JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

Automated Ligand- and Structure-Based Protocol for *in Silico* Prediction of Human Serum Albumin Binding

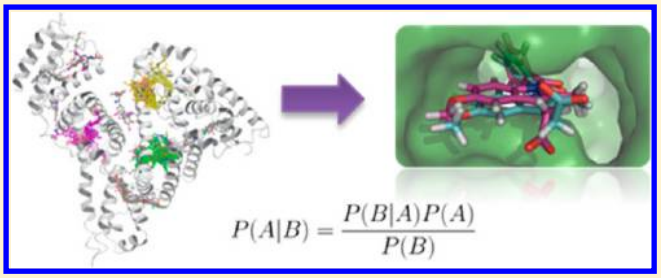
Michelle Lynn Hall,[†] William L. Jorgensen,[‡] and Lewis Whitehead*[,†]

[†]Novartis Institutes for Biomedical Research, 100 Technology Square, Cambridge, Massachusetts 02143, United States

[‡]Department of Chemistry, Yale University, New Haven, Connecticut 06520, United States

Supporting Information

ABSTRACT: Plasma protein binding has a profound impact on the pharmacokinetic and pharmacodynamic properties of many drug candidates and is thus an integral component of drug discovery. Nevertheless, extant methods to examine small-molecule interactions with plasma protein have various limitations, thus creating a need for alternative methods. Herein we present a comprehensive and cross-validated *in silico* workflow for the prediction of small-molecule binding to Human Serum Albumin (HSA), the most ubiquitous plasma protein. This protocol reliably predicts small-molecule interactions with HSA, including a binding affinity calculation using multiple linear regression methods, binding site prediction using a naive-Bayes classifier, and a three-dimensional binding pose using induced fit docking. Furthermore, this workflow is implemented in a portable and automated format that can be downloaded and used by other end users, either as is or with customization.



1. INTRODUCTION

Human Serum Albumin (HSA) is a ubiquitous plasma protein found in millimolar concentrations in the blood. Its native function is the nonspecific transport of nonesterified fatty acids and other hydrophobic endogenous ligands that would not otherwise be soluble in the hydrophilic bloodstream.[1,2] In addition, it often binds drugs with association constants that range from 10^3−10^{10} M, serving as both a friend and foe to drug discovery.[3] Specifically, HSA can bind highly hydrophobic drugs, thus diminishing the high clearance rates or poor distribution that might otherwise render these ineffective.[2] Low binding of these drugs may result in high toxicity or metabolism.[3] Alternatively, very high affinity binding of drug to HSA might prevent the drug from reaching the target at all, resulting in insufficient tissue distribution and efficacy.[1,3] This derives from the fact that the unbound fraction of drug (i.e., that which is not bound to plasma protein) affects the volume of distribution and the drug concentration in the target tissue.[3] Therefore, an understanding of small-molecule HSA binding is an important aspect of structure-based drug design (SBDD).

There are three fundamental questions to ask when considering small molecule binding to HSA: (1) What is the small molecule's affinity for HSA? (2) Which site(s) on HSA does the small molecule bind? (3) What pose does the small molecule adopt in the binding site(s)? An understanding of these facets has the potential to facilitate SBDD such that small molecule HSA interactions can be evaluated for maximum drug efficacy[4] so long as the other pharmacokinetic and pharmacodynamic parameters that govern drug efficacy are also carefully considered.[5]

For example, when efforts to develop a Bcl-2 inhibitor at Abbott Laboratories and Idun Pharmaceuticals were stymied by high affinity binding of their inhibitor to HSA, an understanding of the small molecule HSA interaction was used to effectively remove the interactions that made this binding so strong.[6,7] Regions of the Bcl-2 inhibitor that were solvent-exposed in the inhibitor-bound complex with Bcl-2, but buried in hydrophobic pockets in the inhibitor-bound complex with HSA, were identified by NMR. These regions were functionalized with polar groups, thereby diminishing binding to HSA while affecting minimal degradation in Bcl-2 binding. Ultimately, the inhibitor activity against Bcl-2 was retained and the high affinity binding of the inhibitor to HSA was attenuated, thus salvaging the inhibitor's efficacy. Similar success modulating drug binding to HSA was also achieved at Abbott Laboratories with a COX-2 inhibitor.[8]

In spite of this success, understanding HSA interactions with small molecules remains an elusive and challenging goal owing to the large and diverse set of distinct binding sites on HSA, its high degree of flexibility and binding cooperativity. HSA has at least eight distinct binding sites as shown in Figure 1.[3] Two primary drug binding sites have been described.[1] These are referred to as site 1 (domain IIA, also known as the warfarin site) and site 2 (domain IIIA, also known as the indole-diazepine site). It has been noted that site 1 preferentially binds large, heterocyclic, and negatively charged compounds, whereas site 2 binds smaller carboxylic acids.[4] In addition, a minor drug binding site (site 3)

Received: December 18, 2012
Published: March 10, 2013

dx.doi.org/10.1021/ci3006098 | *J. Chem. Inf. Model.* 2013, 53, 907−922

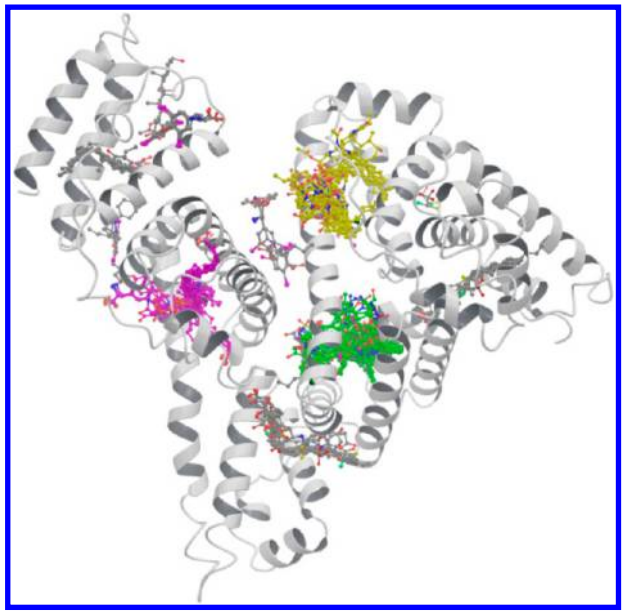


Figure 1. Superposition of all publically available crystal structures solved of HSA to-date with bound ligands. Only one representative protein structure is shown, while all ligands are shown. Ligands are colored by atom type. The carbon atoms of the ligands in site 1, 2, and 3 have been colored green, purple, and yellow, respectively.

has also been described in subdomain IB, where hemin, bilirubin, and fusidic acid are known to bind.[9] There are also up to three other minor potential drug binding sites at the interface between subdomains IIA and IIB, within subdomain IIIB, and at the cleft between domains I and II.[9]

The observed nature of these binding sites can vary greatly depending upon the crystallization conditions and induced-fit interactions with ligands.[1] Furthermore, binding affinity with HSA can vary dramatically in the presence of fatty acids. For example, warfarin binding to HSA is stronger in the presence of low concentrations of fatty acids, presumably owing to fatty acid binding induced conformational changes known to occur in HSA.[1]

We first describe some of the extant *in silico* methods available to study this system, highlighting some of the difficulties presented by the *in silico* study of HSA binding. We then introduce a novel protocol that enables the prediction of the binding affinity, site and pose.

2. *IN SILICO* METHODS TO STUDY SMALL-MOLECULE INTERACTIONS WITH HSA

Three fundamental facets of HSA binding are affinity, site, and binding pose. I.e., what is the affinity of the drug for HSA, what site(s) on HSA does the drug bind, and what pose and protein interaction does the drug adopt in that site? Different computational approaches can be used to tackle each of these aspects. Specifically, as described in the Sections 2.1−2.3 below, affinity and site prediction have shown to be amenable to treatment with ligand-based methods, while pose prediction necessitates treatment with structure-based methods. These three aspects (affinity, site, and pose) are all quite different, and thus very different computational methods must be used to tackle each. Nevertheless, a complete picture of HSA binding can only be realized upon understanding all three aspects.

Many groups have focused on predicting binding affinity to HSA with good success, as summarized in Section 2.1. By contrast, very few studies have investigated site and pose prediction in spite of their importance to understanding HSA-drug binding. All three aspects − affinity, site, and pose − are the focus of our study and are detailed in Sections 2.1−2.3.

2.1. Binding Affinity Prediction. Predictive models of relative HSA affinity have been studied extensively.[1,9−23] Much of this work has been reviewed by Hall,[3] and thus we only briefly discuss it here. A representative example is the work of Hanai, who has developed QSAR models to predict binding affinity with heterogeneous, although small data sets.[17,18] Others have developed models to predict free fraction (i.e., the percentage of drug that is not bound to plasma protein).[3]

An ideal method would facilitate binding affinity predictions across families of compounds and for large data sets that correlates well with *in vivo* binding affinity. This remains an elusive goal for molecular modeling in general, with HSA posing an even larger challenge owing to its plethora of binding sites and their high degree of flexibility and cooperatively. As noted by Murray and Hartshorn, at present, the best solution appears to be to construct local predictive models of binding affinity for chemical classes with known plasma protein binding liabilities.[4]

Numerous groups have developed statistics-based methods capable of predicting retention times of small molecules on HPLC with immobilized HSA.[1,10−16] For a given concentration of drug-binding sites of HSA on the column, and assuming only one binding site per HSA molecule, the retention time is related to the binding constant by a constant factor.[1] However, there is not just one but up to six binding sites on HSA. Additionally, HSA immobilized on an HPLC column is not truly representative of the highly dynamic nature of HSA *in vivo*.[3] Nevertheless, good correlation has been observed with binding affinities and HPLC retention times in spite of the assumptions made.[23] Furthermore, using this method has the advantage that it is medium-throughput and therefore provides a relative abundance of data. Accordingly, it is possible to use these data to build QSAR models that circumvent difficulties associated with using noisy data curated from multiple sources and methods. Thus, we describe a QSAR model for predicting HSA apparent binding affinity using HPLC retention times in Section 3.1.

2.2. Binding Site Prediction. In contrast to the numerous studies on small-molecule HSA binding affinity described above, we also focus on developing a protocol to accurately predict binding sites for small molecules with HSA. There are substantially fewer studies in this area in spite of its high utility to SBDD, presumably owing to the difficulty in describing such a highly flexible and promiscuous system. Zsila has shown the utility of *in silico* methods for the prediction of which site on HSA the small-molecule will occupy using statistical methods with good success and has implemented this predictor as a freely available Web tool that takes the user's chemical structure as input.[9] However, this Web tool remains inaccessible to researchers in the pharmaceutical industry owing to restrictions limiting the distribution of proprietary intellectual property on the Web and elsewhere. Therefore, we have devised an *in silico* statistics-based binding site predictor and implemented this as a workflow that can be downloaded onto the user's computer, thus making it accessible to users in the pharmaceutical industry and circumventing intellectual property concerns.

2.3. Binding Pose Prediction. Compared to studies that predict binding affinity, there are no validated studies that report the use of *in silico* tools for the prediction of the binding pose that do not depend upon some prior knowledge of the specific

ligand−receptor interactions. Mao used NOEs to dock their compounds into a crystal structure of HSA with good results.[8] However, this necessitated prior knowledge of the NOEs and thus the use of NMR. The protocol by Zsila provides the user with a guess of the binding pose within the predicted binding site using standard docking procedures (i.e., rigid receptor and flexible ligand).[9] However, the method was not validated against co-crystal structures of HSA and again suffers from inaccessibility to pharmaceutical users since it is located on an external Web server.

Furthermore, HSA is a highly flexible protein, and our preliminary research led us to conclude that only induced fit docking or other more sophisticated and nuanced docking methods would be capable of capturing the dynamic nature of small molecule interactions with HSA. Receptor conformational dynamics can be captured through the use of Molecular Dynamics simulations run on sufficiently long time scales, for example.[24] However, induced fit docking provides a more computationally efficient method to sample conformational space, especially for highly flexible binding sites such as those observed in HSA.[25] Therefore, we report a validated induced fit docking protocol that is capable of accurately reproducing the crystallographic pose of known HSA binders without prior knowledge of specific ligand−receptor interactions. This workflow is linked to the site predictor described in Section 2.1 (and later in Section 3.2), thus giving the user the ability to answer both the question of which site their small-molecule occupies and what pose the small molecule adopts in that site. Finally, this workflow is implemented such that it can be downloaded and used locally, again making it accessible and customizable to users with intellectual property to protect.

