# 11

# Understanding Protein-Ligand Interactions Using Simulated Annealing in Dimensionally Reduced Fingerprint Representation

Ravi K. Nandigam and Sangtae Kim
*Purdue University School of Chemical Engineering*
*USA*

## 1. Introduction

*Structure-based drug design* is a rational approach for drug discovery based on understanding of the three dimensional structural interactions between a target protein and the drug-like ligands. The underlying premise is that good drug-like molecules must possess structural and chemical features complementary to that of the target receptor, which is usually a protein involved in the disease process. The process first involves identification of the protein target that is of interest. The structure of the target protein is then determined using experimental procedures like NMR, X-ray crystallography or computational approaches like homology modeling. After determining the structure of the target, the structural knowledge is used to systematically search the chemical space for compounds (or ligands) that would bind to the protein in the desired binding mode using docking techniques. These compounds are scored and ranked using scoring functions that take into account factors that could influence the nature of the binding such as steric and electrochemical interactions, exposed surface area, molecular weight, etc. The challenge in the search for the desired ligands is the ability to accurately model and analyze the protein-ligand binding by understanding the structural and chemical characteristics of the protein's binding site from theory, computation and experiment.

The amount of protein-ligand structural data available in public domain and corporate databanks increased exponentially during the last two decades due to significant advances in high throughput experimental techniques and computation power. In addition, there are many more structures that remain undisclosed due to proprietary interests. It is expected to have many more X-ray crystal structures to be available in the near future due to advances in high-throughput techniques and other experimental sophistications. In addition, there are also structures that are computationally generated through docking, or similar techniques. A typical virtual chemical library screen could generate a library of structures containing thousands to millions of small molecules docked onto a target protein in silico (Lyne, 2002). As discussed before the key to success in the rational drug design process is the proper understanding of the receptor site and the mode(s) in which ligands bind to the receptor by leveraging the available structural data. Traditionally, this is done by making logical deductions after visually inspecting the protein ligand complex on a computer or sometimes

aided by software tools like LIGPLOT (Wallace et al., 1995) that generate two dimensional schematic representations of the interactions. However the traditional approach is impractical when the number of structures to be analyzed is very large. In such scenarios, there is requirement for an automated way of detecting the various interaction patterns between the protein and the ligand, representing them in an efficient manner such that different protein ligand complexes can be compared and if possible correlated with their actual binding constants. The interaction patterns so identified from the structural data can eventually help to develop virtual screening and other design tools to aid the search for new drugs i.e. ligands with desired characteristics.

### 1.1 Structural Interaction Fingerprint (SIFt)

Fingerprint based approaches have been developed recently in the cheminformatics domain to mine, analyze, organize and visualize the vast structural binding data. They involve representing the three dimensional protein-ligand structural binding information into a one-dimensional vector by encoding the nature of interactions between binding site residues and the ligand as in Structural Interaction Fingerprint or SIFt (Deng et al., 2004; Chuaqui et al., 2005; Singh et al., 2006). Since the binding information is encoded in a 1D fingerprint, advanced filtering, clustering, and machine learning methods may be applied to identify patterns underlying the binding data, thereby enhancing the ability to make useful implications that are not apparent by looking at individual structures. There are also other fingerprint approaches published in literature such as atom-pairs based interaction fingerprint (Pérez-Nueno VI et al., 2009), pharmacophore based fingerprint (Sato et al., 2010), etc. This chapter demonstrates the use of advanced mathematical and statistical learning techniques to enhance the understanding of binding interactions from fingerprints. Though the methods explained here are in the context of SIFt, they can be applied to other fingerprint approaches.

A SIFt is generated from a protein-ligand complex by first identifying the key residues of the receptor protein, which are the residues that could potentially be involved in binding with a ligand. The key residues are identifying by performing a rigorous search among all known protein-ligand complexes of the target protein for residues that are involved in binding in at least one complex. The next step involves representing of each key residue by a bit pattern corresponding to the kind of interaction that is being made at that residue by the ligand. The first bit is a master bit that checks if an interaction is present at all or not. The second and third bits check if the interaction is with the main chain or side chain portions of the residue. The next four bits characterize the chemical nature of the interaction. The fourth and fifth bits are turned ‘on’ or ‘off’ corresponding to whether the residue is involved in a polar or non-polar interaction respectively, while one of the sixth and seventh bits is turned ‘on’ if there is a hydrogen bond interaction depending on whether the residue has a functional group that is an acceptor or a donor. The bit strings from all the residues are concatenated to form a fingerprint (called SIFt) which is a unique representation for that protein-ligand complex, as shown in Figure 1.

The overall pattern of interactions in a set of structures can be represented by an interaction profile where each element or entry in the profile speaks about the nature of interactions of the entire set. A profile based on the conservation or frequency of a bit over the set of fingerprints was used in (Chuaqui et al., 2005). They demonstrated by comparing the profiles of protein complexes belonging to different kinase targets viz. p38 and CDK2, one

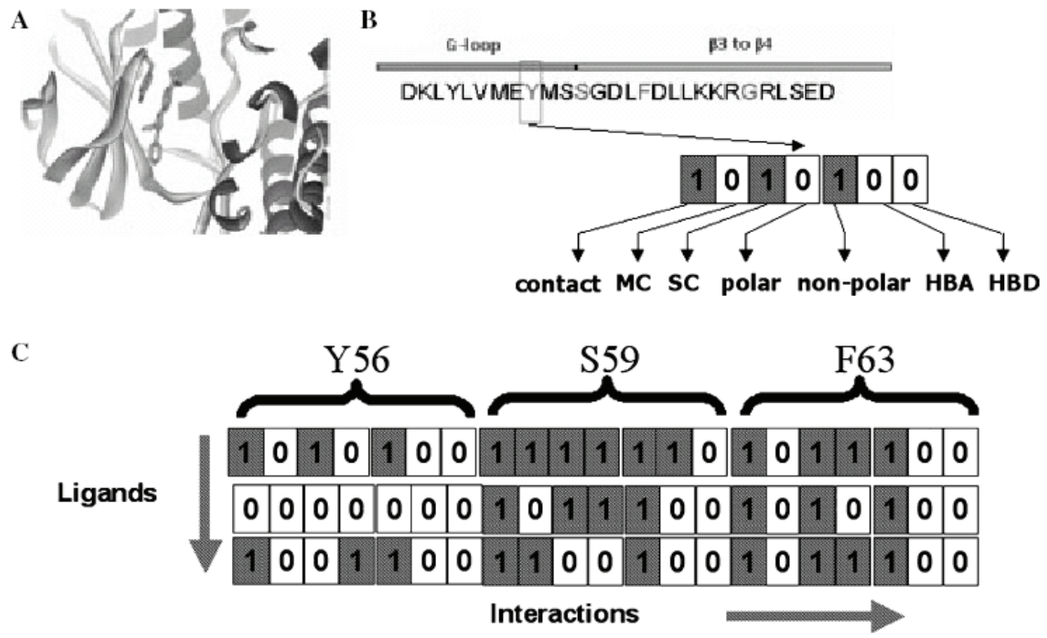



Fig. 1. An illustrative showing the SIFt methodology. (A) identify the key binding residues of the receptor protein in the complex. (B) represent each key residue by a bit string according to the kind of interaction at that residue. (C) concatenate 7-bit strings of all key residues to form a unique fingerprint, called SIFt. (Figure reprinted in part with permission from Singh J et al., 2005).

