A Multivariate Approach to Investigate Docking Parameters' Effects on Docking Performance

C. David Andersson,[†] Elin Thysell,[†] Anton Lindström,[†] Max Bylesjö,[†] Florian Raubacher,[‡] and Anna Linusson*,[†]

Department of Chemistry, Umeå University, SE-901 87 Umeå, Sweden, and AstraZeneca R&D Mölndal, SE-431 83 Mölndal, Sweden

Received December 18, 2006

Increasingly powerful docking programs for analyzing and estimating the strength of protein−ligand interactions have been developed in recent decades, and they are now valuable tools in drug discovery. Software used to perform dockings relies on a number of parameters that affect various steps in the docking procedure. However, identifying the best choices of the settings for these parameters is often challenging. Therefore, the settings of the parameters are quite often left at their default values, even though scientists with long experience with a specific docking tool know that modifying certain parameters can improve the results. In the study presented here, we have used statistical experimental design and subsequent regression based on root-mean-square deviation values using partial least-square projections to latent structures (PLS) to scrutinize the effects of different parameters on the docking performance of two software packages: FRED and GOLD. Protein−ligand complexes with a high level of ligand diversity were selected from the PDBbind database for the study, using principal component analysis based on 1D and 2D descriptors, and space-filling design. The PLS models showed quantitative relationships between the docking parameters and the ability of the programs to reproduce the ligand crystallographic conformation. The PLS models also revealed which of the parameters and what parameter settings were important for the docking performance of the two programs. Furthermore, the variation in docking results obtained with specific parameter settings for different protein−ligand complexes in the diverse set examined indicates that there is great potential for optimizing the parameter settings for selected sets of proteins.

INTRODUCTION

Although systems with sufficient power for effective docking analysis have only been developed recently, they have already become valuable tools in drug discovery programs, and in several cases, docking has made a significant contribution to the discovery of new drugs.[1−3] Docking (essentially, the simulation of protein−ligand interactions based on three-dimensional structures) can provide valuable information on bioactive conformations, protein−ligand interaction points, and estimations of the binding strength of ligands that can greatly facilitate searches for new chemical leads and attempts to discriminate between similar compounds in known active series. Many general reports have been published on the subject,[4−8] and numerous studies have compared different docking software packages,[9−15] scoring functions, and approaches.[11,15−18]

There are two main approaches for evaluating docking performance. In one, the ability of the docking to generate the bioactive binding pose of the ligands is assessed by comparing the docked pose and a crystal structure of the target protein−ligand complex, for example, by calculating the root-mean-square deviation (RMSD) of the heavy atoms. In the other, enrichment studies and the rankings of ligands known to be active in a virtually screened database are used as measures. Both methods have certain advantages and disadvantages. The latter has advantages in its similarity to real case scenarios, where the desire often is to assess the probable interactions of new compounds for which no crystal structure is available. However, in some cases, an active ligand may be highly ranked for the wrong reasons; that is, the simulated binding mode for a specific ligand may have little similarity to the crystal structure of the modeled complex.[12,13,15] In such cases, RMSDs between the docked pose and the crystal structure provide more specific information about how a certain sampling and scoring method works (and flaws in the procedure) since the docked pose is compared directly with the pose of the crystal structure. An RMSD < 2 Å is commonly considered to indicate a successful docking.[19,20] However, this approach may also be misleading in some cases, since a docked ligand with a low RMSD value still can miss key interactions with the target protein that are present in the crystallized protein−ligand complex.[8]

The procedure involved in docking studies usually includes the following steps: (i) representation and definition of the binding site, (ii) conformational analysis of the ligands, (iii) placement of the ligands in the protein (generation of docking poses), (iv) estimation of the fitness of the poses (intermediate scoring), and (v) final scoring of the docked poses. The

* Corresponding author phone: +46-90-7866890; fax: +46-90-138885; e-mail: anna.linusson@chem.umu.se.
[†] Umeå University.
[‡] AstraZeneca R&D Mölndal.

10.1021/ci6005596 CCC: $37.00
Published on Web 06/09/2007

large numbers of currently available docking programs address these tasks in different ways, but the five steps are often controlled by the values of several parameters. In many cases, these parameters can be modified by the user to control the docking performance. However, neither understanding the influence of the different parameters nor finding the optimal settings for a specific case is straightforward. The settings are therefore frequently left at the default values set by the developers, even though modifying certain parameters may often improve the results.[10,21,22]

The traditional approach for investigating the influence of variables is to "tweak" one variable at a time until further changes give no further improvements. Although this may seem to be a systematic and straightforward method, it is often inefficient and, more importantly, it will not necessarily lead to the optimal settings. In order to find the most favorable conditions *and* elucidate the effects of the modifications, all parameters should be changed simultaneously in a controlled and systematic way, that is, with a statistical experimental design.[23] Such approaches have been applied in diverse fields, for example, drug discovery[24−26] and industrial pharmacy.[27,28] In molecular modeling, statistical experimental design has been successfully used to find parameter settings in OMEGA software that generate conformational ensembles of molecules that are likely to include bioactive conformations.[29] In addition, Antes et al.[30] have described an algorithm for parameter optimizations using ensemble methods (POEM) which combines the design of experiments with statistical learning techniques and use of the docking software FlexX. It has also been applied to DOCK, for which Salo et al.[21] suggested a stepwise parameter optimization procedure for virtual screens using Plackett−Burman and central composite designs to redock crystallographic ligands of a small number of protein−ligand complexes.

If the docking performance is to be controlled by changing the parameters, an important aspect to investigate is whether there are simple, linear relationships between alterations in the parameter settings and the reproduction of the crystallographic ligand conformations or whether the effects of such alterations are complex, yielding discontinuous "islands" of good poses. If the relationships are linear, or at least not too complex, it should be possible to obtain information on the extent to which a certain parameter influences the results and, hence, be able to interpret the effects of different docking parameters. In addition, a model could be used to evaluate how good a certain parameter setting is for specific protein−ligand complexes. In this study, we have used statistical experimental design and subsequent regression based on RMSD values using partial least-square projections to latent structures (PLS)[31] to investigate the effects of varying docking parameters on the performance of two docking programs. A general overview of the procedure can be seen in Figure 1. The combinations of parameter settings for the dockings were determined by statistical experimental design (Figure 1a, A-B), and a set of protein−ligand complexes with physicochemically diverse ligands was selected from the PDBbind database[32,33] using principal component analysis (PCA)[34,35] based on 1D and 2D descriptors and space-filling design[36] (Figure 1a, C-E). Two commonly used docking programs that address the five docking steps (see above) in fundamentally different ways

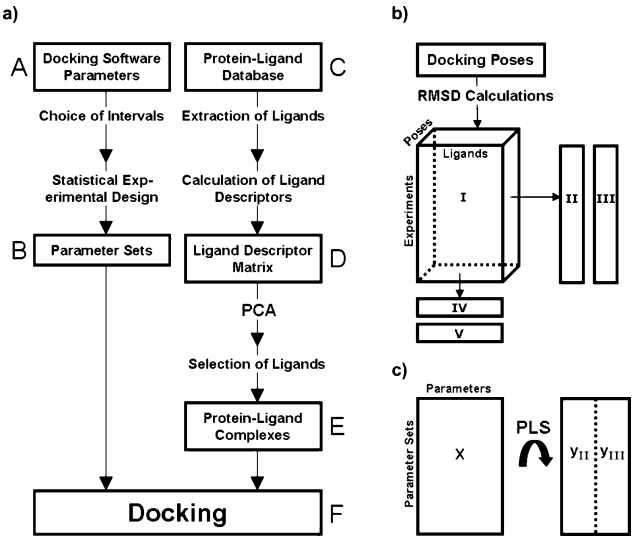

Figure 1. Schematic flowchart of the presented multivariate approach to evaluate docking parameters. (a) Docking parameters, their intervals, and a statistical experimental design were selected, and the design resulted in a number of well-balanced parameter sets (A−B). A set of protein−ligand complexes, with physicochemically diverse ligands, was selected (C−E), and the selected ligands were docked using FRED and GOLD with all of the designed parameter sets (F). (b) Postprocessing involved RMSD calculations between the docked poses and their corresponding crystallographic ligands resulting in matrix I and the extraction of RMSD medians for top poses (II) and best poses (III) across all ligands. Information on docking performance for individual ligands using individual best parameter settings was extracted for both the top pose (IV) and the best pose (V). (c) The parameter sets (matrix **X**) derived from A−B were used in PLS modeling and correlated to the top pose and best pose medians (y_{II}) and (y_{III}).

were selected for the investigation (Figure 1a, F): Fast Rigid Exhaustive Docking (FRED)[37] and Genetic Optimization for Ligand Docking (GOLD).[38−41] FRED relies on pregenerated low-energy conformations of the ligands that are rigidly fitted into a defined binding site and ranked by a scoring function. GOLD, in contrast, generates ligand conformations within the defined binding site in a stochastic manner using a genetic algorithm (GA) and a fitness function. The time consumed for one docking is generally shorter in FRED than in GOLD, but time consumption obviously depends on parameter settings, ligand size, choice of scoring function, and (for FRED) the number of simulated ligand conformations. This investigation was not a comparative study of the two software packages but rather an assessment of a general approach for investigating the effects of varying docking parameters. The generated docking poses' RMSDs to the corresponding crystal coordinates were calculated (Figure 1b), and the influence of the parameters on the docking performance was then evaluated, using PLS regression for the entire set of complexes (Figure 1c).

METHODS

Statistical Experimental Design. In this study, we were interested in analyzing how the docking performance was influenced by the different sets of docking parameters. The objective was to build a model which relates the parameter settings to a response reflecting the docking performance (i.e., the RMSD to the crystallographic coordinates) according to the Taylor expansion linear regression model:

$$y = \sum_{i=1}^{K} x_i b_i + \sum_{j=1}^{K} \sum_{k=1}^{K} x_j x_k b_{jk} + f \qquad (1)$$

where y is the RMSD value for a docking run, x is the parameter settings (values) for K parameters, b_i is the regression coefficient for linear terms, b_{jk} is the model coefficient for the square terms when $j = k$ and for the cross terms when $j \neq k$, and f is the residual. Statistical experimental design offered a possibility to vary all the investigated parameters simultaneously in such a way that the docking outputs could be used for modeling, resulting in minimum model errors in the coefficients both regarding noise and confounding. These coefficients were used to interpret the influence of the parameters on the docking performance (see more details below). Selected high and low values of quantitative parameters (i.e., parameters for which numerical values can be given) gave the range for the investigation and together with a center value constituted three possible levels in the design procedure. This range of values for the parameters' intervals was selected with the objective to be large enough to be able to capture an effect (if there is one) but still be appropriate for the regression model (i.e., Equation 1) and to ensure successful dockings. For parameters where the regression coefficients after modeling were shown to have low influence, two extra levels outside the range were investigated to ensure that the low impact of these parameters was not an effect of the selected ranges (see the Supporting Information). The docking results based on the parameter values from the two extended levels were compared with the results from the original high and low values respectively using the nonparametric and distribution-free Kolmogorov−Smirnov test[42,43] with a 0.05 level of significance. The different options for qualitative parameters (i.e., parameters where an option rather than a numerical value is used) were provided as indicator variables using zeros and ones (note that this was also used in the subsequent modeling).