3. RESULTS AND DISCUSSION

3.1. Binding Affinity Prediction. The use of high performance liquid chromatography (HPLC) screens for HSA binding affinity is a common practice in drug discovery.[23,26] This method is medium-throughput, making it a popular choice for lead optimization projects. Furthermore, the ability to accumulate data on many molecules using the same experimental technique facilitates the development of quantitative-structure activity relationships (QSAR) for apparent affinity to HSA.[3]

Colmenarejo reports HPLC retention times and apparent binding affinities for 95 diverse drug molecules.[26] Numerous authors have used this data to produce QSAR models with good success as reviewed elsewhere.[1] One such method employs QikProp[27] descriptors to describe the apparent affinity constant via multiple linear regression.[28] One advantage of this particular model is that it is commercially available and thus readily employed by the end user for lead optimization purposes.

Similar to the work reported by Colmenarejo, Valko reports retention times, apparent affinity constants, and fraction of drug bound to HSA for 121 diverse drug molecules.[23] (The diversity of this set may be measured, for example, by radial fingerprints[29] and Tanimoto distances.[30] The mean Tanimoto distance for any two molecules of the data set is 0.06 ± 0.08.) Unfortunately, because Colmenarejo and Valko use slightly different methods to obtain their measurements, the agreement between values obtained using the two methods is rather poor. Specifically, the correlation coefficient (R^2) is only 0.47 for the 30 molecules that occur in both sets. (See Supporting Information Figure S1.) Thus, the numerous QSAR models built using the measurements reported by Colmenarejo are not applicable to measurements made using the same technique as Valko. This is problematic in drug discovery campaigns where the latter technique is employed, as reported by Reilly, for example.[31]

We have used the measurements reported by Valko to build a multiple linear regression model with QikProp descriptors capable of describing apparent binding affinities. Unfortunately, Valko does not report error bars for retention times. However, Colmenarejo uses a similar technique and does report error bars of only about 1% in retention times.[26] Because the logarithm of the retention time is used to give apparent binding affinity, we assume that the error in this value is quite small, thus making it ideal for QSAR model development.

QikProp descriptors were chosen because they are readily available, intuitive, and physiologically relevant.[27] Additionally, Jorgensen has shown that a multiple linear regression built using these descriptors gives good predictability for the Colmenarejo set,[28] suggesting their utility in describing the Valko set as well. Multiple linear regression was chosen in particular over other methods to facilitate consistency with the previously developed model.[28] Using measurements obtained in-house as previously described by Reilly,[31] we were also able to validate our model using a completely independent test set of 28 molecules.

To build training and test sets, the molecules of the Valko data set were first prepared by assigning 3D structures and neutralizing all charges (i.e., protonating all acids and deprotonating all bases). Ceftazidime was discarded, as it is a zwitterion and is thus not well-described by QikProp descriptors. The remaining 120 molecules were split randomly five times into 3:1 training and test partitions. Each of the five training sets was used to build multiple linear regression models (referred to as "submodels") with a subset of 1, 2, ..., 7 QikProp descriptors from the available pool of 36 total QikProp descriptors. These five submodels were then used to predict the five corresponding test sets. The average correlation coefficients, R^2 and Q^2, for the training and test sets are shown in Figure 2 along with their

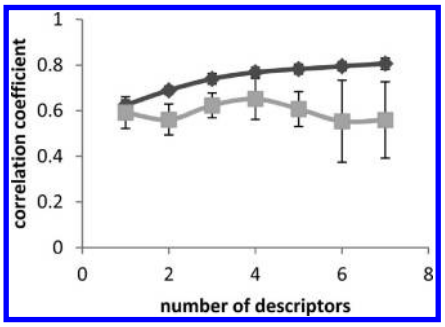


Figure 2. Performance of predictive models as a function of the number of descriptors employed for the training (dark gray) and test (light gray) sets. Error bars show standard deviation. Data tabulated in Table 1.

standard deviations. The first significant drop in predictability for the test set occurs in the transition from 4 to 5 descriptors, where overfitting of the training set begins to degrade the overall performance. Thus, the final model is built using only four descriptors.

The final predictive model, built using all 120 data points published by Valko, is given by

$$\log K = -1.01 + 0.43(\#acid) - 0.26(CIQPlog\ S) - 0.0011(FOSA) + 0.28(QPlogPo/w) \quad (1)$$

where *#acid* is the number of carboxylic acids, *CIQPlogS* is the predicted conformation-independent aqueous solubility, *FOSA*

dx.doi.org/10.1021/ci3006098 | J. Chem. Inf. Model. 2013, 53, 907−922

Table 1. Performance of Predictive Submodels As a Function of the Number of Descriptors Employed for the Training and Test Sets[a]

number of descriptors	average R^2 across training sets	average Q^2 across test sets
1	0.63 ± 0.02	0.59 ± 0.07
2	0.69 ± 0.02	0.56 ± 0.07
3	0.74 ± 0.02	0.62 ± 0.05
4	0.77 ± 0.02	0.65 ± 0.09
5	0.78 ± 0.02	0.61 ± 0.08
6	0.79 ± 0.02	0.55 ± 0.18
7	0.81 ± 0.02	0.56 ± 0.17

[a]Standard deviations are also given. Data shown graphically in Figure 2.

is the hydrophobic component of the solvent accessible surface area, and *QPlogPo/w* is the predicted octanol/water partition coefficient. We can ask the question of how similar this final model (eq 1) is to the five independently trained submodels described above. Specifically, how often did the descriptors that appear in eq 1 also appear in the five independently trained submodels? We find that all four of the descriptors in eq 1 appeared in all five of the submodels, with the exception of *FOSA*, which appeared in four out of five submodels. This suggests the stability of the resultant model described in eq 1. (Here we consider only the four-descriptor models for simplicity and transferability.) The model's predictability is shown in Figure 3a.

The two descriptors *CIQPlogS* and *QPlogo/w* make the largest contribution to the predicted log*K*. In fact, building a one-descriptor regression model using *QPlogo/w* or *CIQPlogS* alone gives R^2 values of 0.62 and 0.46, respectively, compared to $R^2 \leq$ 0.02 when using *#acid* or *FOSA* alone. Additionally, the descriptors and their corresponding scalars describe well-known phenomena in HSA binding. Specifically, HSA is known to bind hydrophobic ligands with carboxylic acids. This is well-captured by the model, where the dominant descriptors for hydrophobicity, *QPlogo/w* and *CIQPlogS*, are scaled such that hydrophobic and poorly soluble compounds will have higher predicted log*K*, i.e., higher predicted apparent affinity for HSA. These two descriptors (*QPlogo/w* and *CIQPlogS*) are loosely correlated as might be expected, with a small but noteworthy correlation coefficient of only $R^2 = 0.4$.

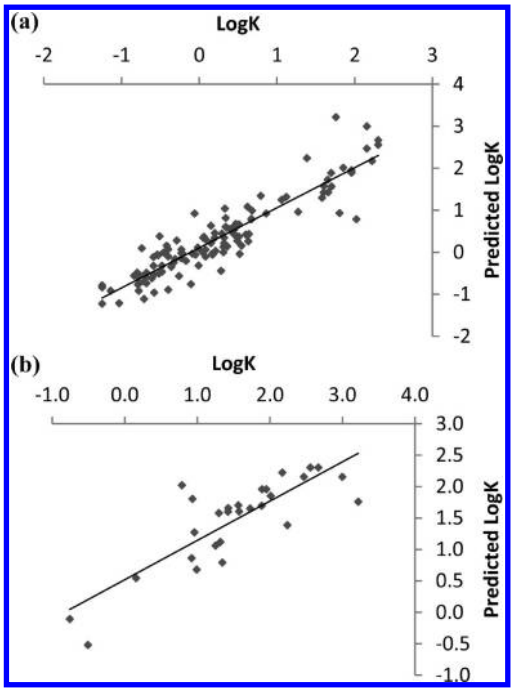


Figure 3. Predicted Log*K* versus measured Log*K*. (**a**) Training set (i.e., all Valko data). The correlation coefficient (R^2) is 0.75, while the line of best fit is given by Predicted Log*K* = 0.75*Log*K* + 0.0472. (**b**) External test set (i.e., in-house data). The correlation coefficient (R^2) is 0.68, and the line of best fit is given by Predicted Log*K* = 0.62*Log*K* − 0.52. The mean unsigned error is 0.4 log units.

The number of acidic functional groups in a molecule (*#acids*) is weighted by 0.43 in eq 1. It is important to note the range of *#acid* for the training group employed in this study is 0−1. Stated otherwise, our training set consists entirely of nonacidic molecules (*#acid* = 0) or monoacids (*#acid* = 1). Thus, one should exercise caution when using eq 1 to study molecules that may lie outside the domain of applicability of the developed model, including poly acids (*#acids* > 1) in particular.

The performance of this model is comparable to that reported by others.[1] Valko reports a multiple-linear regression model to describe apparent binding affinity data using five Abraham descriptors with similar results.[23] Additionally, QikProp descriptors have also been used to describe the Colmenarejo set using a linear regression model with slightly better performance than that reported here.[28] This is unsurprising given that that this model employs seven descriptors, while the one described herein only employs four. Of the seven descriptors employed in the QikProp model trained on Colmenarejo data (*volume*, *donorHB*, *acceptorHB*, *ACxDN̂0.5/SA*, *#acid*, *#amide*, and *#rotor*), only *#acid* also occurs in our model as well. As described above, the presence of this descriptor in both models highlights the fact that acids are pathological HSA binders. Using the same descriptors reported in this study (*#acid*, *CIQPlogS*, *FOSA*, and *QPlogPo/w*) to build a four-descriptor QSAR model based upon the Colmenarejo data set gives similar performance (R^2 = 0.63 across the whole training set), suggesting the robustness of this descriptor set to describe HSA binding affinity.

In addition to the five random-split test sets from Valko's data described above, we also tested the model on an external set of in-house data that has been described elsewhere.[31] These data were not previously included in the training of the model whatsoever and thus constitute a genuine, external validation test set. These data were measured using a similar protocol to that of Valko and provide a convenient test of the model's applicability to drug discovery and lead optimization. The 28 external test molecules were chosen by their similarity to the Valko set to ensure applicability of the model. Specifically, the similarity score of every molecule in the Valko set against molecules in our in-house collection was quantified using extended connectivity fingerprints[32] and Tanimoto distances.[30] Only those molecules with similarities of ≥0.75 to at least one molecule of the Valko set were chosen as members of the external validation set. This was done to ensure applicability, as the diversity of chemical space described by the in-house data is vast and not described by simple models.[31] While some outliers exist, the model shows good applicability to the external validation set, as shown in Figure 3b.

As described in further detail in Section 3.4, this model has been implemented into a portable workflow. Upon user input of one or many molecules in SMILES format, the protocol prepares the molecules and predicts their binding affinities. Both the predicted log*K* and the qualitative binding affinity (low, medium,

or high) are output. The motivation in developing qualitative binding affinity categories was to provide a more interpretable metric for the end-user than log*K*, since these values are not necessarily intuitive. The categories were chosen based upon the reported distribution of log*K* values reported by Valko, where bins of $(-\infty, -0.5)$, $[-0.5, 0.5)$, and $[0.5, \infty)$ were chosen for low, medium, and high affinity, respectively. These cutoffs were chosen such that the majority of compounds will fall into the "medium" affinity category, allowing one to focus only on those molecules which may have problematically low or high affinity HSA binding.

3.2. Binding Site Prediction. In contrast to the classical QSAR model to describe binding affinity (section 3.1, above), to predict the binding site within HSA, we have built a naïve Bayesian classification model. The classifier utilizes Bayes's theorem[33] to predict a best category (i.e., binding site) based upon independent variables. Bayes's theorem utility to chemistry has been described extensively.[34,35]

This classifier utilizes a publically available data set of 125 ligands with known binding site occupancies with respect to the two primary drug binding sites, site 1 and site 2.[21] The data set and assigned binding categories are provided in the Supporting Information, while a few representative examples are shown in Figure 4. Of the 125 ligands, there are 26 and 34 molecules which bind to sites 1 and 2, respectively.