can identify the characteristic role played by the individual interactions in the overall binding. Interaction fingerprints and profile-based methods have been applied to virtual screening, library design, and the analysis of large numbers of X-ray structures to identify interaction patterns that may influence inhibitor potency and selectivity. The evolution of interaction fingerprint and profile approaches and their application to docking, scoring, and the analysis of ligand-receptor interactions has been comprehensively reviewed recently by Brewerton (Brewerton, 2008).

### 1.2 Weighted interactions profile

The original plain fingerprint is a simplified representation of protein-ligand interactions with all interactions being treated identical. But in reality the various possible interactions at different residues might have different contributions towards the overall binding. As an example, it is well-known that in kinases the interactions at the hinge region are critical for binding compared to interactions at other regions. Likewise, a hydrogen bond interaction could have a different impact compared to a polar or nonpolar interaction. By not capturing the information pertaining to the interactions differently from each other, their relative importance information is in effect lost. Hence the fingerprint representation is inefficient due to underrepresentation of significant interaction information and overrepresentation of insignificant interaction information. A new weighted interactions based approach called weighted Structural Interaction Fingerprint (wSIFt) was introduced in (Nandigam et al., 2009) to address this inefficiency of fingerprint representation. In the wSIFt method a robust representation signifying the relative importance of ligand receptor interactions is captured

in the form of a weights vector called weighted profile, where each weight corresponds to the importance of that interaction in overall binding. The weighted profile incorporates empirically determined weights fit from inhibitor potency data. The profile weights are determined such that the fingerprint similarity between docked poses and the weighted profile is in effect a residue-specific QSAR based on the relative importance of ligand-receptor interactions for determining potency.

The chapter describes the wSIFt methodology developed by Nandigam et al. to determine a weighted profile capturing the significance of interactions. The weights are determined using a statistical learning technique from structural data and experimental potency data such that the similarity between the weighted profile and a SIFt (called wSIFt score) is positively correlated with its experimentally determined inhibition potency. The mathematical formulation to determine the weights is an optimization problem with the objective to be maximized being the correlation between the wSIFt score and the inhibitor potency. Since the objective function is complex and non-linear, and the number of variables (i.e. weights) to determine is very large, a stochastic optimization technique (Simulated Annealing) is applied. The dimensionality of SIFt interaction bits is large and the representation contains linearly interdependent interaction bits and hence a dimensionality reduction technique called Nonnegative Matrix Factorization (NMF) is combined with the stochastic optimization stage. The subsequent sections of the chapter describe the methods including the strategy of the overall algorithm, dimensionality reduction and Simulated Annealing, followed by results and analysis of the weights.

## 2. Methods

### 2.1 Overall approach

The weighted profile is assumed to contain non-negative weights with values ranging between 0 and 1 at positions that have a 1 in at least one of the SIFts, and a value of 0 at positions that do not have a 1 in at least one of the SIFts (as shown in Figure. 2).

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| SIFt 1 | 1 | 0 | 0 | 1 | 1 | 0 | 0 | 1 | 0 | 1 |
| SIFt 2 | 1 | 1 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 1 |
| SIFt 3 | 1 | 1 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 1 |
| SIFt 4 | 1 | 1 | 0 | 0 | 1 | 0 | 1 | 1 | 0 | 1 |
| SIFt 5 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 1 | 0 | 1 |
| W-Profile | $w_1$ | $w_2$ | 0 | $w_4$ | $w_5$ | 0 | $w_7$ | $w_8$ | 0 | $w_{10}$ |

Fig. 2. Illustration of weighted profile for a set of interaction fingerprints.

The objective is to determine the weights such that the computed weights will represent the significance of each interaction in contributing toward overall protein-ligand binding. This can be achieved by statistically learning the weights from a training set such that the similarity between the weighted profile and SIFt is positively correlated with the inhibition potency. The reasoning behind the proposed approach is that the interactions appearing more frequently in high potent compounds are supposedly more important, and so in order to boost the w-SIFt score of the high potent compounds the weights for those interactions

will be calculated to be higher. Likewise, interactions that appear more frequently in less potent compounds are supposedly less important, and so in order to decrease the w-SIFt score of the less potent compounds these interactions' weights will be lower. Thus the overall weights in the weighted profile so determined will represent the importance associated with each SIFt interaction bit in the protein-ligand binding potency. The Tanimoto score is used here as the metric to measure the similarity between the weighted profile and SIFt, and for a given SIFt we call this metric the w-SIFt score. Thus a protein-ligand complex with a higher w-SIFt score implies that it comprises of interactions predominantly at higher weight bit positions and so the ligand would be a strong inhibitor of the protein, and likewise a complex with lower w-SIFt score implies that it comprises of interactions mainly at lower weight bit positions and so the ligand would be a weak inhibitor. The proposed strategy to determine the weighted profile can be graphically visualized as in Figure 3. Suppose the SIFts can be represented as points in a high dimensional hyperspace, the desired weighted profile should be more similar to the high potent compounds and less similar to the low potent compounds. In other words the weighted profile should be as closer as possible to the high potent compounds and as farther as possible to the low potent compounds in the SIFt coordinate space.