The chosen docking parameters in FRED[37] and GOLD[41] were varied in a three-level full factorial design[23] and a D-optimal design,[44] respectively, using MODDE software[45] to ensure a good estimation of the model coefficients in eq 1 (Figure 1a, A-B). Full factorial design exhaustively enumerates combinations of all values for all selected levels and all options of the parameters. Applying a full factorial design to the GOLD GA parameters and annealing parameters was considered to lead to a docking study that would be too time-consuming. Hence, the D-optimal design with a Bayesian modification[46] was used. The candidate set (all combinations of the parameter values at the selected levels) used was expanded with cross and square terms (resulting in matrix **X**) in order to support a quadratic model with interaction terms. MODDE software[45] uses an automatic search algorithm, Federov's algorithm,[47] to select the number of wanted parameter sets (N) from the candidate set so that the N different parameter settings maximize the determinant of the $\mathbf{X}'\mathbf{X}$ matrix, that is, minimize the determinant of $(\mathbf{X}'\mathbf{X})^{-1}$ and hence minimize the model error of the coefficients in the subsequent regression model. Correlation matrices of the design terms (linear, square, and cross terms) were computed to investigate confounding patterns in the final selection of parameter sets, which is important since substantial confounding of terms makes it hard to interpret the importance of coefficients in the subsequent models. In this study, we considered correlations > 0.3 between design terms in the selected set of parameter settings to preclude detailed interpretation of the model coefficients.

Selection of Protein−Ligand Complexes. Protein−ligand complexes (Figure 1a, C-E) were selected from the refined version of the PDBbind database,[32,33] a database containing 3D information on 800 protein−ligand complexes as well as binding data. All of the complexes have been determined at sub-2.5-Å resolution, and only proteins binding ligands with molecular weights lower than 1000 are included. The strategy used to select the protein−ligand complexes was designed to generate a diverse subset of complexes by choosing physicochemically diverse ligands, assuming that their diversity would be mirrored by the diversity of the binding sites of the proteins. For each complex, the ligands were extracted and duplicates deleted and characterized by 1D and 2D descriptors, calculated using the Molecular Operating Environment (MOE) software,[48] representing physicochemical properties such as molecular weight, connectivity, and hydrophobicity (see the Supporting Information for a full list of descriptors). To ensure a sufficient spread of hydrophilicity/hydrophobicity, the data set was split into three groups, based on calculated lipophilicity (SlogP[49]) as defined by MOE software:[48] hydrophilic, intermediate, and hydrophobic ligands with SlogP values in the ranges −15 to −3, −3 to +3, and +3 to +9, respectively. PCA models[34,35] were calculated on the basis of the descriptor matrix for each of the three groups using SIMCA-P+ software.[50] Each descriptor variable was mean-centered and scaled to unit variance prior the PCA.

The protein−ligand complexes were selected using a space-filling design based on the distances between the ligands in the resulting principle component (PC) space of the three groups (hydrophilic, intermediate, and hydrophobic ligands). Ligands that proved to be extreme outliers according to the distance to the model in **X** (DmodX; see Validation of PCA and PLS Models section below) and/or score values were excluded. The specific method selects a predefined number of ligands through an iterative procedure to maximize the minimum Euclidean distance between currently selected pairs of ligands[36] and hence gives an even distribution of the selected objects, that is, the ligands. The score values resulting from the PCA based on the ligands' physicochemical properties were used in this design, and the algorithm was programmed in Matlab.[51]

Partial Least-Square Projections to Latent Structure. PLS regression was used to investigate the relationships between the docking parameters and the docking results. The PLS method maximizes the covariance between the latent variables of the **X** and **Y** matrixes (multi-**Y**) and correlates these latent variables through linear combinations to a regression model.[31] PLS models were developed from N observations (different docking runs based on a specific parameter set) and K variables (docking parameters) constituting a matrix **X** and N observations and M responses (represented by RMSD values) constituting a matrix **Y** (Figure 1c). The parameters having qualitative settings were represented in the model by one variable for each option; that is, for FRED, the parameter *score* was represented by three variables, one for each available scoring function, and

the parameter *neg_im* by three variables, one for each size option, while the parameters with quantitative settings were represented with one variable for each parameter (cf. the design variables). One advantage with PLS is the possibility to incorporate more than one response variable in the modeling procedure (multiple **Y**). Here, we have used two measures to describe the docking performance (see details below) in the development of the models. Before the regression, each variable in **X** was scaled to unit variance and mean-centered.

The PLS regression coefficients as described by eq 1 were used for interpretation of the models. A large positive or negative coefficient value for a certain model term, that is, docking parameter, shows that it has strong influence over the model and indicates that it is important for the relationship between **X** and the docking results **Y**. Hence, quantitative parameters with large negative coefficient values should have high parameter values, while coefficients with large positive values should have low values for optimal docking performance (resulting in low RMSD values). Correspondingly, negative coefficient values for a specific qualitative setting reveal that docking runs using these options lead to lower RMSD values. Nonsignificant coefficients, according to PLS coefficient plots, of square and cross terms were iteratively deleted from the models, and the model was recreated until all remaining coefficients were significant. Model terms were excluded on the basis of their respective coefficient significance with a 95% confidence interval calculated by jack-knifing[52] using SIMCA-P+ software.[50] No linear terms of the docking parameters were deleted, even if they proved to be nonsignificant. In addition to the coefficient plots (i.e., PLS regression coefficients corresponding to the centered and scaled **X** and the scaled but not centered **Y**), the score and loading plots were also used to interpret the PLS models.

Validation of PCA and PLS Models. The significance of the PCs generated by the PCA was determined by two different statistical tools. A component was considered significant if its eigenvalue was larger than 2.0 and had a Q^2 value for the individual component greater than 0.05. For the cross-validated Q^2 values,[34,53] the "leave-many-out" method was used, in which a seventh of the data set was excluded from the data set in an iterative manner, for generating a new model where the excluded part was predicted. The overall significance of the PCA models was determined by inspecting the cumulative R^2X and Q^2 values, that is, the cumulative sum of the variation in all **X** variables explained by all extracted PCs and the cumulative Q^2 for all extracted PCs. To identify observations not covered by the PCA model, DmodX[50] was used.

A PLS component was considered significant if the Q^2 value was larger than zero. The overall significance of the model and its predictive power were evaluated by inspecting the cumulative values of R^2Y and Q^2 and docking runs not covered by the PLS model; that is, outliers were investigated by DmodX. Furthermore, plots of observed responses versus calculated responses were studied and root-mean-square error of estimation (RMSEE) values were calculated. In addition, PLS models were investigated using a permutation methodology in which the order of values in **y** is randomized and a new model is calculated.[54,55] This was repeated 200 times for the two variables in **Y**, and plots were generated showing the correlation coefficients between the original **y** and the permuted **y** versus the cumulative R^2 and Q^2 values. The intercept (R^2 and Q^2 when the correlation coefficient is zero) is a measure of the fit, that is, the significance of the model. Intercept values for Q^2 values less than zero and low intercept values on R^2 indicate a nonsignificant model, which is desirable for a PLS model with randomized **y**. Furthermore, two sets of protein−ligand complexes were processed separately throughout the study as a validation procedure to assess the similarities of the results obtained with two independent regression models describing how docking parameters influence the docking performance. All calculations were performed with SIMCA-P+ software.[50]

Preparation of Proteins and Ligands. The protein structures were prepared by semiautomatic procedures as described by Lindström et al.[56] Water coordinates were removed from the Protein Data Bank (PDB) files but were added again for complexes 1EX8, 1M1B, 1GJ9, and 2XIS to enable the metal coordination geometry recognition in GOLD.[41]

The ligands were represented as chiral SMILES[57] using the BABEL[58] file format converter, and 3D structures were then regenerated with CORINA software.[59] Atom and bond types for the ligands were set according to GOLD specifications using the SYBYL software.[60] The ionization states of functional groups in the ligands were calculated at physiological pH using the MARVIN software.[61] Questionable ionization states and tautomers were examined by manual inspection of the crystallographic structures, taking into account the protein environment around such functional groups. If uncertainty still remained about ionization states and tautomerism of the ligands, several versions of the ligands were docked.

Docking in FRED. The conformational search of the ligands was performed using OMEGA.[62] This software uses a rule-based rotatable bond search and a template-based ring conformer mapping to produce a conformational ensemble. The default parameter values were used in OMEGA with two exceptions; the maximal number of low-energy conformations generated for each ligand was set to 200, and the maximum number of allowed rotatable bonds was set to 50. The crystallographic ligand conformations were not added to the ensemble in order to mimic a real-case scenario.

FRED rigidly docks the pregenerated conformations of each ligand into a nonflexible target protein. The software employs a shape-fitting algorithm in the docking step in concert with an exhaustive systematic search by translation and rotation of the conformations and then ranks the resulting poses. Four parameters controlling the docking procedure (three quantitative and one qualitative) and one determining the choice of scoring function for the exhaustive search were selected for the study (Table 1). The parameter Clash Checking specifies the threshold at which the fraction of the sum of overlapping radii between the ligand and the protein's heavy atoms represents a clash. This information is used to create the negative image used to describe the binding site, and the qualitative parameter Negative Image Size governs the overall size of the negative image by controlling the distance that atoms are allowed to be placed from the protein with the options small, normal, and large. The parameters Translational Step Size and Rotational Step Size determine the placement of the ligand within the negative image by

Table 1. Settings of FRED Docking Parameters Used in the Study

		settings[a]		
parameter	abbreviation	1	2[b]	3
exhaustive scoring	score	Plp	Chemgauss	Shapegauss
negative image size	neg_im	small	normal	large
clash checking	clash	0.50	0.75	1
translational step size	tstep	0.75	1	1.25
rotational step size	rstep	0.75	1	1.25

[a] For parameters *clash*, *tstep*, and *rstep*, setting nos. 1 and 3 represent the high and low levels in the design, while setting no. 2 corresponds to the center level. [b] Setting no. 2 represents the software parameter default setting for the specific parameters.

moving and rotating the ligand. For the exhaustive search, the three available scoring functions were used: Piecewise Linear Potential (Plp),[63] Shapegauss,[64] and Chemgauss.[37] The docking region in FRED was represented by a hexahedron (box) enclosing the binding site of the protein. The box dimensions were determined by automated calculations using the crystallographic ligand coordinates. The Euclidean distances for every ligand atom to the mean coordinates of the ligand atoms were calculated in angstroms. The size of the box for each individual ligand was defined by letting the largest Euclidean distance plus 4.0 Å be the distance between the origin and a corner of the hexahedron. The conformational ensemble for each ligand was docked into the relevant protein using the parameter sets defined by the designed docking experiment protocol for FRED. The scoring of the docked poses during the exhaustive search was also used as final scoring without any postrefinement of the poses, and the 15 top-ranked poses were extracted.