Molecules that bind to minor drug binding site 3 or other, even less common drug-binding sites, are also of interest. However, there is a lack of publically available data on small-molecule occupancy of these minor sites, thus limiting the ability to build a reliable predictive protocol.

To describe the molecules, seven standard descriptors were computed including molecular weight (MW); AlogP; number of charged acceptor groups (CAG), e.g., carboxylic acids, a group known to contribute to HSA binding; hydrogen bond donors (HBD); hydrogen bond acceptors (HBA); molar refractivity (MR), a measure of the molecule's polarizability; and polar surface area (PSA). As discussed by Duan, the correct choice of fingerprint method is not always obvious.[29] Because radial fingerprints (also known as extended connectivity fingerprints or ECFPs)[32] are designed specifically for use in quantitative structure−activity relationships, and these were investigated in addition to dendritic fingerprints.[29] We found that varying the radial fingerprint radius (from 4 gradually up to 12) had little effect on the performance of the model, and therefore we report only the results for radius 4 in Table 2 for simplicity. We also computed the Bayesian model with no fingerprints, i.e., using only molecular descriptors, for reference. The total number of bits produced from the 32-bit fingerprints is in the range of $[1,2^{32}]$. Before building the model, these bits are filtered using a Kullback−Leibler distance (also known as relative entropy) cutoff of 0.1.[36] Specifically, binary features with a significance ≤ 0.1 (i.e., those with relatively high correlation) are excluded from the model. We found that using a larger cutoff (i.e., filtering off more bits) only degraded the performance of the model on the various test sets.

The data set was split five times randomly into 3:1 training and test sets, and the classifiers were validated based upon their ability to obtain correct predictions, where 100% would correspond to a perfect model and 50% would correspond to a model with results no better than random. This was quantified for both the five randomly chosen test sets and for the entire data set of 125 data points with results shown in Table 2.

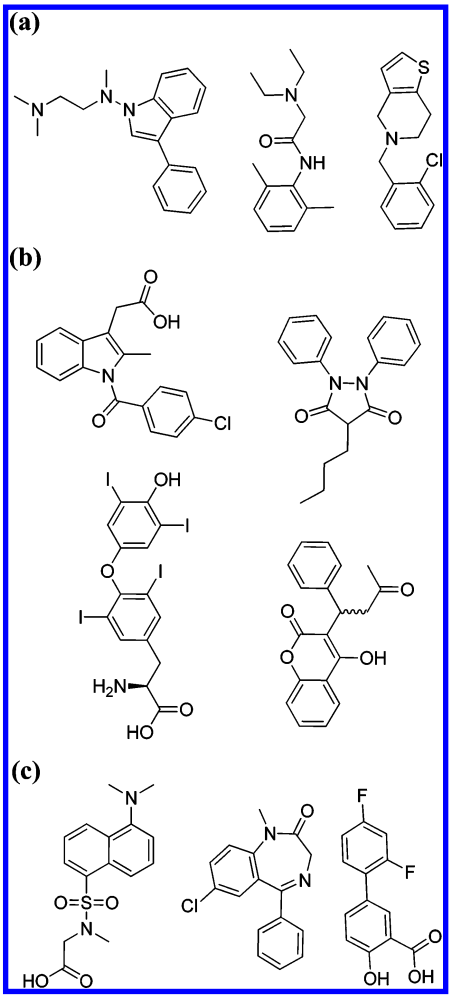


Figure 4. Representative molecules of the data set. (**a**) Molecules that bind to neither site 1 nor site 2 on HSA. From left to right: binedaline, lidocaine, and ticlopidine. (**b**) Molecules that bind to site 1. Indomethacin (top left), phenylbutazone (top right), thyroxine (bottom left), and warfarin (bottom right). (**c**) Molecules that bind to site 2. From left to right: dansylsarcosine, diazepam, and diflunisal.

Table 2. Bayesian Classifier Performance Quantified As the Average Percent of Correct Predictions Across the Five Independent Test Sets Including the Standard Deviation[a]

binding category	no fingerprints	dendritic fingerprints	radial fingerprints
binds to site 1	74.0 ± 15.3% (83.1%)	79.3 ± 3.7% (100.0%)	81.3 ± 5.6% (100.0%)
binds to site 2	66.0 ± 9.2% (66.9%)	83.3 ± 5.8% (99.2%)	79.3 ± 8.2% (100.0%)
average	70.0 ± 5.7% (75.5%)	81.3 ± 2.8% (99.6%)	80.3 ± 1.8% (100.0%)

[a]The results of training the model on all the data (i.e., no test set) are also shown in parentheses. All models used the descriptors AlogP, CAG, HBA, HBD, MR, MW, PSA, and the fingerprints specified.

Using no fingerprints, i.e., using only the descriptors, gives reasonable results, where the accuracy of predictions for the test sets is 70.0% on average. The addition of dendritic or radial fingerprints results in an improvement of approximately 10% for the test sets. Interestingly, the performance of the dendritic and radial fingerprints is negligibly different, and both methods give predictions with accuracies of approximately 80%. The most marked improvement in performance is seen for the training sets,

dx.doi.org/10.1021/ci3006098 | J. Chem. Inf. Model. 2013, 53, 907−922

where the accuracies are consistently perfect or near-perfect with the addition of fingerprints. This has also been observed using Support Vector Machines as reported by Zsila.[9]

We asked whether the use of a simple maximum-similarity lookup table could outperform either machine-learning method − Support Vector Machine or Bayesian classification. We constructed a similarity matrix based upon dendritic fingerprints for all compounds of the training set. We then assigned the predicted binding site for each molecule of the training set to be the experimental binding site reported for the molecule with the highest similarity. Interestingly, this method only predicted the correct binding site approximately 50% of the time and is thus much less accurate than either machine learning method which have accuracies of approximately 80%. In the maximum similarity lookup, each molecule's binding site is predicted solely based upon the data provided by one other molecule, i.e., the one to which it is most similar. In contrast, the Bayesian classifier exploits the data of all molecules in the training set in the form of a probability distribution function built for each independent variable, fingerprint bit, or otherwise. Therefore, it is not surprising that the Bayesian classifier outperforms the maximum similarity lookup.

Our final Bayesian model employs dendritic fingerprints since these give slightly better performance on the test sets on average versus radial fingerprints. In the automated workflow (described in Section 3.3), only those molecules which are predicted to bind to site 1 or 2 are passed onto the subsequent pose prediction stage.

In contrast to our study, Zsila used Support Vector Machine learning on a slightly different data set with different descriptors.[9] They report accuracies of 78% on their test set, comparable to what we report here. This shows the ability to predict binding sites is robust across different statistical methods and descriptor sets; i.e., this is a highly predictable phenomenon in spite of the high degree of flexibility in sites 1 and 2. This is consistent with the empirical and qualitative observation of differences in binding site preferences for site 1 and site 2. Specifically, site 1 preferentially binds large, heterocyclic, and negatively charged compounds where the negative charge is centrally located. Conversely, site 2 binds smaller, more linear compounds with negative charge located peripherally.[4,9] Both our method and that of Zsila are then quantifying the chemical knowledge that was carefully learned in the past from detailed analysis of experimental data.

A key advantage of our method is that it can be used on a desktop computer. Another advantage is the customizable nature of the workflow. This is particularly useful for Bayes classifiers and is a justifiable reason for using this type of classifier instead of Support Vector Machines. In Bayes classifiers, the prior is given by the distribution of molecules across the various binding categories in the training set. Thus, if one is interested in studying a particular chemotype that is a site 1 binder, for example, one can bias the priors toward this and exploit expert knowledge simply by adding the additional training data for these site 1 binders.

3.3. Binding Pose Prediction. *Overview.* The ligand-based methods described in Sections 3.1 and 3.2 predict binding affinity and binding site through the use of QSAR and Bayesian classification models, respectively. Equipped with these, we sought to develop a reliable method to predict the pose of the ligand in the binding site using structure-based methods. To test our protocol, it was necessary to assemble a suitable test set; therefore, we have assembled a data set of crystal structures where the ligand is found crystallized in the same binding site that is also experimentally validated by an alternative technique. This was done to ensure that a protocol was not validated based upon noise, since observed crystallographic binding site occupancy is sensitive to conditions employed. The molecules of the data set occupy site 1 (indomethacin, phenylbutazone, thyroxine, and (*R*)- and (*S*)-warfarin) and site 2 (dansylsarcosine, diazepam, and diflunisal) shown in Figure 4b,c.

As discussed in Section 1, HSA is a highly flexible protein. Thus, our preliminary studies suggested that conventional docking techniques like those employed by Zsila[9] (i.e., rigid receptor and flexible ligand) would not adequately describe ligand-HSA interactions. Ensemble docking (i.e., multiple rigid receptors and flexible ligands) is a popular alternative to conventional docking and is suitable in cases where the receptor is known to take on a few discrete geometries. However the binding site of HSA takes on a multitude of fairly continuous geometries. As an illustrative example, consider just site 1. In site 1, there are six flexible residues (E153, K199, W214, R218, R222, and E292) whose movements are not necessarily coupled. If each of these were to take on only two independent rotomers, this would give $2^6 = 64$ distinct receptor structures, each requiring a separate run in ensemble docking. Alternatively, we have explored the use of induced fit docking (IFD) to study this problem since it has been shown to yield promising results.[25]

Induced fit docking combines conventional docking (i.e., rigid receptor and flexible ligand) and methods to capture receptor conformational changes iteratively to not only dock the ligands but also to obtain an image of the specific ligand−receptor interactions that contribute to binding. Within the Schrödinger suite, this takes the form of first docking into a rigid receptor with Glide[37] using a softened potential to generate an ensemble of poses, followed by refinement of the members of this ensemble with Prime.[38] The refinement stage involves sampling side chains in dihedral space and allowing for small protein backbone adjustments of the receptor. The refined receptor structures are then used again for rigid receptor docking, this time with a hard potential. The resultant structures are then scored using a composite energy function that approximates the receptor−ligand interaction energies (GlideScore) in addition to receptor strain energies and solvation energies using implicit solvent (Prime energy).

Critical to the success of any docking protocol is the choice of a proper receptor. There are 53 and 50 co-crystal structures of ligands bound to site 1 and site 2 of HSA, respectively. Each of these co-crystals can potentially serve as a receptor, giving up to 50−53 separate induced fit docking runs per site. However, since IFD is more expensive than conventional docking, we seek the smallest subset of receptors possible that will also span maximum conformational space of the binding site. Therefore, we first clustered all of the crystal structures for site 1 and site 2 independently. This was done by computing the matrix of RMSDs for all amino acids within 5 Å of the crystallographic ligand and iteratively ordering the crystal structures into clusters until only a few clusters were formed. From there, the representative member of each cluster was chosen to be the one with the largest ligand since this would facilitate properly executing the first step of the induced fit docking (i.e., docking the flexible ligand into the rigid receptor) better than a structure with a smaller ligand. The representative crystals chosen were those with PDB codes 2bxh, 2bxn, 2xsi, 2xw0, and 3cx9 for site 1 and 2bxh, 2xvq, and 2xvu for site 2. Notice that five crystals were chosen for site 1, while only three were chosen for site 2. The

larger size of site 1 made it less amenable to clustering than site 2, thus giving more representative crystal structures.

All IFD runs were performed with default settings as described in Section 5.3. It is important to note that all final reported docking was cross docking (i.e., the ligand being docked is not the same ligand that appears in the co-crystal structure). However, we have found that self-docking (i.e., docking the native ligand back into its native receptor) also works well with our protocol.

There are two important aspects of docking: pose generation and pose scoring. The first involves whether the docking algorithm can produce a pose that is close to the known pose at all. The second is whether or not that pose, once produced, will be scored highly, thus distinguishing it from all the other poses that may also be produced. In general, our protocol is capable of reproducing the crystallographic pose of the test set ligands and ranking this pose highly, thus successfully accomplishing both aspects necessary for a rigorous docking protocol. Furthermore, the side-chain interactions that are important for ligand-HSA binding are also well described by our method. Describing these side-chain interactions is especially important for ultimately enabling structure-based drug design around small-molecule HSA interactions. The results are summarized in Table 3, while a detailed discussion of the results for each ligand follows.