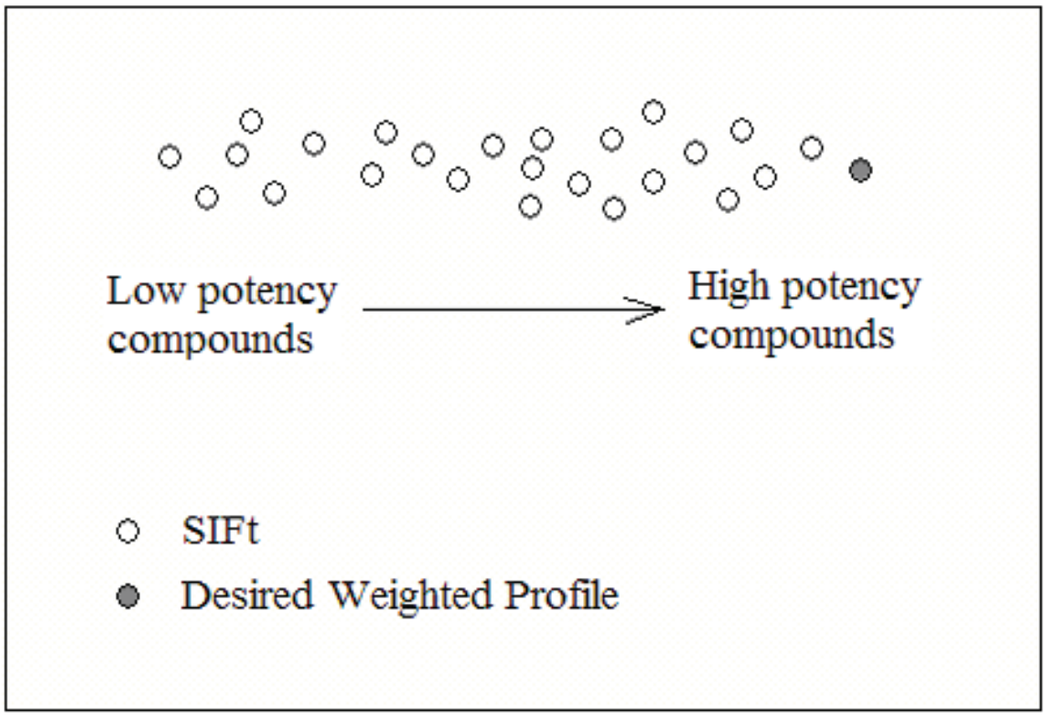



Fig. 3. Illustration of proposed weighted profile relative to the high potent and low potent SIFts in a hypothetical high dimensional hyperspace.

### 2.2 Mathematical objective function

Assume $\mathbf{s}$ represents a SIFt in a vector form and $\mathbf{w}$ represents the weighted profile. The w-SIFt score, let us call $T_w$, is defined as the Tanimoto similarity between the SIFt and the profile. i.e.

$$T_w = \frac{\mathbf{s}\cdot\mathbf{w}}{\mathbf{s}\cdot\mathbf{s}+\mathbf{w}\cdot\mathbf{w}-\mathbf{s}\cdot\mathbf{w}}\,.$$

The weights of the profile will be determined so as to obtain a w-SIFt score that correlates well with the experimentally determined potencies. We constrain the weights to be positive since in principle they represent the significance of the corresponding interactions. The objective of determining the weights can be mathematically stated as follows.

*To determine* $\mathbf{w}$ *so that* $T_w \propto -Log(IC50)$, *with* $w_i \geq 0$.

*i.e.* find $\mathbf{w}$ that corresponds to a straight line fit between $T_w$ and $-Log(IC50)$ with highest correlation. The Pearson's correlation coefficient is chosen here to measure the extent of correlation. So, the objective function is formulated as,

$$\underset{w_i \text{ s.t. } w_i \geq 0}{Maximize} \quad CorrCoef(T_w, -Log(IC50)) \tag{1}$$

where $CorrCoef(x,y) = \dfrac{\mathrm{cov}(x,y)}{\sqrt{\mathrm{var}(x)\cdot\mathrm{var}(y)}}$

Since the objective function is complex and non-linear, and the number of variables (i.e. weights) to alter is very large, we apply a stochastic optimization technique *viz.* Simulated Annealing (Kirkpatrick, Gelatt et al. 1983). The energy function for Simulated Annealing is defined here as the negative of the objective function defined in Equation 1.

$$E_w = -CorrCoef\left(Tw, -Log(IC50)\right) \tag{2}$$

### 2.3 Linear dimensionality reduction

The dimensionality of the SIFt bits is the number of binding region residues times the number of interaction bits per residue. Typically this number is high, for e.g. in the case of P38α the number of bits in SIFt is 560 as discussed in the Dataset Generation subsection. Even after eliminating the zero valued bit positions the number of bit positions whose weights need to be determined is large. However not all the interactions at the non-zero bit positions are independent of each other, as there could be co-occurrences (i.e. two bits simultaneously 'on' or 'off') and cross-occurrences (i.e. bits that are complementary to each other). There could also be additional statistically significant dependencies between bit pairs, i.e. two bit positions positively or negatively highly correlated within the data. So a dimensionality reduction technique is used here to reduce if not eliminate these interdependencies and eventually compress the number of SIFt bits to a considerably smaller number without losing significant information. Thus, by doing so the number of weight parameters to be determined in the weighted profile is also significantly reduced. As the interdependencies in the SIFt are linear, we choose a linear dimensionality reduction technique for the data compression. The values of the SIFts in the reduced space need not be binary, but have to be positive. We now only have as many weights to be determined as the dimension of the reduced space. After determining the weights in the lower dimensional space, the weights in the higher dimensional space (i.e. the original weights of the SIFts) can be obtained by an inverse transformation.

A linear dimensionality reduction technique involves transformation or rotation of the vector coordinate space such that the original data vector of higher dimensionality can be represented by another vector of lower dimensionality. Assume the original SIFt vector is represented as $\mathbf{s}^h$ of dimensionality $n$, let $\mathbf{s}^l$ be its representation in a lower vector space of dimensionality $r$, and $\mathbf{L}$ be the dimensionality reduction transformation. Then,

$$\mathbf{S}^h \approx \mathbf{L} \cdot \mathbf{S}^l$$

Suppose the full SIFt dataset is represented as an $n \times m$ matrix, $\mathbf{S}^h$ where $m$ is the number of SIFts. During linear dimensionality reduction the matrix $\mathbf{S}^h$ is in effect factorized into two sub-matrices $\mathbf{L}$ and $\mathbf{S}^l$ of size $n \times r$ and $r \times m$ respectively *i.e.*

$$\underset{n\times m}{\underbrace{\mathbf{S}^h}} \approx \underset{n\times r}{\underbrace{\mathbf{L}}} \cdot \underset{r\times m}{\underbrace{\mathbf{S}^l}} \qquad \text{where } (n+m)r < nm$$

The matrix $\mathbf{S}^l$ represents the $m$ SIFts in the lower dimensional space.

Dimensionality reduction techniques such as Nonnegative Matrix Factorization (NMF), Principal Component Analysis (PCA), and Vector Quantization (VQ) differ in the nature of the factor matrices. NMF involves a factorization such that the end sub-matrices are nonnegative. PCA involves a factorization such that the $\mathbf{L}$ matrix corresponds to a transformation into the Eigen vector coordinate system, whereas in VQ the factorization is such that the vectors of the transformed matrix are all unary. NMF is used here for dimensionality reduction of the SIFt space as the nonnegative constraint imposed in this method helps to preserve the underlying physical interpretation of the weights.

