, C. R. Corbeil, *Br. J. Pharmacol.* **2008**, *153* (S1), S7–S26, https://doi.org/10.1038/SJ.BJP.0707515.

[18] J. Wang, R. M. Wolf, J. W. Caldwell, P. A. Kollman, D. A. Case, *J. Comput. Chem.* **2004**, *25* (9), 1157–1174, https://doi.org/10.1002/JCC.20035.

[19] R. A. Friesner, J. L. Banks, R. B. Murphy, T. A. Halgren, J. J. Klicic, D. T. Mainz, M. P. Repasky, E. H. Knoll, M. Shelley, J. K. Perry, D. E. Shaw, P. Francis, P. S. Shenkin, *J. Med. Chem.* **2004**, *47* (7), 1739–1749, https://doi.org/10.1021/JM0306430/SUPPL_FILE/JM0306430_S.PDF.

[20] O. Trott, A. J. Olson, *J. Comput. Chem.* **2010**, *31* (2), 455–461, https://doi.org/10.1002/JCC.21334.

[21] M. L. Verdonk, J. C. Cole, M. J. Hartshorn, C. W. Murray, R. D. Taylor, *Proteins Struct. Funct. Bioinforma.* **2003**, *52* (4), 609–623, https://doi.org/10.1002/PROT.10465.
[22] A. Jain, *Curr. Protein Pept. Sci.* **2006**, *7*, 407–420, https://doi.org/10.2174/138920306778559395.
[23] C. P. Mpamhanga, B. Chen, I. M. Mclay, P. Willett, *J. Chem. Inf. Model.* **2006**, *46* (2), 686–698, https://doi.org/10.1021/ci050420d.
[24] Q. Shen, B. Xiong, M. Zheng, X. Luo, C. Luo, X. Liu, Y. Du, J. Li, W. Zhu, J. Shen, H. Jiang, *ACS Publ.* **2011**, *51* (2), 386–397, https://doi.org/10.1021/ci100343j.
[25] S. Y. Huang, X. Zou, A *J. Comput. Chem.* **2006**, *27* (15), 1876–1882, https://doi.org/10.1002/JCC.20505.
[26] M. Xue, M. Zheng, B. Xiong, Y. Li, H. Jiang, J. Shen, *ACS Publ.* **2010**, *50* (8), 1378–1386, https://doi.org/10.1021/ci100182c.
[27] I. Muegge, *Perspect. Drug Discov. Des.* **2000**, *20*, 99–114, https://doi.org/10.1023/A:1008729005958.
[28] J. Spiegel, H. Senderowitz, *Int. J. Mol. Sci.* **2020**, *21* (21), 1–20, https://doi.org/10.3390/ijms21217828.
[29] G. Jones, P. Willett, R. C. Glen, A. R. Leach, R. Taylor, *J. Mol. Biol.* **1997**, *267* (3), 727–748, https://doi.org/10.1006/jmbi.1996.0897.
[30] O. Korb, T. Stützle, T. E. Exner, *J. Chem. Inf. Model.* **2009**, *49* (1), 84–96, https://doi.org/10.1021/CI800298Z/SUPPL_FILE/CI800298Z_SI_001.PDF.
[31] D. Wang, C. Cui, X. Ding, Z. Xiong, M. Zheng, X. Luo, H. Jiang, K. Chen, *Front. Pharmacol.* **2019**, *10* (JULY), 924, https://doi.org/10.3389/FPHAR.2019.00924/BIBTEX.
[32] D. Xu, L. Li, D. Zhou, D. Liu, A. Hudmon, S. O. Meroueh, *ChemMedChem* **2017**, *12* (9), 660–677, https://doi.org/10.1002/CMDC.201600636.
[33] V. P. Berishvili, A. E. Voronkov, E. V. Radchenko, V. A. Palyulin, *Mol. Inform.* **2018**, *37* (11), 1800030, https://doi.org/10.1002/MINF.201800030.
[34] L. Li, M. Khanna, I. Jo, F. Wang, N. M. Ashpole, A. Hudmon, S. O. Meroueh, *J. Chem. Inf. Model.* **2011**, *51* (4), 755–759, https://doi.org/10.1021/CI100490W/SUPPL_FILE/CI100490W_SI_002.PDF.