Docking in GOLD. In GOLD, the ligand is initially placed in the protein binding site on the basis of fitting points; then, a GA is used to explore the conformational space of the ligands on the basis of fitness scores,[39] as follows. The initial population of possible ligand docking poses is set up at random. Each member of the population is encoded as a chromosome, which contains information on the binding between the ligand and the protein, for example, mapping of H-bonding atoms, hydrophobic interactions, and conformations around rotatable ligand bonds. Each chromosome is assigned a fitness score and ranked accordingly. The population is iteratively changed by point mutations and mating of the chromosomes to maximize the fitness score. Eight parameters controlling the GA (the Number of Operations, Selection Pressure, Population Size, Number of Islands, Niche Size, Crossover, Migrate, and Mutate) and two annealing parameters (Van der Waals and Hydrogen Bonding) were selected for the study, resulting in 10 quantitative parameters for the design (Table 2). The two annealing parameters control the clash and hydrogen-bond distance acceptance, allowing poor van der Waals interactions and hydrogen bonds at the start of the GA. The force-field-based Goldscore scoring function was used in the study.[39] During the docking process, the ligand and hydrogens of the protein were regarded as being flexible. The docking region in GOLD was enclosed by a sphere centered on the center of gravity of the crystallographic ligand. The sphere-dimensions were calculated automatically from the crystallographic ligand coordinates in a procedure similar to that described for FRED (see above). Each ligand was docked into the relevant protein 15 times (i.e., in 15 GA runs) using the designed parameter sets, resulting in 15 docking poses for each parameter set.

Table 2. Settings of the GOLD Genetic Algorithm and Annealing Parameters Used in the Study

		settings			
parameter	abbreviation	level 1	level 2	level 3	default[a]
population size	popsize	50	75	100	100
number of islands	num_isl	1	3	5	5
niche size	nichesize	1	2	3	2
number of operations	num_ops	1000	50 500	100 000	100 000
mutate	mutate	95	97.5	100	95
crossover	cross	95	97.5	100	95
migrate	migrate	0	5	10	10
selection pressure	selpress	1.1	1.1125	1.125	1.1
hydrogen bonding[b]	Hbond	2	3.75	5	4
van der Waals[b]	vdWaal	4	7	10	2.5

[a] This is the standard default setting, which is the one of five choices of default settings that is the most accurate according to the vendors. [b] These are annealing parameters and only available if the Goldscore scoring function is used.

Evaluation of Docking Poses. Docking of the selected ligands using the designed combinations of parameters resulted in a large number of docking poses that can be displayed as a 3D plot, in which the designed parameter sets are located on the *x* axis, the docked protein−ligand complexes on the *y* axis, and the 15 extracted docked poses on the *z* axis (Figure 1b). In this study, we evaluated the poses using their RMSD values to obtain quantitative numerical values that were applicable to all of the complexes. For the two programs used, we extracted two ligand poses for each complex and each parameter set from this three-way plot: the docked pose that was most highly ranked by the docking program, denoted the "top pose", and the docking pose that, according to RMSD, was most similar to the crystallographic ligand structure, denoted the "best pose". The RMSD values were calculated using the OEChem Toolkit.[65] The extraction yielded two matrices with RMSD values of *L* docked ligands with *N* different parameter sets of top pose and best pose, respectively. From these result matrices, the median RMSDs for the top poses and best poses for all of the complexes generated by each of the two programs were calculated [Figure 1b (II and III)]. These two RMSD median vectors were used as responses in the PLS modeling to evaluate the effects of the docking parameters on the docking performance (Figure 1c). In addition, the best docking pose, according to the RMSD values, for the top pose and best pose were extracted for each complex over all parameter sets to investigate the optimization potential for individual protein−ligand complexes when modifying parameters [Figure 1b (IV and V)].

RESULTS AND DISCUSSION

Statistical Experimental Design of Docking Parameters. The full factorial design of the FRED parameters at the three levels for the quantitative parameters and all options for the qualitative parameters (Table 1) resulted in 243 different parameter sets (see the Supporting Information for a complete list). The parameter values at the three levels in the design for the three quantitative parameters studied in FRED were chosen so that the default settings specified in the software manual represented the center level (Table 1).

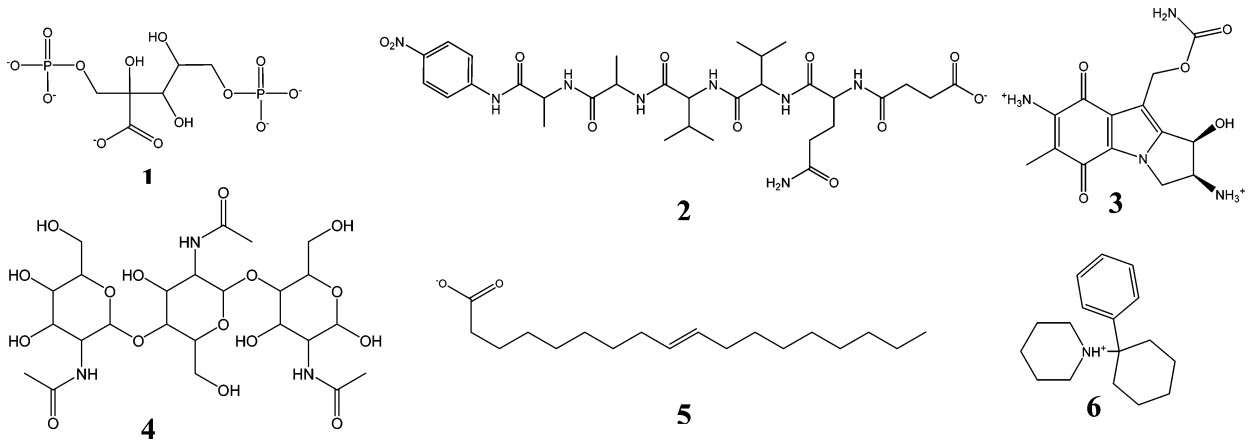

Figure 2. Examples of ligands present in the data sets illustrating their physicochemical diversity. **1**: 1RBO, MW = 351, SLogP = −8.0, rotatable bonds = 12. **2**: 1PIP, MW = 706, SLogP = −1.9, rotatable bonds = 33. **3**: 1KLL, MW = 322, SLogP = −1.5, rotatable bonds = 10. **4**: 1AT6, MW = 628, SLogP = −7.1, rotatable bonds = 24. **5**: 1HMS, MW = 281, SLogP = 4.8, rotatable bonds = 16. **6**: 2PCP, MW = 244, SLogP = 4.6, rotatable bonds = 2. SLogP and number of rotatable bonds are calculated as defined by MOE.[48]

For GOLD, the selected parameter values for the three levels in the design of the eight GOLD GA parameters and two annealing parameters are displayed together with the default settings in Table 2. The three-level D-optimal design used to select parameter settings of the 10 GOLD parameters resulted in 125 parameter sets. The standard default setting available in GOLD was also added, to give a total of 126 parameter sets (see the Supporting Information for a complete list). There was no confounding between the linear, cross, and square terms according to the correlation matrix, showing that these 126 parameter sets represented a qualitative selection for subsequent evaluation of the parameters' effect by a quadratic regression model.

Protein−Ligand Complexes. PCA based on the 1D and 2D descriptors of the ligands from the PDBbind database,[32,33] divided into three groups based on their lipophilicity, resulted in three models describing 80−89% of the original variation by four PCs (score vectors). The cross-validation gave Q^2 values of 0.71−0.85. Hence, the individual score values for the ligands constituted a four-dimensional space from which the selection was made using a space-filling algorithm. For validation purposes, two sets of protein−ligand complexes were selected sequentially; the first (set 1) consisted of 39 complexes[66] and the second (set 2) of 29 complexes.[67] The two sets were processed separately throughout the docking process to assess the similarities of the patterns obtained with two independent sets of protein−ligand complexes. Since the selected sets of protein−ligand complexes were selected to be physicochemically diverse, many ligand properties in the chosen sets extend far beyond the "druglike" ranges defined by Lipinski et al.[68] For example, ligands with up to 50 rotatable bonds were selected as well as ligands with weights up to 900 and logP values between −7 and +9. Sets 1 and 2 were equally diverse with respect to ligand properties, and some representatives are shown in Figure 2.

Docking Evaluation of FRED. In FRED, the pregenerated conformations for each of the 68 ligands (set 1 containing 39 complexes and set 2 containing 29 complexes) were docked back into the corresponding proteins using the 243 different parameter sets, where 15 docked poses were saved for each of the parameter sets (244 215 docking poses in total). For two ligands (1AJ6 and 1KLL), no RMSD values were calculated due to significantly different docking poses compared to the crystallographic ligand. Close examination of the poses of the 1KLL ligand revealed that it occurred at a 90° angle compared to the crystallographic ligand, thereby missing important π-stacking interactions with a histidine ring which are present in the crystal structure. Furthermore, for a few ligands, certain parameter sets did not result in any docking solutions. Most such cases concerned very large ligands in combination with parameter sets where the Negative Image Size was set to small and the Clash Checking value was high. This combination of settings resulted in a too-small negative image which, in turn, resulted in all of the poses being discarded since they simply could not fit within the negative image.

The RMSD values for the top pose (the most highly ranked pose) and best pose (the pose most similar to the crystallographic ligand) for 66 ligands and the 243 parameter sets were extracted, and the RMSD medians of the ligands for each parameter set were calculated and used for PLS modeling (see separate section below). The top and best poses with the lowest RMSD value for each ligand out of all parameter sets were also extracted together with the top and best poses from the default parameter setting (Figure 3). The RMSD values for the top pose obtained using FRED and the default parameter settings were < 2.0 Å for 17 of the 66 docked ligands (26%). The corresponding proportion for the best pose was 29 out of 66 (44%). It is important to note that the pregenerated ensemble of ligand conformations determines the conformational search space of the ligands, and thus, the correct binding mode of the ligands can only be identified if the bioactive conformation is present in the ensemble. Hence, the number of successful dockings may increase if the number of ligand conformations used in FRED is increased.