Lee and Seung (Lee and Seung 1999) demonstrated that NMF involves parts based learning of objects, and is very effective and meaningful for dimensionality reduction in applications like image processing and text mining. NMF has been applied in several recent works in the context of computational biology and bioinformatics. Gao and Church (Gao and Church 2005) applied NMF as an unsupervised classification method for cancer identification based on gene expression data, and found the method to be effective over other clustering techniques. Brunet et al. (Brunet et al., 2004) have also used NMF on cancer related microarray data. The basis vectors in their work, called meta genes, represented distinct molecular patterns thus enabling them to extract meaningful biological information. In Ref. (Kim and Tidor 2003) NMF was used on a large dataset of genome-wide expression measurements of yeast and was able to detect local features in the expression space that mapped to functional cellular subsystems. Recently, Devarajan (Devarajan 2008) provided a review of recent NMF applications in the context of biological informatics

When NMF is applied to SIFts, the basis vectors represent underlying patterns of interactions between protein and the ligands as explained in Results section. The algorithm for solving NMF based on the following update rules as described by Lee and Seung (Lee and Seung 2001) is used here for the dimensionality reduction.

$$L_{ia} \leftarrow L_{ia} \frac{\left(S^h S^{lT}\right)_{ia}}{\left(LS^h S^{lT}\right)_{ia}} \tag{3}$$

$$L_{ia} \leftarrow \frac{L_{ia}}{\sum_j L_{ja}} \tag{4}$$

$$S^l{}_{a\mu} \leftarrow L_{a\mu} \frac{\left(L^T S^h\right)_{a\mu}}{\left(L^T L S^l\right)_{a\mu}} \tag{5}$$

The update in Equation (4) is for ensuring uniqueness of the NMF submatrices. The convergence criterion for the algorithm is the Euclidean distance $\left\| S^h - LS^l \right\|$.

### 2.4 Determining weights using Simulated Annealing

After the NMF dimensionality reduction is completed the SIFts training data is initially transformed into the reduced space. Initial guess values are assigned to the weights in the lower *r*-dimensional space which are back transformed into the higher *n*-dimensional space using the equation $\mathbf{w}^h = \mathbf{L} \cdot \mathbf{w}^l$. The w-SIFt score is then calculated which is used to evaluate the objective function in Equation 2. A new weights vector $\mathbf{w}^l + \Delta \mathbf{w}$ is determined and the objective function is reevaluated. The new weights vector is accepted if the new objective value is better, otherwise it is accepted with a probability $p = \exp(-(E_{w,new} - E_w) / T)$ where $T$ is a global parameter called the temperature which is gradually reduced to a very small value ~ 0 during the course of the algorithm.

Since the weights in the higher dimension are supposed to be nonnegative, the weights in the lower dimension are constrained to be nonnegative. The nonnegativity constraint of the NMF algorithm helps to retain the nonnegative values of the SIFt data in the lower dimension and conversely nonnegative weights in the lower dimensions ensures nonnegative weights in the higher dimension. Maintaining the constraint of nonnegative weights in the higher dimension would have been a challenge, if other dimensionality reduction techniques such as Principal Component Analysis were used because of the possible encoding of the data to negative values in the lower dimension. Fig. 4. summarizes the overall workflow involving NMF dimensionality reduction stage and the determination of weights stage using simulated annealing.

### 2.5 Dataset generation

A dataset of P38α inhibitors whose potency (IC50) values and two-dimensional chemical structure have been reported in literature is considered to begin with. However, in order to generate SIFts for these inhibitors we should identify the *accurate* three-dimensional structure of the ligands binding into the protein which is a huge challenge. A rigorous search to determine the most likely binding pose of the ligand by binding energy minimization is not practical because of the combinatorial complexity of the conformational and positional search space of the ligand and the protein. The six degrees of translational and rotational freedom of the ligand, along with the internal conformational degrees of freedom of both the ligand and the protein, makes the search space extremely large. Consider as an example a simple system comprising a ligand with four rotatable bonds and six rigid body alignment parameters, the search space can be estimated as follows (Taylor et al., 2002): The alignment parameters are used to place the ligand relative to the protein in a

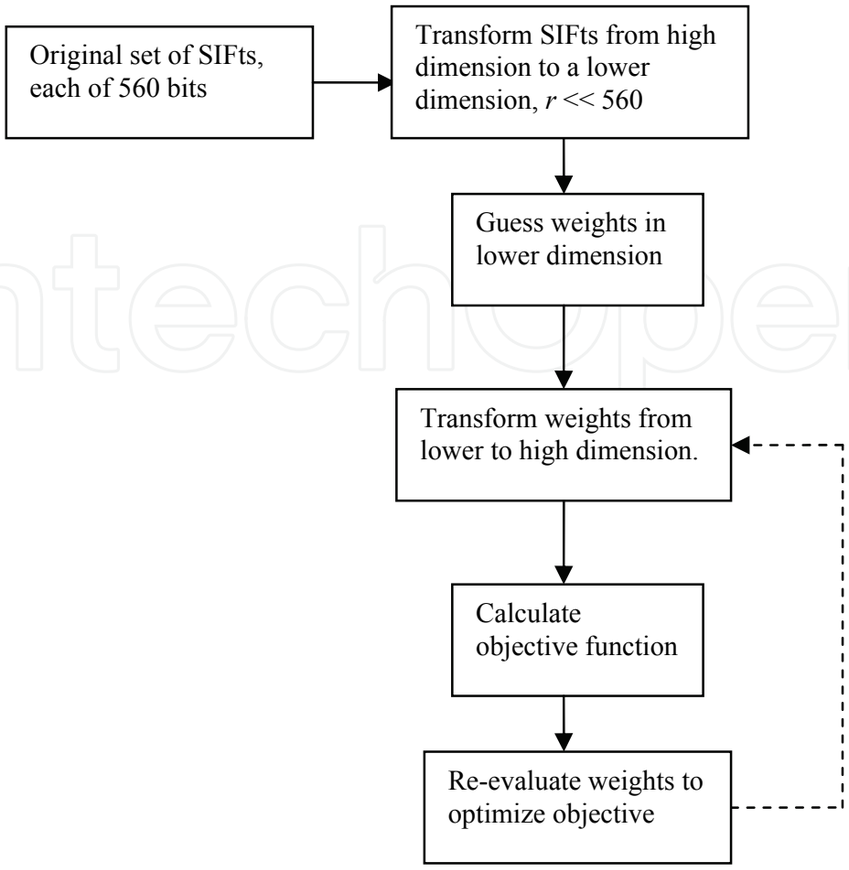



Fig. 4. The workflow involving dimensionality reduction and weights calculation.

cubic active site of size $10^3$ Å$^3$. If the angles are considered in 10 degree increments and translational parameters on a 0.5 Å grid there are approximately $4\times10^8$ rigid body degrees of freedom to sample, corresponding to $6\times10^{14}$ configurations (including the four rotatable torsions) to be searched. The search would take approximately 2,000,000 years of computational time at a rate of 10 configurations per second. So, the search process of docking algorithms implement some way of exploring only a partial region of the search domain thereby making the search implementation feasible. Molecular docking programs use heuristic search approaches based on molecular dynamics, monte carlo methods, genetic algorithms, fragment based methods, point complementarity methods, distance geometry methods, etc. However since these programs are heuristic the docked structure results are not reliable and so need to be cross-checked by other means such as comparing with experimentally determined binding poses of structurally similar ligands.