Finally, strict comparisons of the IFD solutions obtained herein to the crystal structures should be regarded with caution. Specifically, HSA is notoriously difficult to crystallize, especially in high resolution.[2] The crystal structures used as benchmarks here (see Table 3) have an average resolution of approximately 2.5 Å. Thus, unfavorable contacts and steric clashes are commonly observed. A solution that is often employed to overcome this difficulty is to perform a restrained molecular mechanics minimization of the crystal structure to eliminate these contacts and obtain a structure that is more easily compared to those produced using *in silico* methods. However, the resultant minimized structure is a function of the particular force field employed for the minimization. One may argue this biases the benchmark crystal structure to be more similar to the modeled structure, especially if this is the same force field employed for docking and refinement. For this reason, we present the benchmark crystal structures as is (i.e., without minimization) for comparison to the IFD solutions. Note that unfavorable and unreasonable interactions which may appear in crystal structures will not be reproduced by any well-parametrized force field, and thus one-to-one comparisons between the IFD solution and crystal structure may often be impossible. This is further complicated by the convention of using RMSD as a quantitative metric. Among other systematic difficulties, RMSD is a function of size of the molecule.[39] Thus, its ability to rigorously measure the "goodness" of a pose is questionable. Nevertheless, we have provided the RMSD to facilitate comparison with other studies. Owing to the difficulties described in both crystal structure resolution and RMSD, we argue that a more meaningful analysis may be realized by careful, qualitative examination of the IFD solutions in comparison to the crystal structures. In such analysis, it is important to pay attention to not only the ligand pose but also the ligand interactions with protein side chains that contribute to binding. Detailed analyses of our IFD solutions follow.

Table 3. Induced Fit Docking Results on the Molecules Shown in Figure 4b,c

site	ligand	reference crystal structure	IFD receptor crystal structure	RMSD (Å)[a]	IFD rank[b]
1	indomethacin	2bxk, 2bxm, 2bxq	2bxn	1.87 ± 0.18	1 (15)
1	indomethacin	2bxk, 2bxm, 2bxq	2bxn	1.31 ± 0.07	2 (15)
1	phenylbutazone	2bxp, 2bxq	3cx9	1.96 ± 0.30	1 (13)
1	phenylbutazone	2bxp, 2bxq	3cx9	2.44 ± 0.30	2 (13)
1	thyroxine	1hk1, 1hk2, 1hk3	2xw0	2.75 ± 0.16	3 (15)
1	(*R*)-warfarin	1h9z, 2bxd	3cx9	1.53 ± 0.60	1 (13)
1	(*S*)-warfarin	1ha2	3cx9	2.10	3 (14)
2	dansylsarcosine	2xvq	2bxh	1.76	1 (14)
2	diazepam	2bxf	2bxh	0.82	1 (7)
2	diflunisal	2bxe	2bxh	2.01	1 (11)
2	diflunisal	2bxe	2bxh	1.46	2 (11)

[a]Computed by comparing the coordinates of the reference ligand to that produced using IFD. The structures were first aligned using the Align Binding Sites module within Maestro. In some cases, there are multiple reference crystal structures with very similar ligand positions. In these cases, the reference crystal structures are considered together, and the average RMSD and standard deviation are reported. Standard deviations are also given. [b]This is the relative rank of the complex described with respect to all IFD complexes that use the same combination of receptor and ligand. The IFD complex with the best score is ranked 1, the second best is ranked 2, and so forth. The number of complexes output is listed in parentheses.

Site 2. In addition to drug binding, site 2 binds two, medium-chain fatty acids (e.g., myristate and palmitate), and thus competitive binding of drugs and fatty acids is observed. A comparison of the crystallographic orientation of bound fatty acids and drug molecules reveals that many pharmacophore features are shared by these seemingly disparate classes of molecules. Specifically, the carboxylic acids of the nonesterified fatty acids occupy the same regions of the binding pocket as other charged or polar groups on various drug molecules. Conversely, the nonpolar tails of the fatty acids occupy the same regions of the binding pocket as the more hydrophobic regions of the drug molecules. In contrast to site 1, site 2 preferentially binds relatively small, linear molecules with terminal negative charges. However, like site 1, site 2 is comprised of an apolar cavity flanked by polar residues on helices 1, 2, and 6.[2] These polar residues are solvent-exposed, and some form stabilizing salt bridges with residues in subdomain IIB of HSA, i.e., E450 and R485.[2]

For site 2, all three receptors (2xvq, 2bxf, and 2bxe) used for IFD produced good results. Specifically, the docked poses agree well with the co-crystal structures, and the ligand RMSDs are within 2.0 Å. This speaks to the sampling efficiency of the IFD protocol and also reflects the relatively small size of the binding pocket for site 2. While all three receptors performed well, there are subtle differences between the docked poses. As such, we have used our discretion and chosen 2bxh as the best receptor and included this in the automated workflow discussed in Section 3.3.

Dansylsarcosine. In crystal structure (Figure 5a) dansylsarcosine is anchored into the binding site by hydrogen bonds with charged or polar residues, i.e., K414 and Y411. This is well

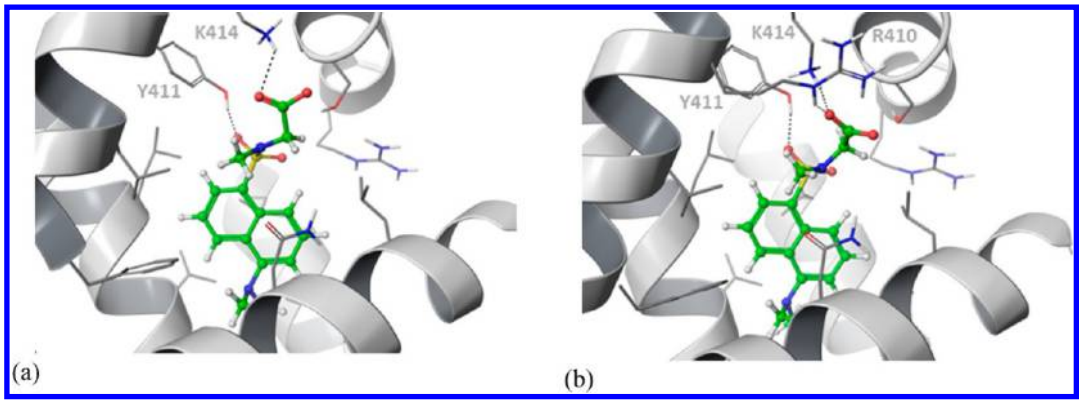


Figure 5. Dansylsarcosine crystal structure versus induced fit docking solution. (a) Crystal structure of dansylsarcosine within site 2 of HSA. PDB id 2xvq. Hydrogens have been added. R410 is not shown since this was not solved. (b) Induced fit docking structure of dansylsarcosine within site 2 of HSA. Note that R410 is shown since this was added using Prime.

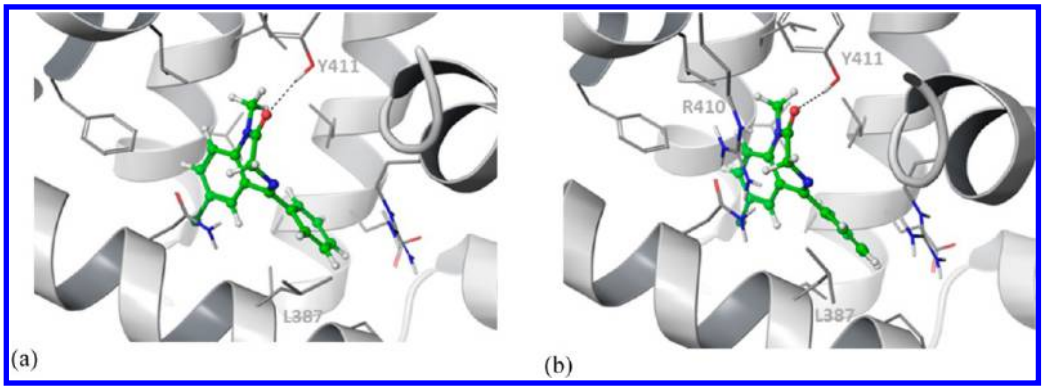


Figure 6. Diazepam crystal structure versus induced fit docking solution. (a) Crystal structure of diazepam within site 2 of HSA. PDB id 2bxf. Hydrogens have been added. R410 is not shown since this was not solved. (b) Induced fit docking structure of diazepam within site 2 of HSA. Note that R410 is shown since this was added using Prime.

reproduced by the IFD solution shown in Figure 5b. In the IFD solution, we see that R410 also serves to anchor dansylsarcosine into the site via hydrogen bonds. As the density of R410 is not available in the crystal structure, it is impossible to say with certainty whether this actually contributes to binding or is an artifact of the IFD. In both the crystal structure and IFD solutions shown in Figure 5, the *N*,*N*-dimethyl-naphthalino moiety of dansylsarcosine is buried in a hydrophobic pocket formed by leucine, isoleucine, phenylalanine, valine, and cysteine residues, where it forms hydrophobic contacts that are critical to its binding.

Diazepam. As with dansylsarcosine, diazepam is anchored into the binding site by hydrogen-bonding with Y411 as shown in Figure 6a. Furthermore, inspection of the crystal structure reveals how the diazepinone ring of dansylsarcosine is buried in a hydrophobic pocket comprised of phenylalanine, leucine, valine, isoleucine, and cysteine residues. The agreement between the crystal structure (Figure 6a) and the IFD solution (Figure 6b) is excellent. Hydrogen bonding to Y411 properly anchors the ligand into the binding site and hydrophobic interactions dominate in the apolar cavity. Minor rotation of L387 is observed, but all other critical contacts are reproduced with high integrity. Note that the side chain of R410 was not available in the crystal structure (Figure 6a). In our IFD structure (Figure 6b), this residue is not making any particularly strong contacts, nonbonded or otherwise, commensurate with it being solvent exposed.

Diflusinal. As with other site 2 ligands, diflusinal is anchored into the binding site by hydrogen-bonding interactions with polar and charged ligands, i.e., K414, S489, and Y411, as shown in Figure 7a. Further inspection of Figure 7a reveals that the difluoro phenyl moiety is buried within a hydrophobic pocket formed by phenylalanine, leucine, valine, and isoleucine residues, thus forming critical hydrophobic contacts that drive ligand binding. These features are well captured by our IFD solutions, where the top two solutions give ligand RMSDs of 2.01 and 1.46 Å. Compared to the crystal structure, the diflusinal in the highest-ranked IFD structure (Figure 7b) is shifted deeper into the hydrophobic pocket. Thus, while the critical hydrophobic contacts are well described, hydrogen bonding of the carboxylic acid to S489 is noticeably absent, as N391 serves as a surrogate hydrogen bond donor. The second-ranked IFD solution (Figure 7c) more closely mimics the crystallographic diflusinal pose. In this structure, the hydrogen-bonding contact to S489 is captured, and N391 serves again as a hydrogen bond donor. Interestingly, the temperature factor of N391 is quite high (ca. 65) suggesting that this residue may in fact have the flexibility to move into the position shown in our IFD solutions. This is reasonable considering the resolution of the reference crystal structure shown in Figure 7a is 2.95 Å, and the movement of N391 in our IFD solutions is less than 1 Å versus the crystal.