As indicated by the PLS model presented below, the default parameter settings appear to be a good parameter set for the full, diverse set of complexes used in this study. However, these proportions were increased by an additional 3% for the top pose and 5% for the best pose when the best-performing set of parameters from the designed sets was used. This high-performing parameter set was also consistent with the PLS model (see below), indicating that its parameter settings (with the parameter *score* set to Chemgauss, *neg_im* to large, *clash* to 0.5, *tstep* to 1.25, and *rstep* to 1) may be

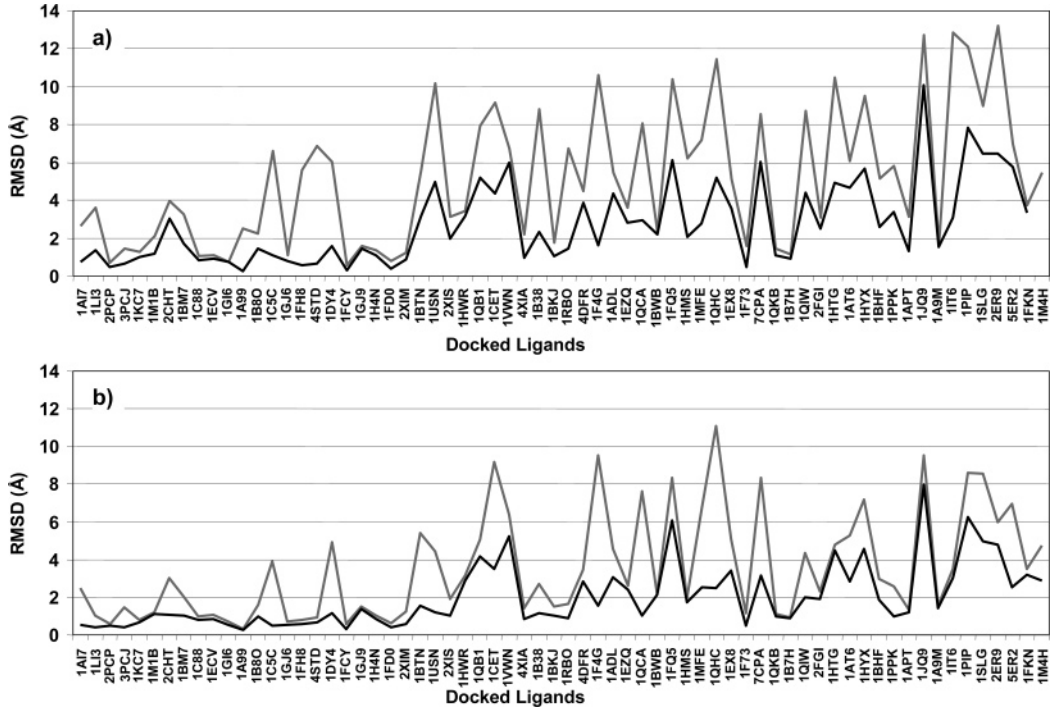

Figure 3. Docking results obtained using FRED for individual protein−ligand complexes presented as RMSD values for the top pose (a) and the best pose (b) of the docked ligands. Ligands are ordered from left to right according to the number of rotatable bonds present in them. Results when the default settings were used are shown by a gray line, and the docking results based on the best parameter sets for each complex are shown as a black line.

somewhat more robust than the default settings for our set of ligands.

It was not surprising that this well-established docking software has default settings that proved to be close to optimal for a diverse set of complexes. A more interesting facet to investigate was the potential scope for using different parameter settings for specific kinds of protein−ligand complexes. Figure 3 gives an overview of the docking performance for individual complexes showing the RMSD values for both their top poses (Figure 3a) and best poses (Figure 3b) and a comparison of their default settings and the parameter settings that resulted in the most accurately reproduced binding modes. The data show that the docking can be sufficiently improved for 15 ligands (1AI7, 1APT, 1A99, 1B8O, 1BM7, 1C5C, 1DY4, 1F4G, 1FH8, 1LI3, 1M1B, 1RB0, 2XIS, 4STD, and 4XIA) by using individual parameter settings rather than the default set to reduce the RMSDs for the top pose from > 2.0 Å to < 2.0 Å. Furthermore, using individual parameter settings rather than default settings also yielded substantial improvements for many ligands that never docked with RMSD values < 2.0 Å. In addition, although FRED successfully docked ligands with as many as 31 rotatable bonds (1A9M) with RMSDs < 2.0 Å compared to the crystallographic ligand, the program was generally less likely to successfully dock ligands with more than 15 rotatable bonds than smaller ligands. Hence, improvements were seen for ligands of all sizes, but the docking performance was generally better for relatively small ligands, which was probably due to limitations of the ligand conformational ensemble. The trends are similar for the best pose, with an additional 23% of the ligands docked < 2.0 Å when modified parameter sets were used (Figure 3b).

Closer examination of the poor dockings revealed that failure commonly occurred when the ligands in the crystallographic structure had large and/or hydrophobic groups projecting into the solvent (e.g., 1CET, 1QB1, and 1VWN), probably because FRED has a known tendency to position ligands to match the shape of the protein surface and hence does not generally handle such poses well. Ligands with symmetric features also caused problems for the docking software. In a few cases (e.g., 1FQ5 and 1HWR), these kinds of molecules were docked at a 180° angle to the crystallographic binding mode, resulting in high RMSD values. Kellenberger et al.[10] found that docking using FRED commonly fails for small, hydrophilic, deeply buried ligands. However, we did not see such trends; docking results were generally acceptable for small, hydrophilic, deeply or partially buried molecules (e.g., 1A99, 1B8O, 1H4N, 1M1B, 2XIS, and 4XIA), especially when individually best parameter settings were used.

Docking Evaluation of GOLD. In GOLD, each of the 68 ligand structures were docked into their corresponding proteins in two batches (set 1 followed by set 2) using the 126 designed parameter sets and Goldscore as the scoring function. GOLD was able to dock 38 of the 39 docked ligands in set 1 and all 29 of the ligands in set 2. The use of 15 GA runs gave 15 poses for each parameter set (126 630 poses in total). The failed docking in set 1, 2XIS, was a magnesium-containing complex where the metal geometry recognition failed despite attempts to leave appropriate water molecules in the protein to correct the problem. Furthermore, RMSD calculations were not completed for two complexes: 1AJ6 in set 1 and 1KLL in set 2 for similar reason as for dockings with FRED. In the former case, the ligand in 1AJ6, the program did not find the crucial hydrogen-bond interactions present in the crystal structure. A possible explanation for this is that a sequestered water molecule that is important for binding to the protein of 1AJ6 had been removed in the

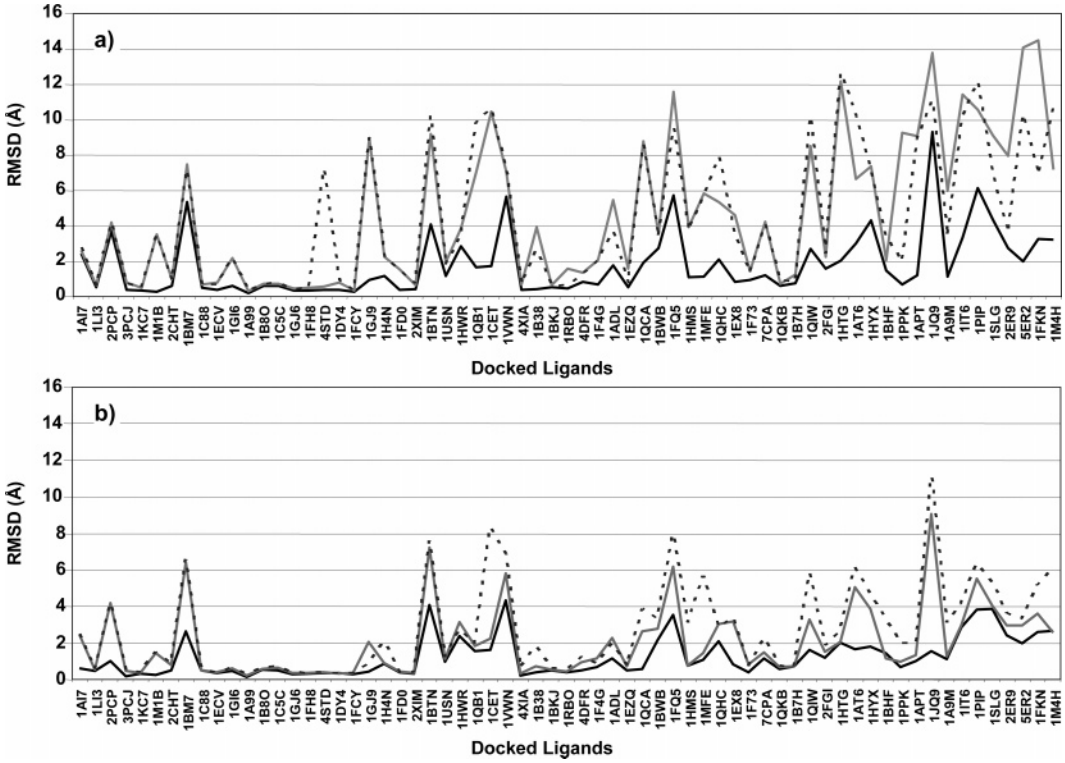

Figure 4. Docking results obtained using GOLD for individual protein−ligand complexes presented as RSMD values for the top pose (a) and the best pose (b) of the docked ligands. Ligands are ordered from left to right according to the number of rotatable bonds present in them. Results when the default settings with 15 and 100 GA runs were used are marked by dotted and solid gray lines, respectively. The docking results based on the best parameter sets for each complex are shown as a black line.

preprocessing of the protein. For 1KLL, GOLD generated docking poses in another area of the binding pocket, 10 Å away from the binding site of the crystallographic ligand.

The RMSD values for the top pose and best pose for 65 ligands and the 126 parameter sets were extracted, and the RMSD medians of the ligands for each parameter set were calculated and used for PLS modeling (see separate section below). The top and best poses with the lowest RMSD value for each ligand out of all parameter sets were also extracted together with the top and best poses from the default parameter setting (Figure 4).

The GA in GOLD introduces an element of haphazardness in the docking; hence, multiple dockings under identical conditions may not necessarily give identical results,[39,69] which could complicate the interpretation of the docking results. In a study by Jones et al.,[39] the outcomes obtained with 2, 5, 10, and 20 GA runs were compared, and the cited authors concluded that GOLD generally requires less than 20 GA runs to reproduce a binding mode (note that the default setting is 10 GA runs). Thomas and co-workers[69] performed a study in which 132 CDK2 inhibitors were docked into 10 different proteins five times using 10 GA runs and two default settings (standard mode and library screening mode) and concluded an existing but not substantial difference in reproducibility. In our study, we chose to run the GOLD dockings with 15 GA runs, which seemed appropriate, taking into account the results from the studies by Jones and Thomas and their respective co-workers.

An investigation was performed in which the 65 ligands for which RMSD values had been computed were docked with 15 and 100 GA runs, while all the other parameters were kept at their standard default settings. The results, shown in Figure 4, indicate that the number of ligands that docked with < 2.0 Å RMSDs to the corresponding crystallographic ligands did not increase when 100 GA runs were used to determine the top pose (Figure 4a), where both 15 and 100 GA runs resulted in poses with < 2.0 Å RMSDs for 40% of the ligands. However, 100 GA runs gave best poses with < 2.0 Å RMSDs for a significantly higher number of ligands than 15 GA runs: 60% and 54%, respectively (Figure 4b). This way of analyzing data only takes into account the ligands that are prone to dock with a RMSD under 2.0 Å, and for these ligands, the reproducibility was fairly consistent. However, for ligands that were never docked with a RMSD lower than 2.0 Å, the results were less consistent. The differences in the outcome between 15 and 100 GA runs were higher for ligands with more than 10 rotatable bonds (i.e., from 1QB1 in Figure 4). This reflects the problem with reproducibility for large flexible ligands. Clearly, poor reproducibility can significantly decrease the information/noise ratio in calculated RMSD values, so further investigations regarding reproducibility are warranted.

The number of successfully docked poses in our set of protein−ligand complexes (see above) is lower than previously presented proportions in the literature.[10,11,14] However, in a recent study by Warren et al.,[15] a large variation in success rates of docking performance was seen for different software packages depending on the type of protein that the ligands were docked into. The relatively low rates here are probably due to the large diversity in our set of protein−ligand complexes, including ligands with distinctly nondrug-like properties.