[35] W. J. Wang, Q. Huang, J. Zou, L. L. Li, S. Y. Yang, *Chem. Biol. Drug Des.* **2015**, *86* (1), 1–8, https://doi.org/10.1111/CBDD.12470.
[36] Y. Yan, W. Wang, Z. Sun, J. Z. H. Zhang, C. Ji, *J. Chem. Inf. Model.* **2017**, *57* (8), 1793–1806, https://doi.org/10.1021/ACS.JCIM.7B00017/SUPPL_FILE/CI7B00017_SI_001.PDF.
[37] I. Antes, C. Merkwirth, T. Lengauer, *J. Chem. Inf. Model.* **2005**, *45* (5), 1291–1302, https://doi.org/10.1021/CI050036G.
[38] M. M. Mysinger, M. Carchia, J. J. Irwin, B. K. Shoichet, *J. Med. Chem.* **2012**, *55* (14), 6582–6594, https://doi.org/10.1021/jm300687e.
[39] All Targets | DUD-E: A Database of Useful (Docking) Decoys – Enhanced http://dude.docking.org/targets (accessed July 17, 2021).
[40] J. Mistry, S. Chuguransky, L. Williams, M. Qureshi, G. A. Salazar, E. L. L. Sonnhammer, S. C. E. Tosatto, L. Paladin, S. Raj, L. J. Richardson, R. D. Finn, A. Bateman, *Nucleic Acids Res.* **2021**, *49* (D1), D412–D419, https://doi.org/10.1093/nar/gkaa913.
[41] Schrödinger Release 2020–2: LigPrep, Schrödinger, LLC, New York, NY, **2020**. Schrödinger, L. L. C.: New York, NY.
[42] G. Madhavi Sastry, M. Adzhigirey, T. Day, R. Annabhimoju, W. Sherman, *J. Comput. Aided. Mol. Des.* **2013**, *27* (3), 221–234, https://doi.org/10.1007/s10822-013-9644-8.
[43] J. Spiegel, H. Senderowitz, *Int. J. Mol. Sci.* **2021**, *23* (1), 43, https://doi.org/10.3390/IJMS23010043.
[44] D. M. Krüger, A. Evers, *ChemMedChem* **2010**, *5* (1), 148–158, https://doi.org/10.1002/CMDC.200900314.
[45] A. Evers, G. Hessler, H. Matter, T. Klabunde, *J. Med. Chem.* **2005**, *48* (17), 5448–5465, https://doi.org/10.1021/JM050090O.
[46] D. Callegari, D. Pala, L. Scalvini, M. Tognolini, M. Incerti, S. Rivara, M. Mor, A. Lodola, *Mol.* **2015**, *20* (9), 17132–17151, https://doi.org/10.3390/MOLECULES200917132.
[47] Validation of (Q)SAR Models – OECD https://www.oecd.org/chemicalsafety/risk-assessment/validationofqsarmodels.htm (accessed Jul 12, 2021).
[48] C. Nantasenamat, *Methods Pharmacol. Toxicol.* **2020**, 55–75, https://doi.org/10.1007/978-1-0716-0150-1_3.
[49] T. J. Hou, J. M. Wang, N. Liao, X. J. Xu, *J. Chem. Inf. Comput. Sci.* **1999**, *39* (5), 775–781, https://doi.org/10.1021/ci990010n.
[50] T. C. Le, D. A. Winkler, *Chem. Rev.* **2016**, *116* (10), 6107–6132, https://doi.org/10.1021/acs.chemrev.5b00691.
[51] V. Namasivayam, J. Bajorath, *J. Chem. Inf. Model.* **2012**, *52* (11), 2848–2855.https://doi.org/10.1021/ci300402 g.
[52] S. Saremi, S. Mirjalili, A. Lewis, *Adv. Eng. Softw.* **2017**, *105*, 30–47, https://doi.org/10.1016/J.ADVENGSOFT.2017.01.004.

Received: February 10, 2022
Accepted: July 5, 2022
Published online on July 26, 2022