Site 1. Like site 2, site 1 is composed of an apolar cavity framed by polar features of helices 1, 2, and 6 of the subdomain.[2] Further, while the entrance to site 2 is open to solvent, the entrance to site 1 faces subdomain IIIA and has noticeably less solvent exposure.

dx.doi.org/10.1021/ci3006098 | *J. Chem. Inf. Model.* 2013, 53, 907−922

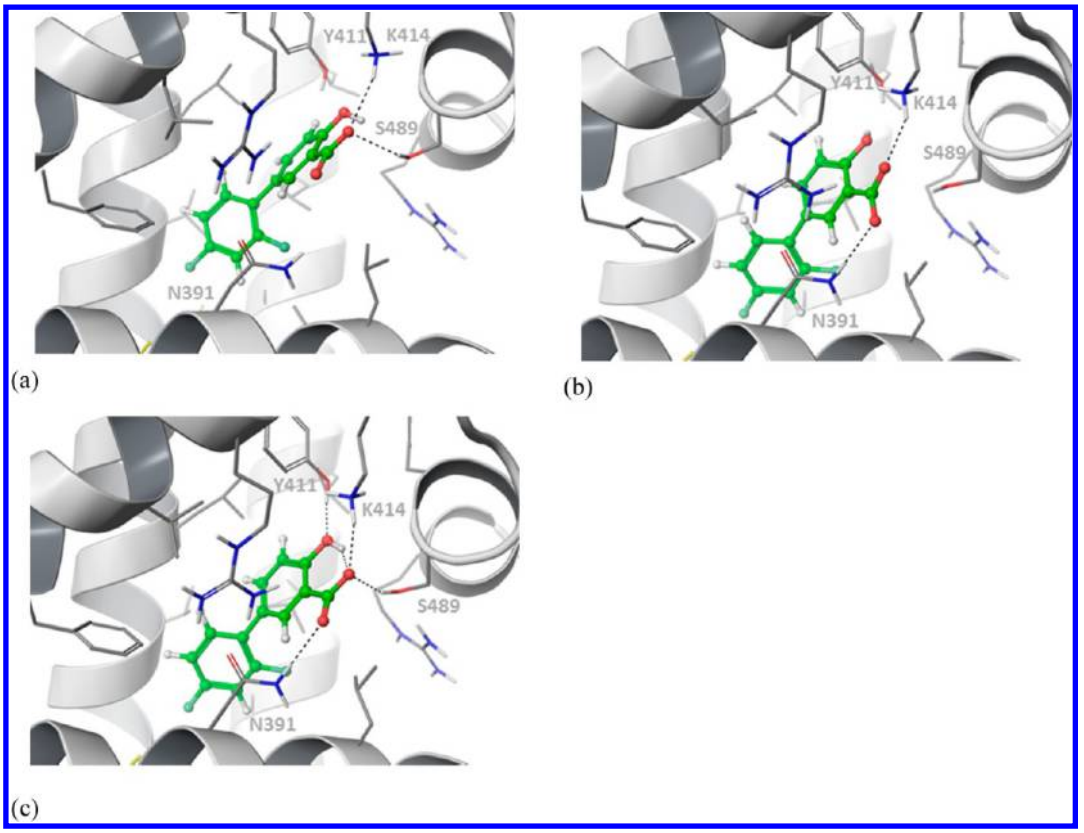


Figure 7. Diflusinal crystal structure versus induced fit docking solution. (a) Crystal structure of diflusinal within site 2 of HSA. PDB id 2bxf. Hydrogens have been added. (b) Highest ranked induced fit docking structure of diflusinal within site 2 of HSA. (c) Second-highest ranked induced fit docking structure of diflusinal within site 2 of HSA.

In contrast to site 2, site 1 preferentially binds large, heteroaromatic negatively charged ligands, where the charge is centralized.

Unlike the results for site 2, we see that correct prediction of ligand pose and interactions within site 1 is far more sensitive to the choice of receptor. This is consistent with the observation that site 1 is far more complex in structure than site 2.[2] The five receptors can be clustered, as they generally only give two ligand conformations in the binding site. These two clusters are comprised of (1) receptors 2bxh, 2xsi, and 3cx9 and (2) receptors 2bxn and 2xw0. There is in general no agreement in ligand pose between these two clusters. Inspection of Table 2 shows that for phenylbutazone and (*R*)- and (*S*)-warfarin, the first group of receptors (2bxh, 2xsi, and 3cx9) give the best results, whereas for indomethacin and thyroxine, the second group of receptors (2bxn and 2xw0) give the best results. Thus it is necessary to perform IFD with two receptors for site 1 (that is, ensemble IFD) and use discretion when choosing which resultant IFD structures to carry forward. While there are many similarities between the complex structures produced by the two groups of receptors (2bxh, 2xsi, and 3cx9 versus 2bxn and 2xw0), subtle differences nevertheless remain. Therefore, we have used our discretion and chosen the two best receptors from each group (2xw0 and 3cx9) for all subsequent discussion.

The necessity to perform ensemble IFD reflects the much larger size of the binding pocket in site 1 versus site 2.[2] In fact, site 1 can accommodate multiple ligands simultaneously (e.g., azapropazone and phenylbutazone, azapropazone and indomethacin, and indomethacin and myristic acid), and many co-crystal structures of ligands bound to site 1 show how these ligands do not completely fill the binding site.[9] In theory, the binding of multiple ligands simultaneously in site 1 can be simulated using iterative IFD runs, although we have not explored that option here.

Indomethacin. Indomethacin binding to site 1 is particularly interesting, as indomethacin occupies a different region of site 1 than the other ligands studied. This is facilitated by W214 flipping conformations to both open up the pocket and leave an exposed face for π-stacking interaction with the indole ring of indomethacin as shown in Figure 8a. By occupying the periphery of the binding pocket, indomethacin leaves room for the binding of additional ligands as seen in 2bxk and 2bxq, where indomethacin is cobound with azapropazone and phenylbutazone, respectively. To facilitate a direct comparison to the IFD results, we show only the crystal structure that has indomethacin bound without a coligand in Figure 8a.

Both the first- and the second-highest ranked IFD solutions (Figure 8b,c) agree well with the crystal structures. In all cases, indomethacin is anchored into the binding side by hydrogen bonding of protein side chains with the carboxylic acid, i.e. R218 and W214 in the crystal structures and K195 in the IFD solutions. Comparison of Figure 8a to Figure 8b,c shows how a 180° torsion of the exocyclic bond leads to the carboxylic acid hydrogen bonding to K195 instead of R218. The temperature factors for R218 and K195 in the crystals structures are quite high (ca. 70), suggesting the flexibility of these residues, making this solution one that is seemingly physically viable. The critical π-stacking interaction of the indole of indomethacin with W214 is captured in both IFD solutions, although W214 adopts a different conformation in Figure 8b,c versus the crystal structures. Additional hydrophobic interactions include the cation-π interaction of K199 with the indole ring of

dx.doi.org/10.1021/ci3006098 | *J. Chem. Inf. Model.* 2013, 53, 907−922

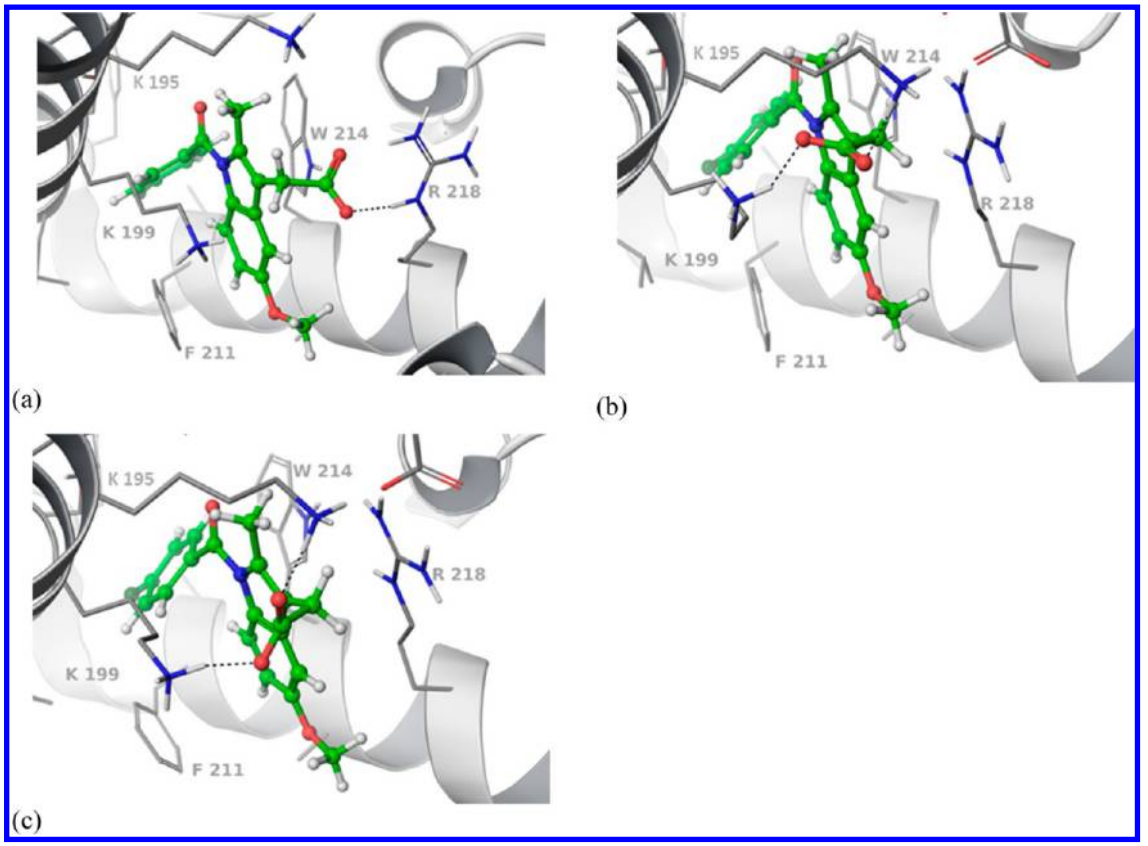


Figure 8. Indomethacin crystal structure versus induced fit docking solution. (a) Crystal structure of indomethacin within site 1 of HSA. PDB id 2bxm. Hydrogens have been added. (b) Highest ranked induced fit docking structure of indomethacin within site 1 of HSA using 2bxn as the receptor. (c) Second-highest ranked induced fit docking structure of indomethacin within site 1 of HSA using 2bxn as the receptor.

indomethacin and the edge-to-face π-stacking interaction of F211 with the para-chloro phenyl ring of indomethacin. Both of these critical interactions are captured by both IFD solutions as shown in Figure 8b,c.

Phenylbutazone. The three crystal structures of phenylbutazone in site 1 of HSA are highly conserved with the exception of the *n*-butyl tail of phenylbutazone. In 2bxc (Figure 9a), the *n*-butyl tail resides in the charged and polar pocket formed by E153, S192, and Q196 (not shown). This unfavorable interaction is presumably mitigated by stacking of the aliphatic chain against the face of Y150. Conversely, in structures 2bxp (Figure 9b) and 2bxq, this tail has rotated 180° to occupy a hydrophobic pocket made up by L238 along with F211 and H242 (not shown). By conventional wisdom, burying the *n*-butyl group in the hydrophobic pocket should facilitate stronger binding and indeed this is the pose that our IFD protocol prefers, as shown in Figure 9c,d. Interestingly, the binding of phenylbutazone in these two structures is highly conserved in spite of 2bxq having two ligands (phenylbutazone and indomethacin) in its binding site. In 2bxq, the accommodation of indomethacin in addition to phenylbutazone is accomplished by shifting R218 and R222 and flipping out W214 such that it forms an edge-to-face π-stack with the phenyl ring of indomethacin. This is just one example of the high degree of flexibility and voluminous nature that characterize site 1 and make it a challenging target for docking and other *in silico* studies.

Comparison of the IFD solutions shown in Figure 9c,d to the crystal structure shown in Figure 9b reveals that the ligand pose is well reproduced. The highest and second-highest ranked IFD solutions give RMSDs of 1.96 ± 0.30 and 2.44 ± 0.30 Å, respectively. In both IFD solutions (Figure 9c,d), the hydrogen bonding of phenylbutazone to R222 is reproduced, as seen in Figure 9b. Conversely, the hydrogen bonding interaction of the pyrazolone with H242 seen in Figure 9b is absent in the IFD solutions. However, this is a function of the initial preparation of the receptor structure since opposite protomors are observed for H242 in Figure 9b versus Figure 9c,d, where Nε and Nδ are protonated, respectively. Prior to docking, the receptor is prepared using the native ligand of the co-crystal structure, and thus the optimal protonation state of histidine (i.e., Nε versus Nδ) is determined by the native ligand. In cross-docking, there is no guarantee that the optimal protonation state obtained by optimizing the hydrogen-bonding network for the native ligand will be the same as that in the presence of a non-native ligand. Thus, short of sampling alternate protomers iteratively throughout the docking procedure or running ensemble docking with different protomers, one is limited to the protonation state that best fits the co-crystallized ligand. In promiscuous receptors such as HSA, this is presumably more problematic than in other receptors.

Hydrophobic interactions are a critical component to ligand binding, e.g., the π-stacking of the guanidino group of R257 and a phenyl ring of phenylbutazone. Although the highest-ranked IFD solution has a slightly lower ligand RMSD (1.96 versus 2.44 Å), the second solution better captures this critical interaction.