As stated earlier, comparisons of different parameter sets are complex, due to the stochastic nature of the algorithm. Thus, parameter sets that truly improved docking performance, that is, that performed better than default settings, can only be identified using RMSDs in cases where none of the runs with default settings (including both 15 and 100 GA runs) performed better than or as well as the alternative settings, and even then, uncertainty remains because of the poor reproducibility for certain ligands. We cannot identify a general set of parameters that clearly outperforms the default settings for the total set of diverse complexes. However, large variations were seen in this respect between specific complexes. A comparison between the default setting outcomes (top pose and best pose RMSDs) with 15 and 100 GA runs and those obtained with best individual parameter sets[70] for each complex can be seen in Figure 4. For 26% of the docked ligands—1A9M, 1ADL, 1APT, 1B38, 1CET, 1EX8, 1GI6, 1GJ6, 1H4N, 1HMS, 1M1B, 1MFE, 1QB1, 1QCA, 2FGI, 5ER2, and 7CPA—top poses with RMSDs < 2.0 Å were obtained using an individual parameter set, while default settings with both 15 and 100 GA runs resulted in poses with RMSDs > 2.0 Å (Figure 4a). This means that an additional 26% (to the 40% with default settings) of the ligands were docked with RMSDs < 2.0 Å when parameter sets other than the default settings were used. Furthermore, another 17% of the ligands—1AT6, 1BTN, 1FKN, 1FQ5, 1HYX, 1M4H, 1PIP, 1QHC, 1QIW, 1VWN, and 2ER9—were docked with RMSDs that were lower (but not < 2.0 Å) when individual parameter sets were used rather than the default set. The corresponding percentages for the best pose were 12% and 14%, respectively (Figure 4b). The difference in results between the parameter settings (Figure 4) seems to increase with the number of rotatable bonds in the ligands, which probably reflects the poor reproducibility for large flexible ligands. Nevertheless, despite the difficulties in interpreting the results, the large improvements obtained when using individually best parameter sets indicates that there is interesting potential for optimizing parameters for specific sets of protein−ligand complexes.

In the study by Perola and colleagues,[11] a general tendency for the performance of GOLD to decline as the flexibility of the ligands increased was detected. This was confirmed in our study, since good poses were generally produced for ligands with fewer than 15 rotatable bonds. No clear trends were seen between the hydropobicity (SLogP) of the ligands and the docking outcome, but the docking performance was generally best for ligands with SLogP values between −5 and +1. Successfully docked ligands were often small and rigid and contained hydrogen-bond donors and acceptors, confirming the assertion by Jones et al.[39] that GOLD most reliably docks hydrophilic ligands with hydrogen-bonding capabilities. The ligands for which GOLD was not able to produce poses with RMSDs < 2.0 Å included large peptides, circular peptides, and peptidelike ligands. Although poses were generated with correct interactions between the ligands and proteins in such cases, they were not all present in one single docking pose. In the study by Kellenberger et al.,[10] it was found that GOLD generally did not successfully dock small, hydrophobic ligands. In our case, nine of the 17 ligands in the data set that had a molecular weight below 500 and a SLogP value greater than 2 were successfully docked (RMSD < 2.0 Å for the top pose) with the default settings (with 15 and 100 GA runs). Of the eight failed ligands, six had a molecular weight below 320, supporting the findings by Kellenberger et al.[10]

Table 3. Model Quality Information for FRED and GOLD PLS Models for the Two Independent Sets of Complexes (Sets 1 and 2)

	FRED		GOLD	
PLS model terms	set 1	set 2	set 1	set 2
R^2Y	0.83	0.85	0.74	0.63
Q^2	0.79	0.84	0.67	0.55
RMSEE, top pose	0.64	0.57	0.59	0.70
RMSEE, best pose	0.36	0.43	0.25	0.32
R^2 (top pose), permuted $\mathbf{y}^a$	0.02	0.02	0.08	0.09
Q^2 (top pose), permuted $\mathbf{y}^a$	−0.33	−0.20	−0.23	−0.22
R^2 (best pose), permuted $\mathbf{y}^a$	0.02	0.01	0.07	0.08
Q^2 (best pose), permuted $\mathbf{y}^a$	−0.33	−0.21	−0.25	−0.23

[a] Results when the permuted **y** has no correlation with the response **y**.

Multivariate Evaluation of the Effect of Docking Parameters. PLS regression models were constructed to investigate whether differences in RMSD values for the docked poses could be steered by the changes in parameter settings, or if the effects of such changes are too complex to be modeled by such a relationship (cf. eq 1). The use of two independent sets of complexes (set 1 and set 2) allowed the method to be validated by constructing two models for each investigated program and comparing the similarity of the patterns in their results. Hence, the median RMSD values used as response values in the modeling were calculated across the two sets of complexes separately. For FRED, the PLS regression models were constructed using the parameter values for the 243 parameter sets as the **X** matrix and the median RMSD values of the corresponding docking runs for the best pose and top pose (*BP Median* and *TP Median*) as **Y** values. Correspondingly, for GOLD, the settings for the 10 investigated parameters in the 126 different parameter sets were used as the **X** matrix with the median RMSD values of the top and best poses for the two sets as **Y** (Figure 1c). For both GOLD and FRED, significant models were obtained with similar R^2Y, Q^2, RMSEE, and permutation tests' values for sets 1 and 2, implying that there were indeed quantitative relationships between the docking parameters and the RMSD values (Table 3). The lower explained variation value (R^2Y) for GOLD compared to that of FRED was probably due to increased noise levels resulting from the lower reproducibility associated with using GA. In order to determine if the docking parameters had the same influence on the docking results for both sets of ligands, the PLS regression coefficients for the set 1 and set 2 models were plotted (see the Supporting Information) and the goodness of fit values for the resulting linear regressions were calculated. The PLS regression coefficients showed strong correlations for both responses and both programs, the R^2 values for FRED being 0.77 and 0.88 for the top pose and best pose, respectively, and the corresponding values for GOLD being 0.87 and 0.86, respectively. These findings indicate that the same docking parameters influenced the docking performance for the two different ligand sets in the same way.

The constructed models showed that two independent sets of protein−ligand complexes displayed a simple quantitative relationship between modifications of the docking parameters and the docking performance represented as RMSD to the

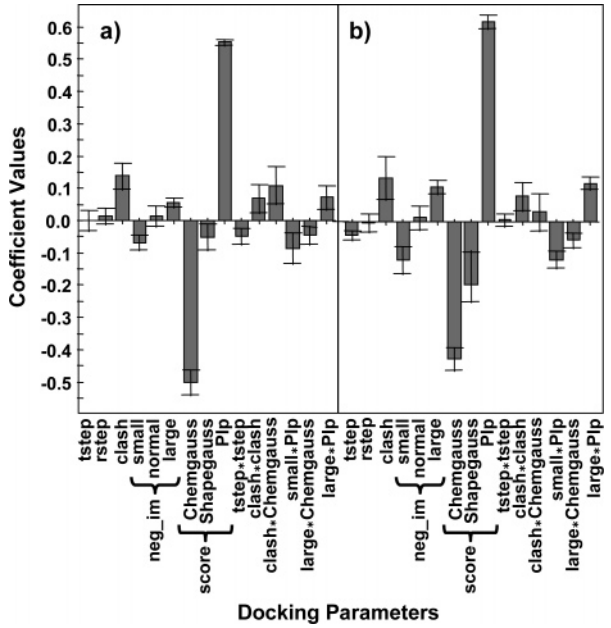

Figure 5. PLS coefficient plot showing the FRED docking parameters' influence on the median RMSDs for the best pose (a) and the top pose (b) on the basis of the parameter span displayed in Table 1. The qualitative parameters *neg_im* and *score* are represented by three variables, one for each option. The confidence intervals for each coefficient are also displayed.

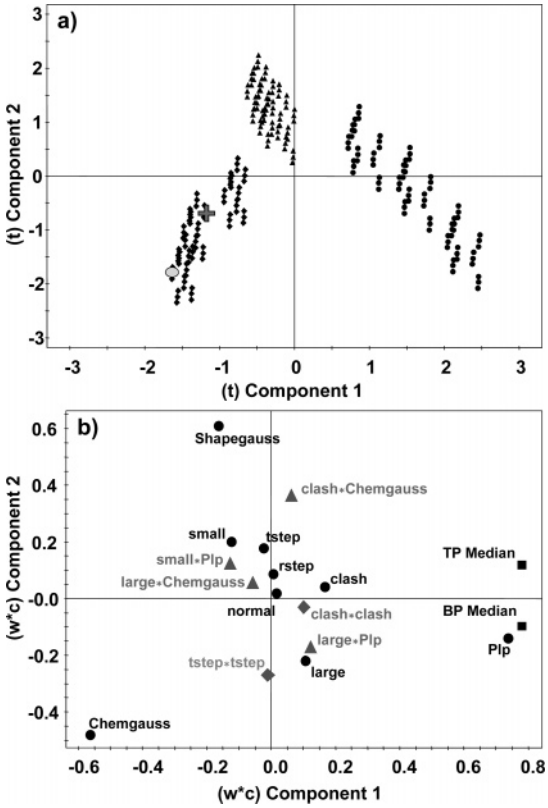

Figure 6. Score (a) and loading (b) plots of the PLS model based on the FRED dockings. The 243 designed parameter sets can be seen in the score plot (a), where the markers illustrate the selection of an exhaustive scoring function (dots, Plp; triangles, Shapegauss; diamonds, Chemgauss). The default parameter set is marked with a cross, and the better performing parameter set is marked with an ellipse. In the PLS loading plot (b), linear terms of the docking parameters are marked with dots, square terms with diamonds, cross terms with triangles, and responses with squares.

crystallographic ligand and that modifying the docking parameters had similar effects for the two sets. These results prompted a more detailed study of the relationships between the parameters and the docking results for the total set of complexes using FRED and GOLD, as discussed below.

Interpretation of the Effects of Docking Parameters in FRED. The PLS model based on the docking results using the 243 designed parameter sets had two significant components describing 91% ($R^2Y = 0.91$) of the variation in the median RMSD values for the total set of 66 complexes, with most of the described variation in the first component. The cross-validation resulted in a Q^2 value of 0.90. The calculated estimation error (RMSEE) was 0.33 Å, and the permutation test gave an $R^2 < 0.05$ and a negative Q^2 for both the top pose and best pose medians, further verifying the significance of the model (plots of observed versus calculated RMSD values and plots of permutation tests can be seen in the Supporting Information). No docking runs were detected as outliers.

The choices of scoring function for the exhaustive search followed by the settings for Clash Checking were the parameter settings that had the greatest impact on docking performance in FRED according to the PLS model (Figure 5). These parameters (*score* and *clash*) had the most dominating regression coefficients, and the signs of these coefficient values show that selecting Chemgauss as the scoring function and low values on the Clash Checking parameter resulted in better docking performance (Figure 5). The use of Plp as the scoring function during the exhaustive search had a negative effect on the docking performance for this set of ligands as indicated by its high positive coefficient values (Figure 5). The coefficient plots show that the trends were similar for both the top pose and best pose. A detailed interpretation of the regression coefficients for the interaction terms reveals that large negative images for binding sites seem to be favorable when using Chemgauss as the scoring function, while the images should be kept small when using Plp. It is somewhat surprising that *tstep* and *rstep* had so little influence over the docking in this case (cf. sizes of coefficient values in Figure 5), since these two parameters control the positioning of the ligand in the binding site. Even though these findings were valid for the docking results of the total set of diverse ligands, we could see noteworthy effects of changes of these parameter values for individual complexes.