For the inhibitors above, all stable three-dimensional conformations are first generated using Omega program (Openeye Scientifc Software, NM) and these conformations are searched against known ligands of P38α using ROCS. The known ligands here are the ligands whose binding conformation with P38α has been confirmed experimentally and is available in the PDB. Only those inhibitors with a conformation closely matching with the known ligands are considered further for docking, whereas the remaining inhibitors are discarded because the resulting poses from docking cannot be verified for accuracy. Glide docking program (Schrodinger, NY) is used here to obtain the likely docking poses for the selected inhibitors.

After comparing the docking results with the binding pose of the corresponding known ligand using SIFts, only those ligands are finally retained whose binding pose matches closely with that of the known ligand and hence can be considered to be accurate. More information regarding the generation of the accurate binding poses from the two-dimensional inhibitor structures can be found in Nandigam et al. The final SIFt dataset considered here consisted of 89 protein-ligand structures of P38α. The active site of P38α consists of 56 residues with each residue being represented by 10 bits, making SIFt a 560 bit vector.

### 2.6 Cross-validating weights

The methodology described in the previous section involves a dimensionality reduction step that requires knowing *a priori* the dimensionality of the reduced space. Since we do not know the exact value of the reduced dimensionality in the case of SIFts, we build weighted profile models based on some guess values of the reduced space dimensionality using a training set and validate the models on a validation set. The guess value that generates a weights vector model that has the least validation error is chosen as the accurate dimensionality of the reduced space. This is because a model with the least validation error would theoretically also generate the least predictive error (Hastie et al., 2003).

Since the available SIFt dataset is small to split into separate training and validation sets, a five-fold cross validation method is used to generate training and test sets. The dataset of 89 SIFts is divided into a training set (both for training and cross validation) of 80 SIFts, and a test set (for final testing) of 9 SIFts. The 80 SIFts are further divided into 5 training-validation set pairs. In the cross validation procedure, a model is built for each of the five training sets and is validated against its corresponding validation set. The validation error of an individual model is the sum of squared differences between the model prediction values and the experimental $-Log(IC50)$ values for the validation set. The overall cross-validation error for a given dimensionality guess is taken as the average of validation errors of the five individual models constructed from the five training-validation set pairs.

The following steps outline the 5-fold cross-validation procedure.

1. Divide the overall dataset into five training and validation sets.
2. Consider a set of $r$ values.
3. For each $r$ , run the dimensionality reduction algorithm and then calculate five sets of weights corresponding to the five training sets.
4. Validate wSIFt scores of the five validation sets calculated based on the above weights by comparing against the experimental potency values.

## 3. Results

The cross validation errors calculated for various guess values of dimensionality for the reduced space are shown in Figure 3. The results show that a value of 20 for the reduced dimensionality corresponds to the least overall cross validation error, implying that the given P38α SIFt data can be efficiently translated as a combination of 20 linearly independent vectors. Figure 4 shows a heat map representation of the transformation $\mathbf{L}$ which is a graphical illustration of the 20 basis vectors in terms of the original 560 bits. Each of these basis vectors represents an 'interaction pattern' which is a combination of individual interactions that were found to co-occur in the original SIFt data. Each entry in

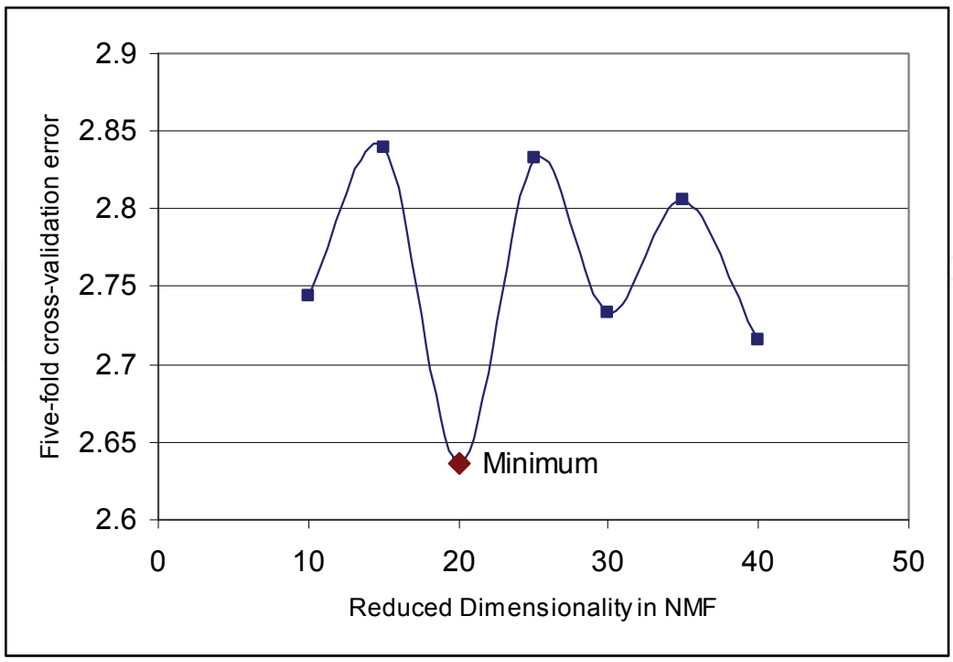



Fig. 3. The cross validation error of models built using different values of the lower dimensionality in NMF.