Thyroxine. The ligand pose in thyroxine is the most difficult to reproduce of all the molecules examined. The best IFD solution has an RMSD of 2.75 ± 0.16 Å versus the crystal structures. However, this problem may partially reside in the crystallographic poses themselves. The average resolution of the crystals

dx.doi.org/10.1021/ci3006098 | J. Chem. Inf. Model. 2013, 53, 907−922

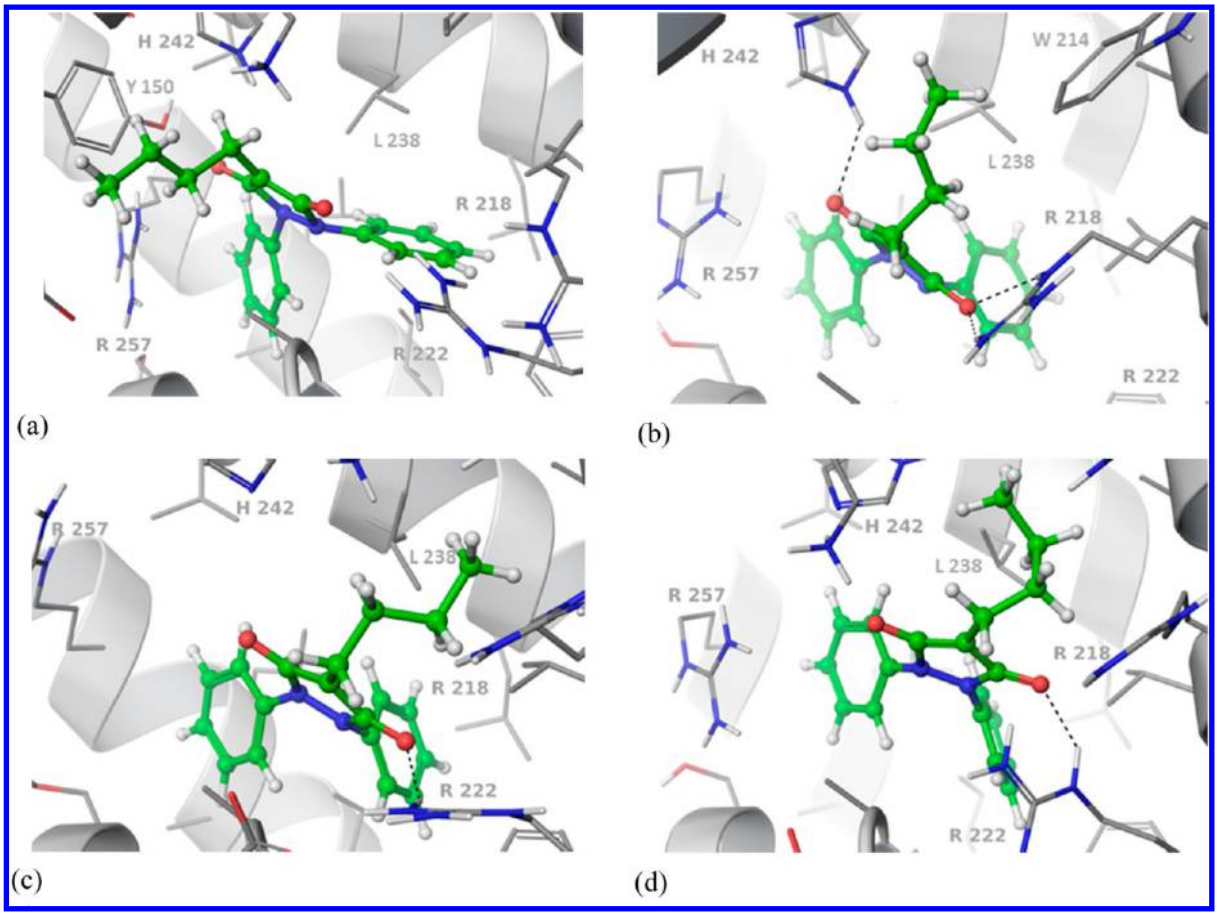


Figure 9. Phenylbutazone crystal structures versus induced fit docking solutions. (a) Crystal structure of phenylbutazone within site 1 of HSA. PDB id 2bxc. Hydrogens have been added. (b) Crystal structure of phenylbutazone within site 1 of HSA. PDB id 2bxp. Hydrogens have been added. (c) Highest ranked induced fit docking structure of phenylbutazone within site 1 of HSA. (d) Second-highest ranked induced fit docking structure of phenylbutazone within site 1 of HSA.

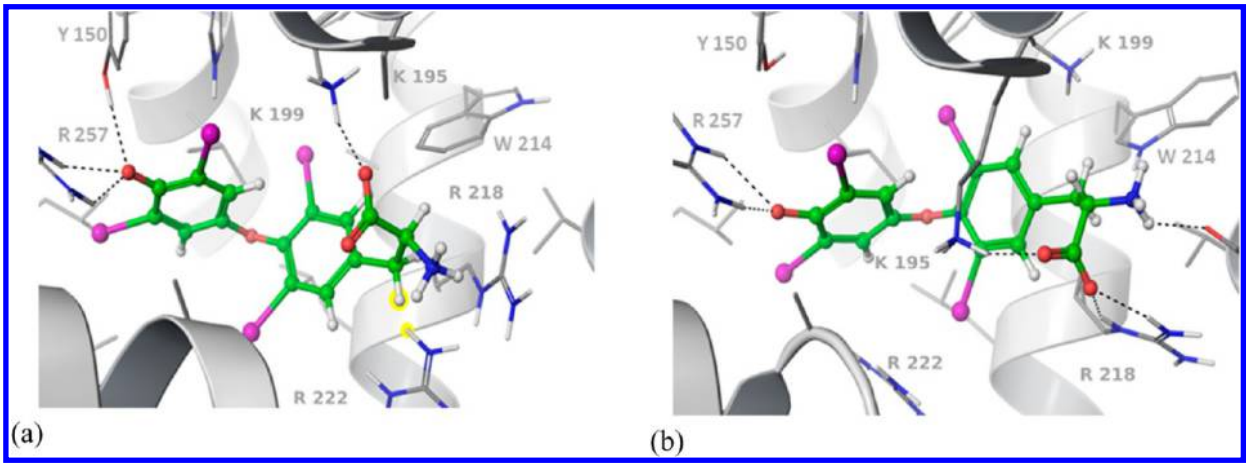


Figure 10. Thyroxine crystal structure versus induced fit docking solutions. (a) Crystal structure of thyroxine within site 1 of HSA. PDB id 1hk1. Hydrogens have been added. Note that the side chain of K195 is not shown since it was not available in the crystal structure. Unfavorable contacts are highlighted in yellow. (b) Highest ranked induced fit docking structure of phenylbutazone within site 1 of HSA. Note that K195 has been added using Prime.

(1hk1, 1hk2, 1hk3) is 2.75 ± 0.09 Å. Furthermore, inspection of the binding sites reveals unfavorable contacts owing to close contacts between nonbonding groups. For example, adopting a standard protonation protocol (e.g., protonating the solvent-exposed R222) places the hydrogens of R222 and thryoxine at a distance of less than 1 Å in the crystal structure 1hk1 as highlighted in Figure 10a. Thus, it is not surprising that in the IFD solution (Figure 10b), R222 has translated away from thyroxine to mitigate this unfavorable contact. Further, R222 has a high temperature factor (ca. 70), making movements of this residue quite physically reasonable.

dx.doi.org/10.1021/ci3006098 | J. Chem. Inf. Model. 2013, 53, 907−922

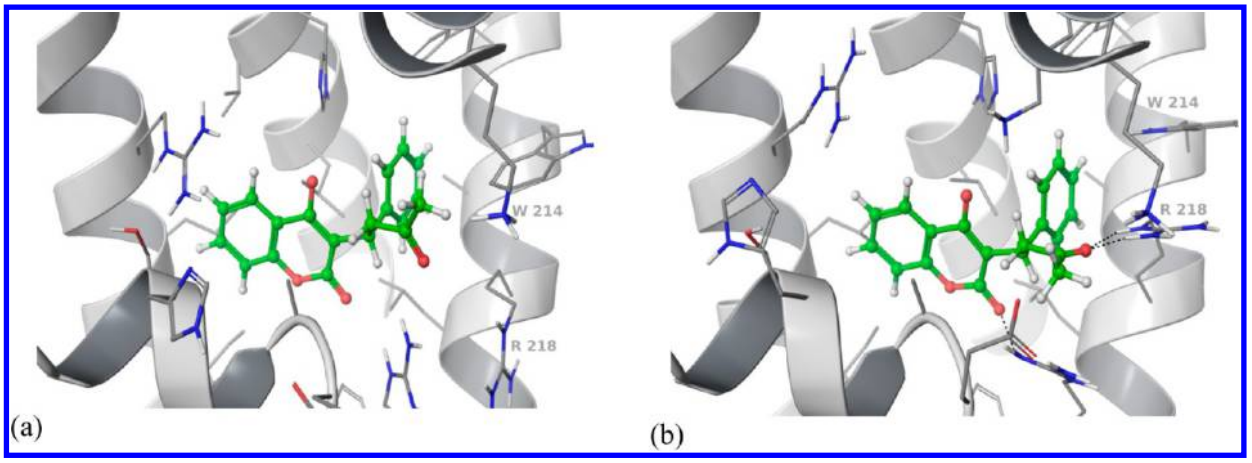


Figure 11. (*R*)-warfarin crystal structure versus induced fit docking solutions. (a) Crystal structure of (*R*)-warfarin within site 1 of HSA. PDB id 1h9z. Hydrogens have been added. (b) Highest ranked induced fit docking structure of (*R*)-warfarin within site 1 of HSA.

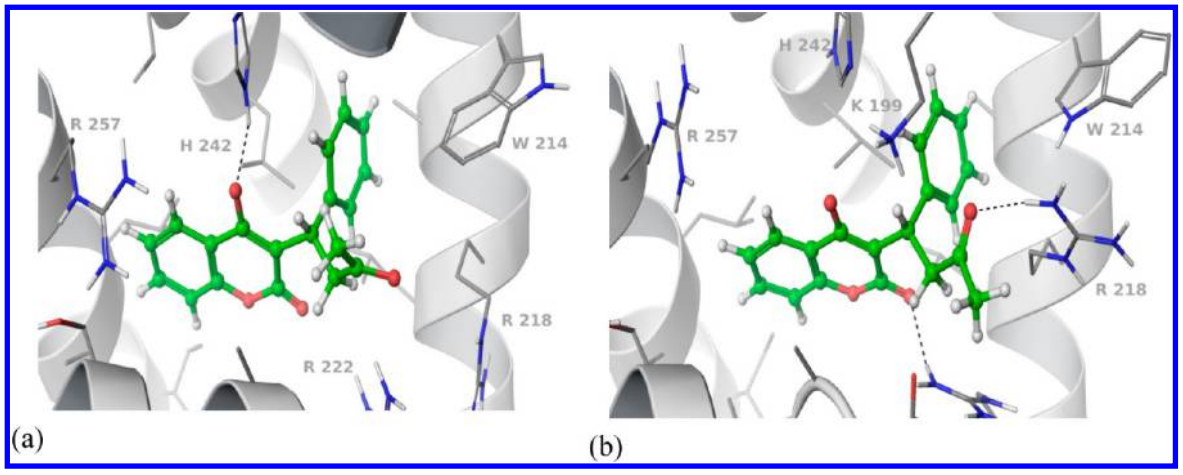


Figure 12. (*S*)-warfarin crystal structure versus induced fit docking solutions. (a) Crystal structure of (*S*)-warfarin within site 1 of HSA. PDB id 1ha2. Hydrogens have been added. The side chain of K199 is not shown since it was not available in the crystal structure. (b) Third-highest ranked induced fit docking structure of (*S*)-warfarin within site 1 of HSA. Note that K199 is shown since its side chain was added with Prime.

In spite of the relatively high ligand RMSD, the qualitative features that contribute to thyroxine binding to site 1 of HSA are well reproduced. Inspection of Figure 10 shows how in both structures, the central phenyl ring is anchored within a hydrophobic pocket formed by leucine, valine, alanine, and tryptophan residues. Furthermore, in both the IFD and crystal structure, the terminal phenoxy moiety is partially solvent-exposed and anchored inside the hydrophilic binding pocket created by R257 and Y150, while the carboxylic acid is also solvent exposed and tethered to a lysine (K199 or K195).