The plot of the two extracted latent variables (scores) from the PLS model (adjusted to best describe their RMSD values) visualizes the relationships between the parameter settings of the 243 docking experiments (Figure 6a). The corresponding plot of the docking parameters' weight values (loadings) reveals the importance of each parameter in the PLS model and their relationships to the medians of the RMSD values (Figure 6b). Three distinct groups of docking experiments can be seen in the score plot (Figure 6a), which relate to the three investigated scoring functions for the exhaustive search (Chemgauss, Shapegauss, and Plp) since the selection dominated the docking results (Figure 6b). Within these groups, subgroups based on settings for the parameters *clash* and *neg_im* can be seen, especially within the *Plp* group. The two responses, *TP Median* and *BP Median*, are located to the far right in the loading plot (Figure 6b) and comparison with the location of the docking parameters reveals that the

parameter sets giving rise to the most successful dockings are in the *Chemgauss* group. This group includes two subgroups: one with parameter sets containing *clash* set at 1 to the upper right and the other with *clash* settings of 0.5 and 0.75 down to the left. On the basis of the two plots, we can conclude that the latter group contains the best parameter settings for our total set of ligands. Both the default parameter setting (Table 1) and the setting from the design that performed slightly better regarding the number of docked poses < 2.0 Å (with the parameter *score* set to Chemgauss, *neg_im* to large, *clash* to 0.5, *tstep* to 1.25, and *rstep* to 1) were among the best settings from a general perspective (Figure 6a).

Interestingly, the PLS model results indicate that there may be a difference between the optimal docking parameters for reproducing the binding mode in the crystal structure (sampling) and for successful ranking of poses (scoring). In the loading plot, the two responses (*TP Median* and *BP Median*) are spaced somewhat apart in the second component (Figure 6b), indicating that the influence of some of the parameters on the two responses differed. Separations in the second component were due to variations in the scoring functions (Shapegauss or Chemgauss), the negative image size, the translational step sizes, and various nonlinear terms. It is noteworthy that the variable *Chemgauss* was more important for a low *TP Median* than for a low *BP Median* and hence more important for accurate scoring. Conversely, the variable *Shapegauss* was more important for a low *BP Median*, that is, good sampling, than for a low *TP Median*. These findings indicate that both the Shapegauss and Chemgauss scoring functions can produce poses with low RMSDs to crystallographic ligand structures (low *BP Median*), but Chemgauss was better at ranking docking poses (*TP Median*). The fact that *Chemgauss* gave the best results was not surprising since the Chemgauss scoring function is a refined form of Shapegauss. However, our results indicate that Shapegauss was as good or maybe even better for generating poses similar to the binding modes in the crystal structure. Therefore, it may be of interest to further investigate the performance of Shapegauss in combination with postrefinement and rescoring.

Interpretation of the Effects of Docking Parameters in GOLD. The PLS model based on the docking experiments using the 126 designed parameter sets had two significant components describing 77% ($R^2Y = 0.77$) of the variation in the median RMSD values for the total set of 65 complexes, with most of the described variation in the first component. The Q^2 value derived from cross-validation was 0.71, the calculated estimation error (RMSEE) was 0.53 Å, and the permutation test gave an $R^2 < 0.05$ and a negative Q^2 for both the top pose and best pose medians, further verifying the significance of the model (plots of observed versus calculated RMSD values and plots of permutation tests can be seen in the Supporting Information). No docking runs were detected as outliers.

The most influential docking parameter was the Number of Operations in the GA, for which high settings (*num_ops* = 100 000) were beneficial according to the regression coefficient values (Figure 7). However, the coefficient value of the square term (cf. size and direction of the *num_ops** *num_ops* in Figure 7) suggests that the optimum for the parameter *num_ops* was within its tested range (1000 ≤ *num_ops* ≤ 100 000). The second most important parameter was the Niche Size, for which high settings seemed to be beneficial (*nichesize* = 3). Furthermore, the PLS coefficient values of square terms for Selection Pressure (*selpress*) and Crossover (*cross*) were of moderate sizes, indicating that optima may exist for these parameters within the tested spans (1.1 ≤ *selpress* ≤ 1.125; 95 ≤ *cross* ≤ 100). The remaining six parameters—Population Size (*popsize*), Number of Islands (*num_isl*), Mutate (*mutate*), Migrate (*migrate*), Hydrogen Bonding (*Hbond*), and Van der Waals (*vdWaal*)—did not affect the docking results for this set of complexes within the selected intervals (Table 2).

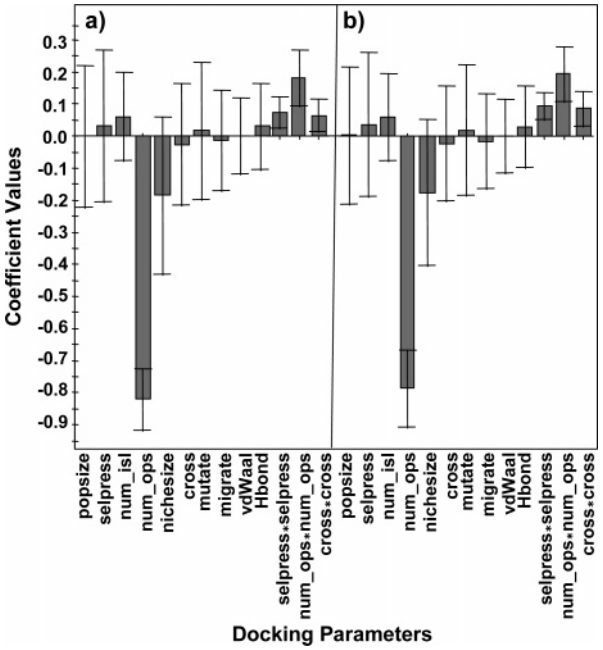

Figure 7. PLS coefficient plot showing the GOLD docking parameters' influence on the median RMSDs for the best pose (a) and the top pose (b) on the basis of the parameter span displayed in Table 2. The confidence intervals for each coefficient are also displayed.

The score and loading plots (Figure 8) for the GOLD PLS model were studied in an attempt to elucidate how the use of the 126 different parameter sets for dockings influenced the median RMSDs of the top and best poses to the corresponding crystallographic ligands. The docking runs are divided into two groups in the score plot (Figure 8a): parameter sets with 1000 GA operations are located in the upper right quadrant, while sets with 50 500 or 100 000 GA operations are located in the left part of the plot. Inspection of the loading plot reveals that this division reflects the docking results, since 1000 operations gave substantially poorer results than the other settings. The standard default setting is positioned among the best parameter sets investigated in this study. It can also be seen in the loading plot (Figure 8b) that the two responses, top pose and best pose medians, are highly correlated in both the first and second PLS components, where the second component mainly reflects the nonlinear relationship between the parameters and the docking results as described by the quadratic terms. In conclusion, the GOLD PLS model based on the D-optimal design showed that *num_ops* was the parameter with the strongest influence on the docking results.

In order to investigate the GOLD parameters further, an additional study was conducted in which a new D-optimal

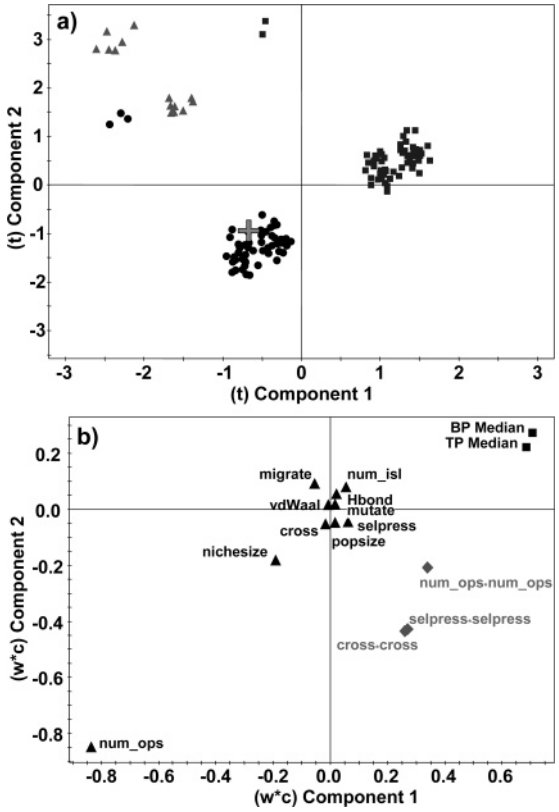

Figure 8. Score (a) and loading (b) plots of the PLS model based on the GOLD dockings. The 126 designed parameter sets can be seen in the score plot (a), where the markers illustrate the number of operations used (dots, 100 000; triangles, 50 500; squares, 1000). The standard default parameter set is marked with a cross. In the PLS loading plot (b), linear terms of the docking parameters are marked with triangles, square terms with diamonds, and responses with squares.

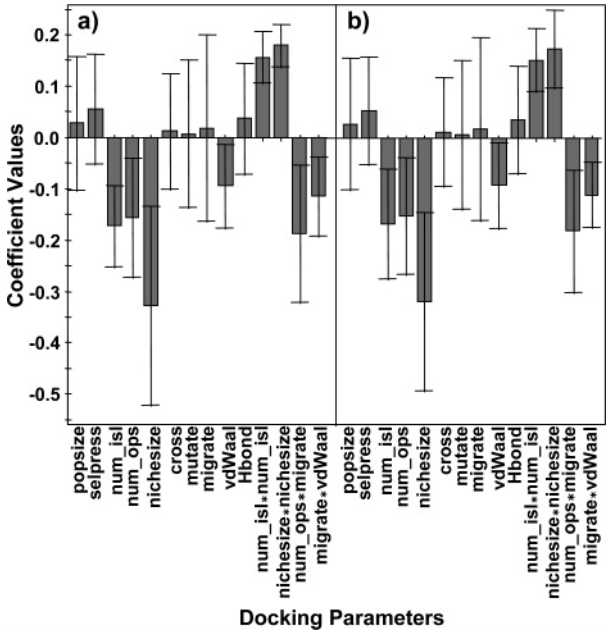

Figure 9. PLS coefficient plot showing the GOLD docking parameters' influence on the median RMSDs for the best pose (a) and the top pose (b) on the basis of the followup study (see text for details). The confidence intervals for each coefficient are also displayed.

design was made where the high and low values for Number of Operations were set to 50 500 and 100 000, respectively (center value of 75 250). This new design included all previous parameter sets containing *num_ops* equal to 50 500 and 100 000 (70 parameter sets), and an expansion of 56 parameters sets resulted in again 126 parameter sets for the followup investigation. In this merged set of parameter sets, based on the first selection, the square terms were to some extent confounded with each other (see the Supporting Information for the complete list and correlation matrix). The 65 ligands were docked with the additional parameter sets and postprocessed according to previously described procedures. PLS modeling based on the parameter sets from the focused D-optimal design and the docking results resulted in a model with R^2Y and Q^2 values of 0.33 and 0.22, respectively, with no deviating docking runs. The RMSEE value was 0.40 Å, and the permutation test gave an R^2 value of 0.05 and negative Q^2 values. Although significant, this model is of poor quality according to R^2Y and Q^2 (probably due to small differences in RMSD medians and the stochastic feature of the GA algorithm), so attempts to draw meaningful, detailed conclusions from it would be futile. However, the coefficient plot (Figure 9) indicates that the number of GA runs is less dominant in this model; in fact, the Niche Size became more influential when the parameter span for Number of Operations was changed. In addition, two more parameters seem to be important: the Number of Islands in the GA algorithm and the annealing parameter Van der Waals, high settings for these parameters being favorable. It can also be seen that the nonlinear terms had a greater impact on this model. However, no firm conclusions should be drawn regarding these terms due to the confounding of some of the square terms. For further investigation, more reliable models need to be obtained, which may require increased reproducibility of the dockings and use of more sensitive response indicators than median RMSD values.