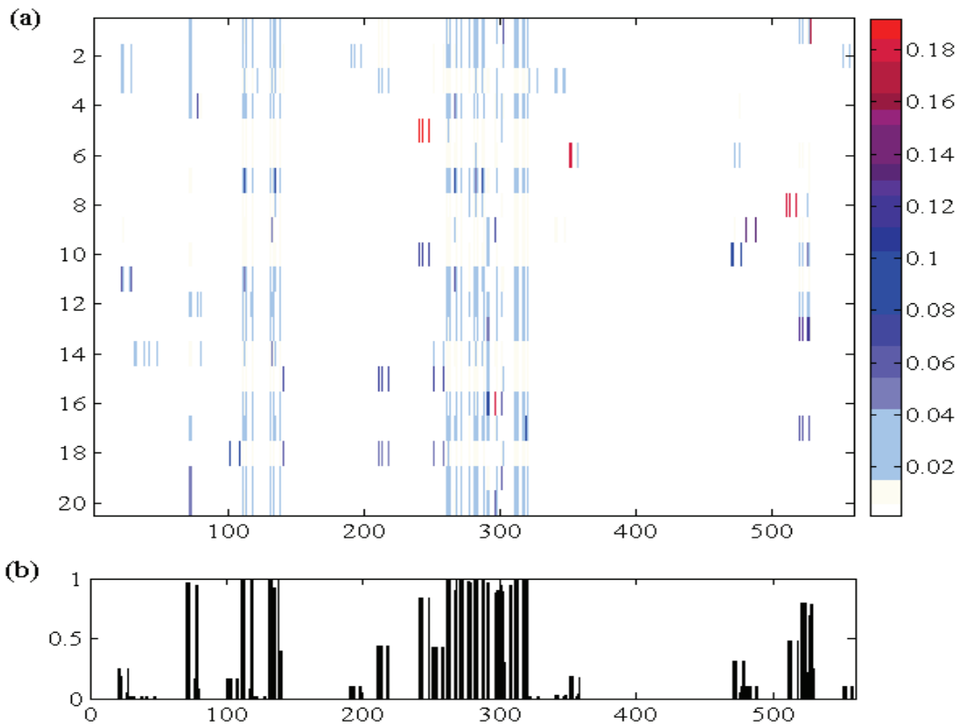



Fig. 4. (a) Heatmap of the transformation matrix ( **L** ) from 560 bit-space to a lower dimensional space (of size 20). The panel on the right shows the numerical value range for the colors in the heatmap. (b) The average of all the SIFts in the entire dataset.

the basis vector corresponds to the importance of that particular bit in that pattern of interactions. Thus the basis vectors represent a meaningful combination of interactions due to the nonnegative restriction on the elements of $\mathbf{L}$ matrix. Also, since the transformation matrix, $\mathbf{L}$, is nonnegative we simply need to restrict our weights in the lower dimensional space to be positive in order to satisfy the criterion that the weights in the original 560 bit space should be nonnegative.

The weight values of the weighted profile are provided as supplementary information in (Nandigam et al, 2009). The weights at the fingerprint positions corresponding to the contact bit of all the residues is shown in Figure 5. By looking at the weight values at the residue positions and the average SIFt values in Figure 5, it can be deduced that the weights are 'learnt' based on the supposed contribution of the interactions towards potency rather than mere frequency of interaction occurrence. In Figure 6(a), the w-SIFt scores of the training compounds, computed using the final weights model, are plotted against $-Log(IC50)$ values. The SIFt training data is categorized into three classes (colored blue, yellow and red in the figure) for better illustration and subsequent box plot analysis. The points in blue, yellow and red correspond to highly potent, moderately potent, and least potent compounds respectively. Figure 6(b) is the corresponding box plot representation showing the mean, quantiles, and outliers of the weighted profile scores (w-SIFt scores) for the three classes.

In Figure 6(c) w-SIFt scores of the 9 SIFts from final test set and the 80 SIFts from the training set are compared with the potencies. The w-SIFt scoring metric seems to perform well on the final test set too. The analysis done in Figure 6(a-b) is repeated for molecular weight and the docking score, in order to compare the performance of w-SIFt against other ligand parameters. Figure 7(a) shows the scatter plot of the molecular weight against the $-Log(IC50)$ values, whereas Figure 7(b) is its corresponding box plot. Figure 7(c) is the scatter plot of –docking score against the $-Log(IC50)$ values with its respective box plot shown in Figure 7(d). The figures show that the molecular weight ( $R = 0.2929$ **)** and docking score ( $R = 0.3415$ ) bear some correlation with the potencies though the deviation from the straight-line fit seems to be higher as evidenced in the respective box plots. The w-SIFt score definitely seems to a better metric for assessing the experimental potency of the ligand from the interaction fingerprint.

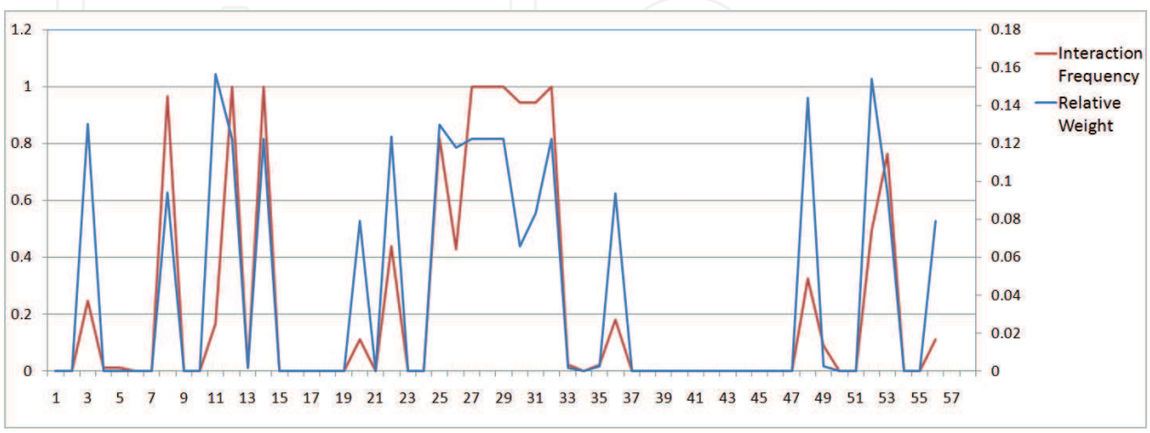



Fig. 5. The weighted profile showing the contact-bit weight at each of the residues as determined from the algorithm.