The density for K195 is not available in the crystal (Figure 10a), but the side chain has nevertheless been added using Prime for the IFD (Figure 10b). In the IFD solution, K199 is forming a cation-π interaction with W214, while K195 tethers to the carboxylic acid of thyroxine. Conversely, in all crystal structures (1hk1, 1hk2, and 1hk3), K199 tethers the carboxylic acid, not K195. The temperature factor of K199 is relatively low (ca. 35), suggesting that this is an interaction that is well conserved. Nevertheless, tethering the carboxylic acid of thyroxine to K195, instead of K199, allows for the carboxylic acid to be further tethered to R218 as shown in Figure 10b. R218 has a relatively high temperature factor (ca. 50), suggesting the flexibility of this residue. Further, this residue also makes unfavorable contacts with thyroxine in the crystal structure, where the positively charged ammonium and guanidino moieties are only separated by ca. 2 Å as highlighted in Figure 10a. Such unfavorable contacts should not be reproduced by any well-parametrized force field, and indeed that is what we observe.

(R)-Warfarin. There is good agreement between the crystallographic and IFD (*R*)-warfarin structure, where the ligand RMSD is 1.5 ± 0.6 Å. This is shown graphically in Figure 11. In both structures, the acetyl moiety is solvent exposed, while the aromatic lactone is buried within a hydrophobic pocket comprised of isoleucine, alanine, valine, and leucine residues. The primary differences in the side-chain interactions of the ligand with the two receptors (crystallographic and IFD) are seen in the solvent exposed residues including R218 and W214. In crystal structures of (*R*)-warfarin bound to HSA (see Figure 11a for example), the phenyl ring of (*R*)-warfarin is making an edge-to-face π-stack interaction with W214. In the IFD solution (Figure 11b), the centroid of W214 has shifted about 1 Å versus the crystal structure. R218 has shifted about 5 Å versus the crystal. Inspection of the crystal structures reveals the flexibility of R218, where its temperature factor is approximately 60, commensurate with the larger side chain movement seen here. In the IFD structure (Figure 11b), a cation-π interaction between W214 and R218 is formed as a consequence of the larger R218 movement. Thus, the phenyl ring of (*R*)-warfarin is still engaged

dx.doi.org/10.1021/ci3006098 | *J. Chem. Inf. Model.* 2013, 53, 907−922

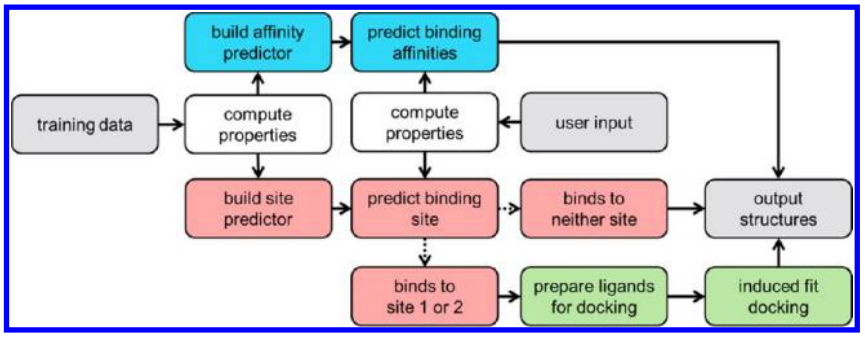


Figure 13. Schematic of automated workflow for predicting binding affinity (blue), site (pink), and pose (green) of any arbitrary, user-provided small molecule with HSA.

in a π-stack interaction; however, it is with R218. Therefore, the hydrophobic interactions which contribute to (*R*)-warfarin's high affinity for site 1 of HSA are well modeled, including burial of the aromatic lactone in the hydrophobic pocket and π-stacking of the phenyl group.

(S)-Warfarin. The binding of (*S*)- and (*R*)-warfarin is quite similar in spite of the difference in chirality as shown by comparison of Figure 11a and Figure 12a. Specifically, the phenyl and lactone rings remain buried in the same pocket in both structures, and only the partially solvent exposed ketone is affected by the difference in chirality. As with (*R*)-warfarin, there is good agreement between the IFD solution and the crystal structure for (*S*)-warfarin as shown in Figure 12 where the ligand RMSD is 2.10 Å. Differences can be seen in the positions of W214 and R218. These residues undergo the movements also seen in the IFD solution of (*R*)-warfarin. R218 is highly flexible (temperature factor of ca. 50) and is seen forming a hydrogen bond to the ketone of both (*R*)- and (*S*)-warfarin in the IFD solutions in Figure 11b and Figure 12b, respectively. The hydrogen bond interaction of the lactone of (*S*)-warfarin with H242 seen in the crystal structure is not present in the IFD solution. As discussed above for phenylbutazone, this presumably owes to a difference in protomer for H242 in the IFD receptor. Importantly, the critical hydrophobic contacts of the phenyl and lactone rings with the hydrophobic pockets are well captured by the IFD solution.

Summary. Our protocol accurately handles the two aspects of docking, pose prediction and scoring, as described above in detail for the test ligands. The average ligand RMSD is 1.8 Å for site 1 and 2, in line with other reports on the accuracy of IFD.[25] This is also reasonable considering the resolution of the crystal structures used for comparison, which are in the range of 2−3 Å. For site 2, the correct pose is correctly ranked by the scoring function as the top-scoring complex each time. Site 1, being larger, poses a bigger challenge, and the proper choice of receptor is critical to obtaining an accurate pose. Nevertheless, the correct pose is predicted within the first few predictions as ranked by score.

Overall our results suggest that the IFD protocol presented herein should be amenable to prospective studies. However, care must be taken, since the voluminous nature of site 1 necessitates that each molecule be docked into two separate site 1 receptors, as described above. These different IFD calculations generally produce dissimilar poses for the same molecule. Thus, chemical intuition must ultimately be used to pick the best receptor for the molecule of interest. Conversely, no such difficulty is encountered when performing IFD within site 2, owing to its relatively small size. Finally, whenever employing a previously developed protocol for a prospective study, one must take care that the system they wish to study is within the domain of applicability of the model previously developed. Therefore, we do not anticipate good results will be obtained when studying molecules substantially different from those shown in Figure 4.

This pose-prediction protocol facilitates a better understanding of the specific ligand-HSA interactions that contribute to binding affinity, as discussed in detail for the seven ligands shown in Figure 4b,c. Further, although IFD can require a few hours per molecule on a single processor,[40] it can be run in parallel (e.g., one processor per molecule or multiple processors per molecule), making the method medium-throughput and amenable to lead optimization. Alternatively, where computational resources are limited, consensus induced fit docking is an attractive and viable way to account for protein flexibility without increasing computational cost as shown by Kalid and Toledo Warshaviak,[41] although we have not explored this option for HSA in particular.

Further insight may be gained from more computationally expensive methods such as hybrid quantum mechanics molecular mechanics (QM/MM) and free energy perturbation (FEP), to name a few. As discussed in Section 1, rigorous *in silico* predictions of binding affinity and its contributing factors remain a challenging goal, especially for highly flexible systems such as HSA. Thus, we reserve the potential exploration of these facets for a subsequent publication.

3.4. Portable Workflow. Since ligand-HSA interactions are of great interest to drug discovery, we sought to provide a workflow that would be readily accessible to the structure-based drug design community. As such, we have implemented the binding affinity prediction (Section 3.1), site prediction (Section 3.2), and binding pose prediction (Section 3.3) into a KNIME workflow as shown in Figure 13. Within KNIME, we have used Schrödinger nodes since the linear regression model, Bayesian classifier, and induced fit docking protocols are seamlessly integrated into KNIME, and the induced fit docking protocol has shown good performance in past studies.[25,42,43] Furthermore, Schrödinger is a common software suite used by many modelers in the structure-based drug design community. Therefore, the tools within the workflow are familiar to many of those that could benefit from use of this workflow, facilitating further customization to suit the end user's need.

The workflow takes any arbitrary user-provided input molecules as SMILES strings and predicts their apparent binding affinity and binding sites. Binding pose prediction is then automatically performed within the appropriate binding sites, and the resultant complex structures are output. This workflow may be downloaded and run locally (on the user's machine or

dx.doi.org/10.1021/ci3006098 | *J. Chem. Inf. Model.* 2013, 53, 907−922

cluster) without an Internet connection and without the need to share the structures of the molecules with an outside server, thus protecting intellectual property.

The workflow is completely customizable. For example, since many pharmaceutical companies may have their own repositories of ligand-HSA binding site data, we have enabled the user to add their own training data to that described herein. The user can then subsequently rebuild the Bayesian classification models to presumably provide even more robust classification models than those described here. This is particularly useful with Bayes classifiers since they are sensitive to the definition of the prior. The user may also optimize the descriptors and fingerprints that are calculated to suit their particular needs. Additionally, if the user has proprietary co-crystal structures with HSA, these may be used as receptors for IFD in addition to those receptors discussed here. This is advantageous where the input molecules are of the same chemotype as those of the co-crystal. These are just a few examples of how the user may optimize the workflow to suit their particular needs; however, numerous other possibilities exist.

The time required to execute each step of the workflow varies depending upon the particular task. Building the Bayesian predictor and using the model to predict binding sites of novel molecules is quite fast and takes only a few minutes on a desktop computer. The binding affinity protocol is equally fast. The induced fit docking is notably slower, consistent with the computationally intensive nature of this task. This step may take several hours per molecule when run on a single processor.[40] However, the IFD step can be submitted to a cluster seamlessly from the KNIME workflow.

4. CONCLUSION

We present an *in silico* workflow for the elucidation of small molecule binding to HSA. This workflow gives a prediction of the binding affinity and site on HSA as well as the small molecule's pose within that particular site with good accuracy as shown on our test sets. The affinity, site, and pose prediction protocols employ disparate computational techniques (QSAR models, Bayesian classification, and induced fit docking, respectively) that together provide a comprehensive picture of HSA binding. The full workflow is available for download, enabling the end user to obtain the affinity, site, and pose of their small molecule within HSA upon input of their structure.

This has implications in structure-based drug design where small-molecule binding to HSA can destroy the efficacy of a drug *in vivo* that might otherwise have excellent activity. Furthermore, this workflow enables the user to obtain useful information about their small molecules' interaction with HSA relatively quickly without the need to set up specialized binding assay or crystallization workflows, thus facilitating lead optimization on a practical time scale.

5. METHODS

5.1. Binding Affinity Prediction. All affinity data were taken from that previously reported by Valko.[23] Molecular descriptors were computed using QikProp.[27] Multiple linear regression models were built using Canvas.[44]

5.2. Binding Site Prediction. All binding site data were obtained from the publically available data set previously curated by Kratochwil.[21] All molecular descriptors and fingerprints for the molecules of this data set were computed using Schrödinger's Cheminformatics program, Canvas.[44] The fingerprints computed were 32-bit, and the radius of the radial fingerprints was tested a various radii (4, 8, and 12). The Bayesian classifier described in Section 3.1 was also built using Canvas with categorical predictions using the aforementioned molecular descriptors and fingerprints.

5.3. Binding Pose Prediction. The coordinates for all proteins were obtained from the RCSB Protein Data Bank (PDB). Structures were prepared using Protein Preparation Wizard.[45] Where side chains were missing, Prime[38] was used to add them. All structures were aligned using the Protein Structure Alignment module in Prime. Hydrogen atoms were added, and all Asp, Glu, Arg, and Lys residues were left in their charged state. All ligands were prepared using LigPrep.[46]The induced fit docking (IFD) described in Section 3.2 was performed using the induced fit docking protocol and OPLS 2005 force field as implemented by Schrödinger using all default parameters and atomic partial charges.[38,47] Specifically, the centroid of the docking box was taken as the co-crystallized ligand in the respective receptor structure. No constraints were applied to the docking. Ligands were sampled during all docking, including ring conformations within a 2.5 kcal/mol window. Amide bonds were penalized in their nonplanar conformation. In the first stage (i.e., Glide docking), the receptor and ligand van der Waals forces were both scaled by 0.50 and 20 poses were output. In the second stage, i.e., Prime refinement, residues (including their side chains) with 5.0 Å of the docked ligand were refined. In the final stage, i.e., Glide redocking, ligands were redocked into complexes that were within 30 kcal/mol of the lowest energy structure and within the top 20 structures overall. All docking was performed using Glide SP.