According to the developers of GOLD, alterations of the GA and annealing parameters can influence docking significantly,[71] but within our spans of parameter values, this was only true for four of the 10 parameters (*num_ops*, *nichesize*, *num_isl*, and *vdWaal*). The fact that six out of 10 parameters had such a minor impact was somewhat surprising. In addition, a fairly small number of nonlinear terms was found to be significant, which surprised us since we anticipated a higher degree of nonlinearity.

Applicability of the Approach. The study conducted here on a diverse set of protein−ligand complexes to investigate the parameters of FRED and GOLD could easily be extended to also cover other software. While the presented methodology as such is applicable to other cases, the user needs to define the set of complexes and which parameters to use, including value ranges and intervals of these, that are of interest for that specific task. Benchmarking of software and development of high-performing default settings for new software could be applications of the presented approach. In such cases, the selected test set of complexes needs to cover all properties of targets and ligands that are potential research cases. Another application of the methodology could be to optimize parameter settings of docking software for a specific study, for example, a virtual screen. The results reported here indicated a high potential for such optimization by looking at the best performing individual parameter setting for each complex (Figure 3 and 4) but should not be seen as an optimization procedure. Instead, if that is the objective,

the parameter optimization could be performed toward multiple complexes, where known 3D structures of ligand−protein complexes of the target of interest or similar proteins constitute the test set. The subsequent PLS model based on the dockings could be used, in a similar fashion as presented here, to find the optimal parameter settings.

CONCLUSIONS

A general procedure based on multivariate techniques to explore the effects of docking parameters on docking performance has been developed. It was successfully applied to two fundamentally different programs: FRED and GOLD. A number of parameters controlling the docking procedure in FRED and GOLD were varied according to statistical experimental designs (full factorial design and D-optimal design, respectively), yielding parameter sets that covered the selected interval of parameter values. The designed parameter sets were applied in a docking study of two sets of protein−ligand complexes (one with 39 complexes and the other with 29) selected from the PDBbind database on the basis of ligand diversity using 1D and 2D descriptors, PCA, and space-filling design. The resulting docking poses' RMSD values to the corresponding crystallographic ligands were used as responses, both as individual values for each complex and as medians for the total set. In each case, two poses were extracted: the highest-ranked pose (top pose) and the pose most similar to the crystallographic ligand (best pose).

For this diverse set of protein−ligand complexes, 26% of the ligands docked using FRED yielded top poses with RMSD values < 2.0 Å to the crystallographic ligands when the default parameter settings were used and 44% when regarding the best pose. By selecting the best parameter settings for each protein−ligand complex from the designed set, an additional 23% of the docked ligands had top poses and best poses < 2.0 Å. Docking of the ligands using GOLD with default settings and 15 GA runs resulted in 38% of the top poses < 2.0 Å and 53% of the best poses. Applying individual parameter settings from the designed set to each complex resulted in an additional 26% with top poses < 2.0 Å and 12% of the best poses. These findings conclude that there is great potential to optimize the docking performance through modification of the docking parameters.

Significant regression models between the docking runs using the designed parameter sets and docking results (median RMSD values) were established for both programs using PLS regression. Similar trends were obtained for both of the independent sets of protein−ligand complexes, in terms of both model statistics and the influence of specific docking parameters. The fact that quantitative models could be obtained for these two independent sets indicates that modifications of the parameters could control the docking in certain directions within the investigated parameter space.

The multivariate analysis of the effect of docking parameters in FRED revealed that the selection of scoring function (*score*) for the exhaustive search had the greatest impact on docking performance, Chemgauss being the most reliable choice, especially in combination with a large negative image size. The parameter Clash Checking (*clash*) was the second most important, and a low setting (*clash* = 0.5) was beneficial. The parameters Translational Step Size (*tstep*) and Rotational Step Size (*rstep*) were of minor importance, but setting *tstep* equal to 1.25 seemed somewhat more beneficial than settings of 1 or 0.75. Furthermore, the model indicates that, even though Chemgauss seems to be the superior choice for yielding top poses similar to the crystallographic ligand, Shapegauss may be as good or even better for sampling of correct binding modes.

The multivariate analysis of the parameters in GOLD revealed that the parameter Number of Operations (*num_ops*) had the greatest influence on the docking performance, and a high setting was beneficial with a potential optimum between 50 500 and 100 000. The second most important parameter was the Niche Size, for which high settings (*nichesize* = 3) seemed to be beneficial. The subsequent focused design also indicated that high settings for Number of Islands (*num_isl* = 5) and Van der Waals (*vdWaal* = 10) may be advantageous.

Investigation of the score and loading plots showed that for both FRED and GOLD the default settings were among the best parameter sets for reproducing the ligand binding modes measured as RMSD values for this diverse set of complexes. However, the investigation in which individual parameter settings for each protein−ligand complex were examined showed that modifying the docking parameters could substantially improve the docking for certain complexes.

In summary, we have shown that statistical experimental design together with PLS regression is a powerful approach for exploring the importance of different docking parameters.

ACKNOWLEDGMENT

The authors thank the Swedish Research Council and AstraZeneca R&D Mölndal for financial support.

Supporting Information Available: List of 1D and 2D descriptors calculated for the ligands and used for diversity based selection, list of the different parameter sets used for docking in FRED, list of the different parameter sets used for docking in GOLD, list of the parameter sets from the expanded design used for docking in GOLD, correlated variables from the extended D-optimal design for GOLD parameters, results of the Kolmogorov−Smirnov test of using parameter values outside the design range, coefficient values for PLS models comparing sets 1 and 2, results of the permutation tests, and plots of observed versus calculated values for the PLS models. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Shoichet, B. K. Virtual Screening of Chemical Libraries. *Nature* **2004**, *432* (7019), 862−865.
(2) Jorgensen, W. L. The Many Roles of Computation in Drug Discovery. *Science* **2004**, *303* (5665), 1813−1818.
(3) Leach, A. R.; Shoichet, B. K.; Peishoff, C. E. Prediction of Protein−Ligand Interactions. Docking and Scoring: Successes and Gaps. *J. Med. Chem.* **2006**, *49* (20), 5851−5855.
(4) Kuntz, I. D.; Blaney, J. M.; Oatley, S. J.; Langridge, R.; Ferrin, T. E. A Geometric Approach to Macromolecule−Ligand Interactions. *J. Mol. Biol.* **1982**, *161*, 269−288.
(5) Halperin, I.; Ma, B.; Wolfson, H.; Nussinov, R. Principles of Docking: An Overview of Search Algorithms and a Guide to Scoring Functions. *Proteins: Struct., Funct., Bioinf.* **2002**, *47* (4), 409−443.
(6) Kitchen, D. B.; Decornez, H.; Furr, J. R.; Bajorath, J. Docking and Scoring in Virtual Screening for Drug Discovery: Methods and Applications. *Nat. Rev. Drug Discovery* **2004**, *3* (11), 935−949.
(7) Krovat, E. M.; Steindl, T.; Langer, T. Recent Advances in Docking and Scoring. *Curr. Comput.-Aided Drug Des.* **2005**, *1*, 93−102.

(8) Cole, J. C.; Murray, C. W.; Nissink, J. W. M.; Taylor, R. D.; Taylor, R. Comparing Protein−Ligand Docking Programs is Difficult. *Proteins: Struct., Funct., Bioinf.* **2005**, *60* (3), 325−332.

(9) Schulz-Gasch, T.; Stahl, M. Binding Site Characteristics in Structure-Based Virtual Screening: Evaluation of Current Docking Tools. *J. Mol. Model.* **2003**, *9* (1), 47−57.

(10) Kellenberger, E.; Rodrigo, J.; Muller, P.; Rognan, D. Comparative Evaluation of Eight Docking Tools for Docking and Virtual Screening Accuracy. *Proteins: Struct., Funct., Bioinf.* **2004**, *57* (2), 225−242.

(11) Perola, E.; Walters, W. P.; Charifson, P. S. A Detailed Comparison of Current Docking and Scoring Methods on Systems of Pharmaceutical Relevance. *Proteins: Struct., Funct., Bioinf.* **2004**, *56* (2), 235−249.

(12) Kontoyianni, M.; McClellan, L. M.; Sokol, G. S. Evaluation of Docking Performance: Comparative Data on Docking Algorithms. *J. Med. Chem.* **2004**, *47* (3), 558−565.

(13) Cummings, M. D.; DesJarlais, R. L.; Gibbs, A. C.; Mohan, V.; Jaeger, E. P. Comparison of Automated Docking Programs as Virtual Screening Tools. *J. Med. Chem.* **2005**, *48*, 962−976.

(14) Chen, H. M.; Lyne, P. D.; Giordanetto, F.; Lovell, T.; Li, J. On Evaluating Molecular-Docking Methods for Pose Prediction and Enrichment Factors. *J. Chem. Inf. Model.* **2006**, *46* (1), 401−415.

(15) Warren, G. L.; Andrews, C. W.; Capelli, A.-M.; Clarke, B.; LaLonde, J.; Lambert, M. H.; Lindvall, M.; Nevins, N.; Semus, S. F.; Senger, S.; Tedesco, G.; Wall, I. D.; Woolven, J. M.; Peishoff, C. E.; Head, M. S. A Crtitical Assessment of Docking Programs and Scoring Functions. *J. Med. Chem.* **2006**, *49*, 5912−5931.

(16) Stahl, M.; Rarey, M. Detailed Analysis of Scoring Functions for Virtual Screening. *J. Med. Chem.* **2001**, *44* (7), 1035−1042.

(17) Wang, R.; Lu, Y.; Wang, S. Comparative Evaluation of 11 Scoring Functions for Molecular Docking. *J. Med. Chem.* **2003**, *46* (12), 2287−2303.

(18) Xing, L.; Hodgkin, E.; Liu, Q.; Sedlock, D. Evaluation and Application of Multiple Scoring Functions for a Virtual Screening Experiment. *J. Comput.-Aided Mol. Des.* **2004**, *18* (5), 333−344.

(19) Kramer, B.; Rarey, M.; Lengauer, T. Evaluation of the FlexX Incremental Construction Algorithm for Protein−Ligand Docking. *Proteins: Struct., Funct., Bioinf.* **1999**, *37* (2), 228−241.