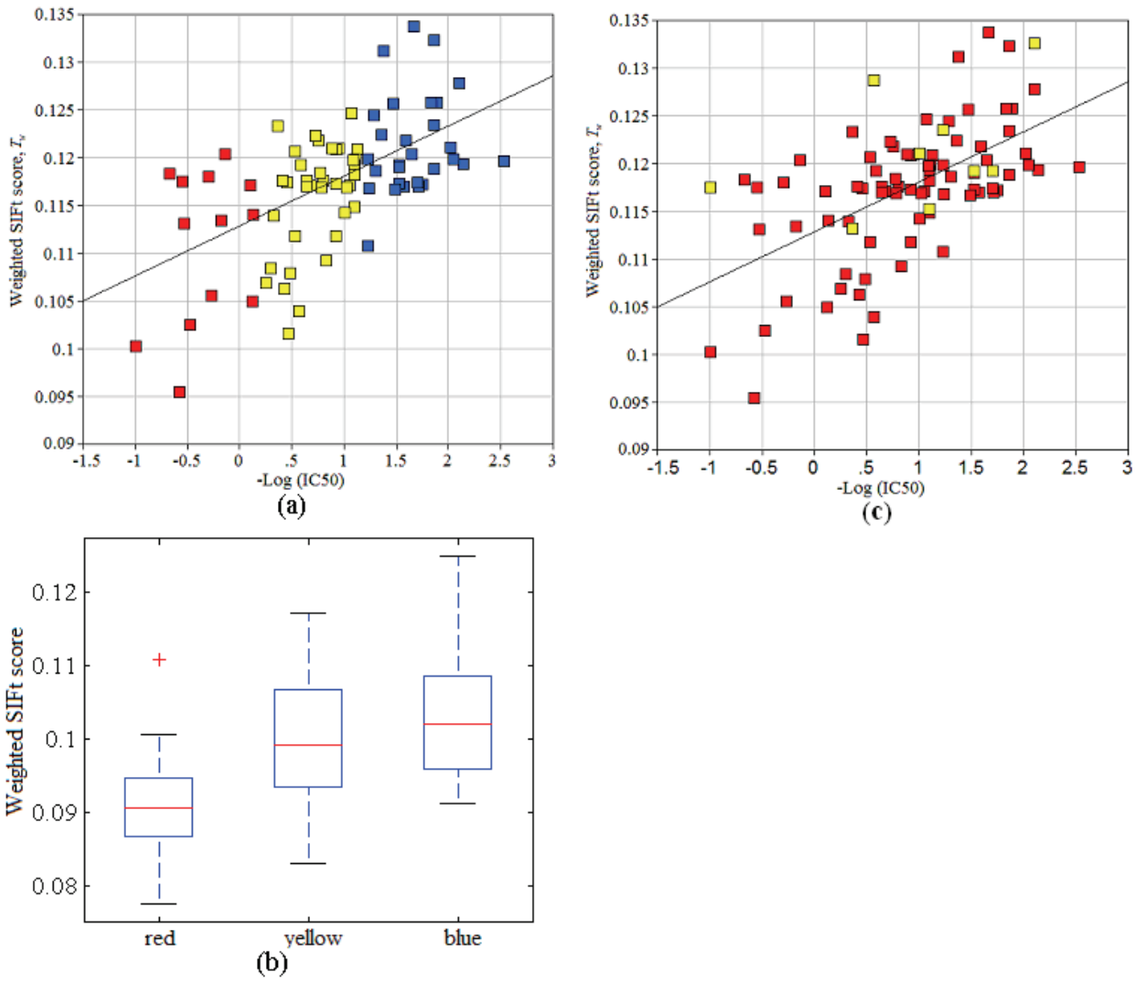



Fig. 6.(a) Scatter plot of the weighted SIFt scores against $-Log(IC50)$ for training data. The points in blue, yellow, and red correspond to the most potent, moderately potent, and least potent compounds. The correlation coefficient, $R = 0.6040$ . (b) Box plots of the distribution of the Weighted SIFt scores with respect to potency classes. (c) Scatter plot of the weighted profile scores against $-Log(IC50)$ for training (in red) and testing (in yellow) compounds.

## 4. Discussion

Typical physics based or empirical scoring functions are difficult to interpret: it is often not possible to extract information on what residues are driving potency and which interactions are more dispensable. The visual interpretation of the profile weights as illustrated in the previous section is perhaps the most powerful feature of the weighted interaction profiles described in this chapter.

The binding pocket of P38α with a ligand bound to it (PDB 1BL7) is shown in Figure 8(a), with the key binding residues highlighted with purple, cyan or white. It is observed that the weights illustrated in Figure 5 in fact reflect the relative importance of specific interactions in determining the potency of the P38α inhibitors considered in this study. In Figure 8(a),