5.4. Portable Workflow. The procedure described has been built as a KNIME[48] workflow and utilizes Schrödinger KNIME nodes to reproduce both the Bayesian classifier and induced fit docking.

■ ASSOCIATED CONTENT

Ⓢ Supporting Information

The data used to train the Bayesian model and binding affinity prediction. This material is available free of charge via the Internet at http://pubs.acs.org. Please contact the corresponding author to obtain a copy of the KNIME workflow.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: lewis.whitehead@novartis.com.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

M.L.H. wishes to thank (in alphabetical order) Greg Bebernitz, John Reilly, and Clayton Springer for insightful discussions.

■ REFERENCES

(1) Colmenarejo, G. In silico prediction of drug-binding strengths to human serum albumin. *Med. Res. Rev.* **2003**, *23*, 275−301.

(2) Curry, S. Lessons from the crystallographic analysis of small molecule binding to human serum albumin. *Drug Metab. Pharmacokinet.* **2009**, *24*, 342−357.

(3) Hall, L. M.; Hall, L. H.; Kier, L. B. Methods for predicting the affinity of drugs and drug-like compounds for human plasma proteins: a review. *Curr. Comput.-Aided Drug Des.* **2009**, *5*, 90−105.

(4) Murray, C. W.; Hartshorn, M. J. New Applications for Structure-Based Drug Design. In *Comprehensive Medicinal Chemistry II*; November 2006: 2007; Vol. *4*.

(5) Liu, X.; Chen, C.; Hop, C. E. C. A. Do we need to optimize plasma protein and tissue binding in drug discovery? *Curr. Top. Med. Chem.* **2011**, *2011*, 450−466.
(6) Wendt, M. D.; Shen, W.; Kunzer, A.; McClellan, W. J.; Bruncko, M.; Oost, T. K.; Ding, H.; Joseph, M. K.; Zhang, H.; Nimmer, P. M.; Ng, S.-C.; Shoemaker, A. R.; Petros, A. M.; Oleksijew, A.; Marsh, K.; Bauch, J.; Oltersdorf, T.; Belli, B. A.; Martineau, D.; Fesik, S. W.; Rosenberg, S. H.; Elmore, S. W. Discovery and structure−activity relationship of antagonists of B-cell lymphoma 2 family proteins with chemopotentiation activity in vitro and in vivo. *J. Med. Chem.* **2006**, *49*, 1165−1181.
(7) Oltersdorf, T.; Elmore, S. W.; Shoemaker, A. R.; Armstrong, R. C.; Augeri, D. J.; Belli, B. A.; Bruncko, M.; Deckwerth, T. L.; Dinges, J.; Hajduk, P. J.; Joseph, M. K.; Kitada, S.; Korsmeyer, S. J.; Kunzer, A. R.; Letai, A.; Li, C.; Mitten, M. J.; Nettesheim, D. G.; Ng, S.; Nimmer, P. M.; O'Connor, J. M.; Oleksijew, A.; Petros, A. M.; Reed, J. C.; Shen, W.; Tahir, S. K.; Thompson, C. B.; Tomaselli, K. J.; Wang, B.; Wendt, M. D.; Zhang, H.; Fesik, S. W.; Rosenberg, S. H. An inhibitor of Bcl-2 family proteins induces regression of solid tumours. *Nature* **2005**, *435*, 677−681.
(8) Mao, H.; Hajduk, P. J.; Craig, R.; Bell, R.; Borre, T.; Fesik, S. W. Rational design of diflunisal analogues with reduced affinity for human serum albumin. *J. Am. Chem. Soc.* **2001**, *123*, 10429−10435.
(9) Zsila, F.; Bikadi, Z.; Malik, D.; Hari, P.; Pechan, I.; Berces, A.; Hazai, E. Evaluation of drug−human serum albumin binding interactions with support vector machine aided online automated docking. *Bioinformatics* **2011**, *27*, 1806−1813.
(10) Andrisano, V.; Bertucci, C.; Cavrini, V.; Recanatini, M.; Cavalli, A.; Varoli, L.; Felix, G.; Wainer, I. W. Stereoselective binding of 2,3-substituted 3-hydroxypropionic acids on an immobilised human serum albumin chiral stationary phase: stereochemical characterisation and quantitative structure−retention relationship study. *J. Chromatogr., A* **2000**, *876*, 75−86.
(11) Andrisano, V.; Booth, T. D.; Cavrini, V.; Wainer, I. W. Enantioselective separation of chiral arylcarboxylic acids on an immobilized human serum albumin chiral stationary phase. *Chirality* **1997**, *9*, 178−183.
(12) Ashton, D. S.; Beddell, C.; Ray, A. D.; Valkó, K. Quantitative structure-retention relationships of acyclovir esters using immobilised albumin high-performance liquid chromatography and reversed-phase high-performance liquid chromatography. *J. Chromatogr., A* **1995**, *707*, 367−372.
(13) Ashton, D. S.; Beddell, C. R.; Cockerill, G. S.; Gohil, K.; Gowrie, C.; Robinson, J. E.; Slater, M. J.; Valko, K. Binding measurements of indolocarbazole derivatives to immobilised human serum albumin by high-performance liquid chromatography. *J. Chromatogr., B: Biomed. Sci. Appl.* **1996**, *677*, 194−198.
(14) Kaliszan, R. Retention data from affinity high-performance liquid chromatography in view of chemometrics. *J. Chromatogr., B: Biomed. Sci. Appl.* **1998**, *715*, 229−244.
(15) Kaliszan, R.; Noctor, T.; Wainer, I. Quantitative structure-enationselective retention relationships for the chromatography of 1,4-benzodiazepines on a human serum albumin based HPLC chiral stationary phase: An approach to the computational prediction of retention and enantioselectivity. *Chromatographia* **1992**, *33*, 546−550.
(16) Markuszewski, M.; Kaliszan, R. Quantitative structure−retention relationships in affinity high-performance liquid chromatography. *J. Chromatogr., B* **2002**, *768*, 55−66.
(17) Hanai, T.; Koseki, A.; Yoshikawa, R.; Ueno, M.; Kinoshita, T.; Homma, H. Prediction of human serum albumin−drug binding affinity without albumin. *Anal. Chim. Acta* **2002**, *454*, 101−108.
(18) Hanai, T.; Miyazaki, R.; Kinoshita, T. Quantitative analysis of human serum albumin−drug interactions using reversed-phase and ion-exchange liquid chromatography. *Anal. Chim. Acta* **1999**, *378*, 77−82.
(19) Deeb, O.; Hemmateenejad, B. ANN-QSAR model of drug-binding to human serum albumin. *Chem. Biol. Drug Des.* **2007**, *70*, 19−29.
(20) Saikhov, R. D.; Stefan, L. R.; Klopman, G. Multiple computer-automated structure evaluation model of the plasma protein binding affinity of diverse drugs. *Perspect. Drug Discovery Des.* **2000**, *19*, 133−155.
(21) Kratochwil, N. A.; Huber, W.; Müller, F.; Kansy, M.; Gerber, P. R. Predicting plasma protein binding of drugs: a new approach. *Biochem. Pharmacol.* **2002**, *64*, 1355−1374.
(22) Hajduk, P. J.; Mendoza, R.; Petros, A. M.; Huth, J. R.; Bures, M.; Fesik, S. W.; Martin, Y. C. Ligand binding to domain-3 of human serum albumin: a chemometric analysis. *J. Comput.-Aided Mol. Des.* **2003**, *17*, 93−102.
(23) Valko, K.; Nunhuck, S.; Bevan, C.; Abraham, M. H.; Reynolds, D. P. Fast gradient HPLC method to determine compounds binding to human serum albumin. Relationships with octanol/water and immobilized artificial membrane lipophilicity. *J. Pharm. Sci.* **2003**, *92*, 2236−2248.
(24) Deeb, O.; Rosales-Hernández, M. C.; Gómez-Castro, C.; Garduño-Juárez, R.; Correa-Basurto, J. Exploration of human serum albumin binding sites by docking and molecular dynamics flexible ligand−protein interactions. *Biopolymers* **2010**, *93*, 161−170.
(25) Sherman, W.; Day, T.; Jacobson, M. P.; Friesner, R. A.; Farid, R. Novel procedure for modeling ligand/receptor induced fit effects. *J. Med. Chem.* **2005**, *49*, 534−553.
(26) Colmenarejo, G.; Alvarez-Pedraglio, A.; Lavandera, J.-L. Cheminformatic models to predict binding affinities to human serum albumin. *J. Med. Chem.* **2001**, *44*, 4370−4378.
(27) *QikProp*, version 3.5; Schrödinger, LLC: New York, NY, 2012.
(28) Jorgensen, W. L. *QikProp, Technical Manual*; Schrodinger, Inc.: Portland, OR, 2006.
(29) Duan, J.; Dixon, S. L.; Lowrie, J. F.; Sherman, W. Analysis and comparison of 2D fingerprints: insights into database screening performance using eight fingerprint methods. *J. Mol. Graphics Modell.* **2010**, *29*, 157−170.
(30) Rogers, D. J.; Tanimoto, T. T. A computer program for classifying plants. *Science* **1960**, *132*, 1115−8.
(31) Reilly, J.; Etheridge, D.; Everatt, B.; Jiang, Z.; Aldcroft, C.; Wright, P.; Clemens, I.; Cox, B.; Press, N. J.; Watson, S.; Porter, D.; Springer, C.; Fairhurst, R. A. Studies in drug albumin binding using HSA and RSA affinity methods. *J. Liq. Chromatogr. Relat. Technol.* **2011**, *34*, 317−327.
(32) Rogers, D.; Hahn, M. Extended-connectivity fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742−754.
(33) Sivia, D.; Skilling, J. *Data Analysis: A Bayesian Tutorial*, 2nd ed.; Oxford University Press, Inc.: New York, NY, 2006.
(34) Armstrong, N.; Hibbert, D. B. An introduction to Bayesian methods for analyzing chemistry data: part 1: An introduction to Bayesian theory and methods. *Chemom. Intell. Lab. Syst.* **2009**, *97*, 194−210.
(35) Hibbert, D. B.; Armstrong, N. An introduction to Bayesian methods for analyzing chemistry data Part II: A review of applications of Bayesian methods in chemistry. *Chemom. Intell. Lab. Syst.* **2009**, *97*, 211−220.
(36) Kullback, S.; Leibler, R. A. On information and sufficiency. *Ann. Math. Stat.* **1951**, *22*, 79−86.
(37) *Glide*, version 5.8; Schrödinger, LLC: New York, NY, 2012.
(38) *Prime*, version 3.1; Schrödinger, LLC: New York, NY, 2012.
(39) Irving, J. A.; Whisstock, J. C.; Lesk, A. M. Protein structural alignments and functional genomics. *Proteins: Struct., Funct., Bioinf.* **2001**, *42*, 378−382.
(40) *Induced Fit Docking Manual*; Schrödinger, LLC: 2009.
(41) Kalid, O.; Toledo Warshaviak, D.; Schechter, S.; Sherman, W.; Shacham, S. Consensus induced fit docking (cIFD): methodology, validation, and application to the discovery of novel Crm1 inhibitors. *J. Comput.-Aided Mol. Des.* **2012**, *26*, 1217−1228.
(42) Farid, R.; Day, T.; Friesner, R. A.; Pearlstein, R. A. New insights about HERG blockade obtained from protein modeling, potential energy mapping, and docking studies. *Bioorg. Med. Chem.* **2006**, *14*, 3160−3173.
(43) Sherman, W.; Beard, H. S.; Farid, R. Use of an induced fit receptor structure in virtual screening. *Chem. Biol. Drug Des.* **2006**, *67*, 83−84.
(44) *Canvas*, version 1.5; Schrödinger, LLC: New York, NY, 2012.

(45) *Schrödinger Suite 2011 Protein Preparation Wizard*, Epik version 2.5; Schrödinger, LLC: New York, NY, 2012.
(46) *LigPrep*, version 2.5; Schrödinger, LLC: New York, NY, 2012.
(47) *Schrödinger Suite 2012 Induced Fit Docking protocol*, Glide version 5.8; Schrödinger, LLC: New York, NY, 2012.
(48) Berthod, M. R.; Cebron, N.; Dill, F.; Gabriel, T. R.; Kötter, T.; Meinl, T.; Ohl, P.; Sieb, C.; Thiel, K.; Wiswedel, B. *KNIME: The Konstanz Information Miner 2007*; Springer: 2007.