(20) Gohlke, H.; Hendlich, M.; Klebe, G. Knowledge-Based Scoring Function to Predict Protein−Ligand Interactions. *J. Mol. Biol.* **2000**, *295* (2), 337−356.

(21) Salo, J. P.; Yliniemelä, A.; Taskinen, J. Parameter Refinement for Molecular Docking. *J. Chem. Inf. Comput. Sci.* **1998**, *38* (5), 832−839.

(22) Mozziconacci, J. C.; Arnoult, E.; Bernard, P.; Do, Q. T.; Marot, C.; Morin-Allory, L. Optimization and Validation of a Docking-Scoring Protocol; Application to Virtual Screening for COX-2 Inhibitors. *J. Med. Chem.* **2005**, *48* (4), 1055−1068.

(23) Box, G. E. P.; Hunter, W. G.; Hunter, J. S. In *Statistics For Experimenters, An Introduction to Design, Data Analysis, and Model Building*; John Wiley & Sons, Inc.: New York, 1978.

(24) Linusson, A.; Gottfries, J.; Olsson, T.; Örnskov, E.; Folestad, S.; Norden, B.; Wold, S. Statistical Molecular Design, Parallel Synthesis, and Biological Evaluation of a Library of Thrombin Inhibitors. *J. Med. Chem.* **2001**, *44* (21), 3424−3439.

(25) Larsson, A.; Johansson, S. M. C.; Pinkner, J. S.; Hultgren, S. J.; Almqvist, F.; Kihlberg, J.; Linusson, A. Multivariate Design, Synthesis, and Biological Evaluation of Peptide Inhibitors of FimC/FimH Protein−Protein Interactions in Uropathogenic *Escherichia Coli*. *J. Med. Chem.* **2005**, *48* (4), 935−945.

(26) Tye, H. Application of Statistical 'Design of Experiments' Methods in Drug Discovery. *Drug Discovery Today* **2004**, *9* (11), 485−491.

(27) Gabrielsson, J.; Sjöström, M.; Lindberg, N.-O.; Pihl, A.-C.; Lundstedt, T. Robustness Testing of a Tablet Formulation Using Multivariate Design. *Drug Dev. Ind. Pharm.* **2006**, *32*, 297−307.

(28) Gabrielsson, J.; Lindberg, N.-O.; Lundstedt, T. Multivariate Methods in Pharmaceutical Applications. *J. Chemom.* **2002**, *16*, 141−160.

(29) Boström, J.; Greenwood, J. R.; Gottfries, J. Assessing the Performance of OMEGA with Respect to Retrieving Bioactive Conformations. *J. Mol. Graphics Modell.* **2003**, *21* (5), 449−462.

(30) Antes, I.; Merkwirth, C.; Lengauer, T. POEM: Parameter Optimization Using Ensemble Methods: Application to Target Specific Scoring Functions. *J. Chem. Inf. Model.* **2005**, *45* (5), 1291−1302.

(31) Wold, S.; Sjöström, M.; Eriksson, L. PLS-Regression: A Basic Tool of Chemometrics. *Chemom. Intell. Lab. Syst.* **2001**, *58* (2), 109−130.

(32) Wang, R.; Fang, X.; Lu, Y.; Wang, S. The PDBbind Database: Collection of Binding Affinities for Protein−Ligand Complexes with Known Three-Dimensional Structures. *J. Med. Chem.* **2004**, *47* (12), 2977−2980.

(33) Wang, R.; Fang, X.; Lu, Y.; Yang, C.-Y.; Wang, S. The PDBbind Database: Methodologies and Updates. *J. Med. Chem.* **2005**, *48* (12), 4111−4119.

(34) Wold, S.; Esbensen, K.; Geladi, P. Principal Component Analysis. *Chemom. Intell. Lab. Syst.* **1987**, *2* (1−3), 37−52.

(35) Eriksson, L.; Johansson, E.; Kettaneh-Wold, N.; Trygg, J.; Wikström, C.; Wold, S. In *Multi- and Megavariate Data Analysis - Basic Principles and Applications, Part 1*, 2nd ed.; Umetrics AB: Umeå, Sweden, 2006.

(36) Marengo, E.; Todeschini, R. A New Algorithm for Optimal, Distance-Based Experimental Design. *Chemom. Intell. Lab. Syst.* **1992**, *16*, 37−44.

(37) *FRED*, version 2.0.1; Openeye Scientific Software Inc.: Santa Fe, NM, 2004.

(38) Jones, G.; Willett, P.; Glen, R. C. Molecular Recognition of Receptor Sites Using a Genetic Algorithm with a Description of Desolvation. *J. Mol. Biol.* **1995**, *245* (1), 43−53.

(39) Jones, G.; Willett, P.; Glen, R. C.; Leach, A. R.; Taylor, R. Development and Validation of a Genetic Algorithm for Flexible Docking. *J. Mol. Biol.* **1997**, *267* (3), 727−748.

(40) Verdonk, M. L.; Cole, J. C.; Hartshorn, M. J.; Murray, C. W.; Taylor, R. D. Improved Protein−Ligand Docking Using GOLD. *Proteins: Struct., Funct., Bioinf.* **2003**, *52* (4), 609−623.

(41) *GOLD*, version 2.2; The Cambridge Crystallographic Datacenter: Cambridge, U. K., 2004.

(42) Massey, F. J. The Kolmogorov−Smirnov Test for Goodness of Fit. *J. Am. Stat. Assoc.* **1951**, *46* (253), 68−78.

(43) Kirkman, T. W. Statistics to Use. http://www.physics.csbsju.edu/stats/ (accessed March 19, 2007).

(44) St. John, R. C.; Draper, N. R. D-Optimality for Regression Designs: A Review. *Technometrics* **1975**, *17* (1), 15−23.

(45) *MODDE*, version 6.0; Umetrics AB: Umeå, Sweden, 2001.

(46) Dumouchel, W.; Jones, B. A Simple Bayesian Modification of D-Optimal Designs to Reduce Dependence on an Assumed Model. *Technometrics* **1994**, *36* (1), 37−47.

(47) de Aguiar, P. F.; Bourguignon, B.; Khots, M. S.; Massart, D. L.; Phan-Than-Luu, R. D-Optimal Designs. *Chemom. Intell. Lab. Syst.* **1995**, *30* (2), 199−210.

(48) *MOE−Molecular Operating Environment*, version 2004.03; Chemical Computing Group Inc.: Montreal, Canada, 2004.

(49) Wildman, S. A.; Crippen, G. M. Prediction of Physicochemical Parameters by Atomic Contributions. *J. Chem. Inf. Comput. Sci.* **1999**, *39* (5), 868−873.

(50) *SIMCA-P+*, version 11.0; Umetrics AB: Umeå, Sweden, 2005.

(51) *Matlab*, version 6.0; The MathWorks Inc.: Natick, MA, 2001.

(52) Efron, B.; Gong, G. A Leisurely Look at the Bootstrap, the Jackknife, and Cross-Validation. *Am. Stat.* **1983**, *37* (1), 36−48.

(53) Stone, M. Cross-Validatory Choice and Assessment of Statistical Predictions. *J. R. Stat. Soc. B.* **1974**, *36*, 111−147.

(54) Lindgren, F.; Hansen, B.; Karcher, W.; Sjostrom, M.; Eriksson, L. Model Validation by Permutation Tests: Applications to Variable Selection. *J. Chemom.* **1996**, *10* (5−6), 521−532.

(55) Eriksson, L.; Verboom, H. H.; Pejnenburg, W. J. G. M. Multivariate QSAR Modelling of the Rate of Reductive Dehalogenation of Haloalkanes. *J. Chemom.* **1996**, *10* (5−6), 483−492.

(56) Lindström, A.; Pettersson, F.; Almqvist, F.; Berglund, A.; Kihlberg, J.; Linusson, A. Hierarchical PLS Modeling for Predicting the Binding of a Comprehensive Set of Structurally Diverse Protein−Ligand Complexes. *J. Chem. Inf. Model.* **2006**, *46*, 1154−1167.

(57) Weininger, D. SMILES, a Chemical Language and Information-System. 1. Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, (1), 31−36.

(58) *BABEL*, version 1.6; OpenEye Scientific Software Inc.: Santa Fe, NM, 1997.

(59) *CORINA*; Molecular Networks GmbH: Erlangen, Germany. http://www2.chemie.uni-erlangen.de/software/corina/free_struct.html (accessed Sept 12, 2005).

(60) *SYBYL*, version 7.0; Tripos Inc.: St. Louis, MO, 2004.

(61) *MarvinView*, version 3.5.2; ChemAxon Ltd.: Budapest, Hungary, 2005.

(62) *OMEGA*, version 1.8b2; OpenEye Scientific Software Inc.: Santa Fe, NM, 2004.

(63) Verkhivker, G. M.; Bouzida, D.; Gehlhaar, D. K.; Rejto, P. A.; Arthurs, S.; Colson, A. B.; Freer, S. T.; Larson, V.; Luty, B. A.; Marrone, T.; Rose, P. W. Deciphering Common Failures in Molecular Docking of Ligand−Protein Complexes. *J. Comput.-Aided Mol. Des.* **2000**, *14* (8), 731−751.

(64) McGann, M. R.; Almond, H. R.; Nicholls, A.; Grant, J. A.; Brown, F. K. Gaussian Docking Functions. *Biopolymers* **2003**, *68* (1), 76−90.

(65) *OEChem*, 1.3.2; Openeye Scientific Software Inc.: Santa Fe, NM, 2004.
(66) Set 1. PDB codes: 1A99, 1AI7, 1AJ6, 1APT, 1AT6, 1B38, 1B7H, 1B8O, 1BHF, 1BKJ, 1BWB 1C5C, 1DY4, 1EX8, 1EZQ, 1F73, 1FCY, 1FH8, 1GJ9, 1H4N, 1HMS, 1HTG, 1IT6, 1M1B, 1MH4, 1PIP, 1QCA, 1QIW, 1RBO, 1SLG, 1VWN, 2CHT, 2PCP, 2XIS, 3PCJ, 4DFR, 4STD, 5ER2, 7CPA.
(67) Set 2. PDB codes: 1A9M, 1ADL, 1BM7, 1BTN, 1C88, 1CET, 1ECV, 1F4G, 1FD0, 1FKN, 1FQ5, 1GI6, 1GJ6, 1HWR, 1HYX, 1JQ9, 1KC7, 1KLL, 1LI3, 1MFE, 1PPK, 1QB1, 1QHC, 1QKB, 1USN, 2ER9, 2FGI, 2XIM, 4XIA.
(68) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and Computational Approaches to Estimate Solubility and Permeability in Drug Discovery and Development Settings. *Adv. Drug Delivery Rev.* **1997** *23*, (1−3), 3−25.
(69) Thomas, M. P.; McInnes, C.; Fischer, P. M. Protein Structures in Virtual Screening: A Case Study with CDK2. *J. Med. Chem.* **2006**, *49* (1), 92−104.
(70) The best parameter sets were selected from the docking runs using the designed 126-parameter sets plus 56 additional parameter sets resulting from a followup design based on the PLS model (see details in the text).
(71) *GOLD user manual*, version 2.2; The Cambridge Crystallographic Datacenter: Cambridge, U. K., 2004.

CI6005596