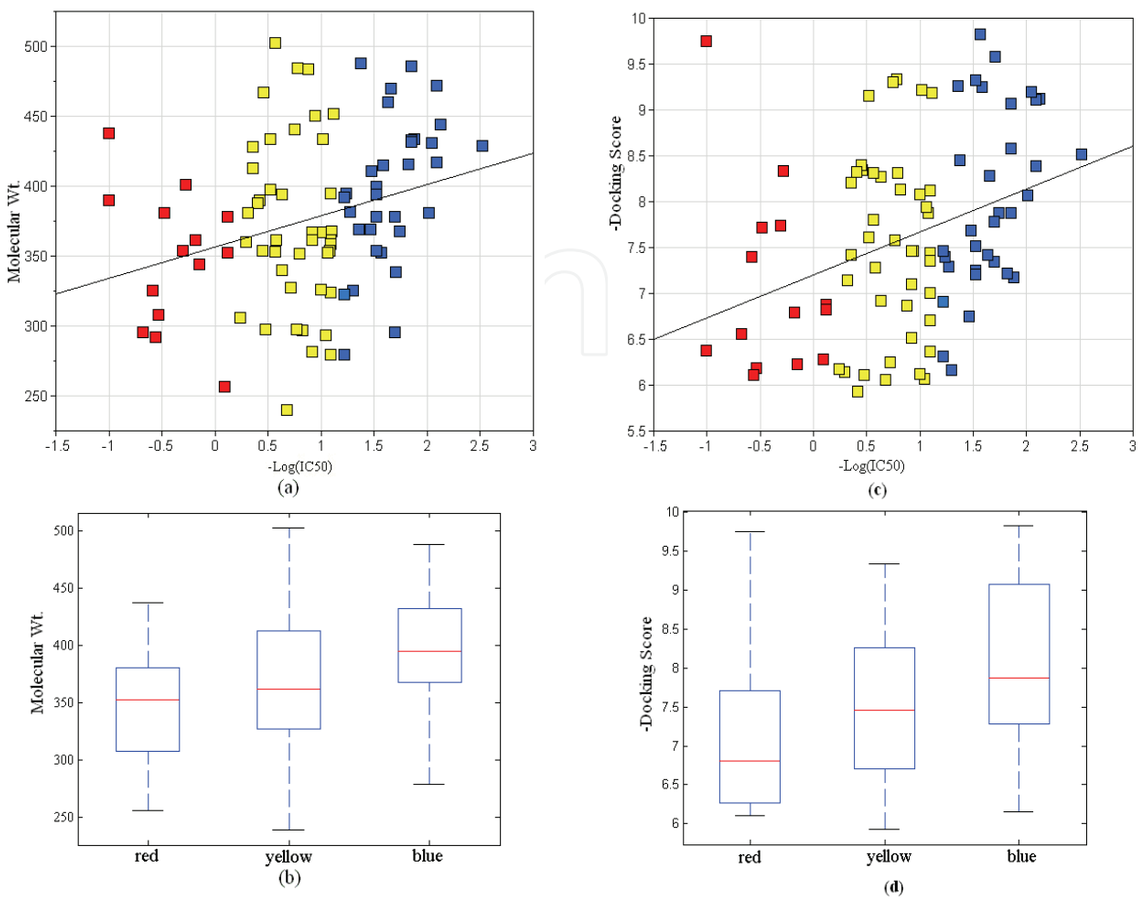



Fig. 7.(a) Scatter plot of the molecular weight against $-Log(IC50)$ . The points in blue, yellow, and red correspond to the most potent, moderately potent, and least potent compounds. The correlation coefficient, $R = 0.2929$ . (b) Box plots of the distribution of molecular weight with respect to potency classes. (c) Scatter plot of the –docking score against $-Log(IC50)$ . The points in blue, yellow, and red correspond to the most potent, moderately potent, and least potent compounds. The correlation coefficient, $R = 0.3415$ . (d) Box plots of the distribution of the -docking scores with respect to potency classes.

the most highly weighted residues are in purple; those with intermediate weight in cyan, and those least important for potency are colored white. The majority of ATP competitive kinase inhibitors interact with the hinge region of the kinase via at least one hydrogen bond (Chuaqui, Deng et al. 2005) mimicking the interactions made by the adenine moiety of ATP. In fact, these interactions are often used as constraints for filtering poses from docking experiments (Lyne, Kenny et al. 2004; Chuaqui, Deng et al. 2005). Not surprisingly, interactions with Met109, the key hydrogen-bonding residue in the hinge for P38α, are weighted heavily. In addition, Ala51 that makes hydrophobic contact with the typically heteroaromatic hinge binding substituents is identified as important for potency. Another nearly canonical interaction observed in the majority of kinase inhibitor co-crystal structures is with the conserved residue Lys53.

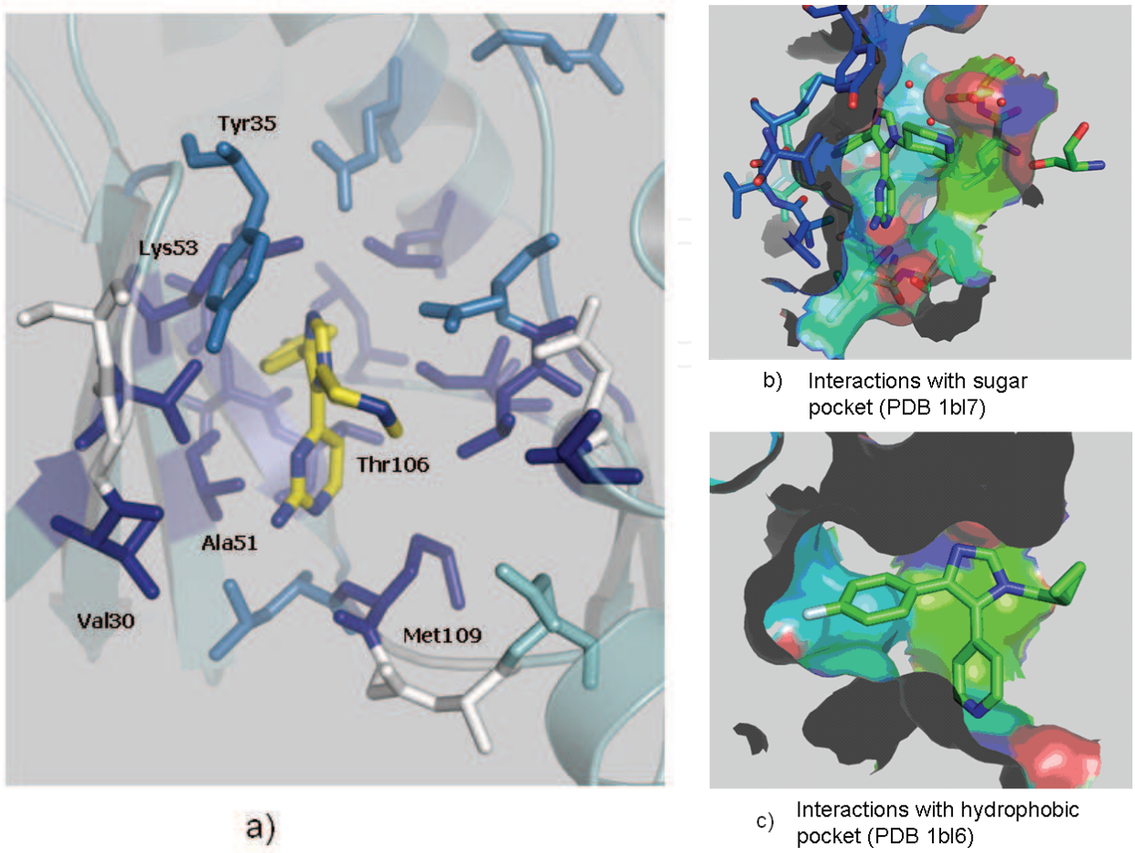



Fig. 8. (a) P38α with the key residues colored according to their weights. The residues in purple are the most highly weighted followed by residues in cyan, and the residues in white are the least weighted residues. Also shown in the figure are the labels for the residues referred to in the Discussion section. The sugar pocket (b) and hydrophobic pocket (c) are identified from the w-SIFt analysis as important regions for potency.

In addition to these highly conserved interactions, the hydrophobic pocket and sugar pocket regions of the ATP binding site received high weights. As is shown for example in Figure 8(b), inhibitors with substituents interacting with sugar pocket residues demonstrated increased potency over unsubstituted examples. Targeting the sugar pocket is a common strategy in kinase inhibitor design although it is not necessary to achieve potent activity in many kinases. The current analysis, however, indicates that this is an important region for p38α inhibition. In contrast, interaction with the P-loop of the kinase is not as important. The hydrophobic (or selectivity) pocket shown in Figure 8(c) was the final region that was identified in our analysis as being critical for potency. The small Thr106 gatekeeper residue in P38α permits access to the hydrophobic pocket unlike in kinases with bulky gatekeeper residues, e.g., CDK2 (Phe) or Akt (Met). Many P38α inhibitors exploit this region with substituted phenyl groups that contact a cluster of hydrophobic residues lining the pocket. The weights determined from our analysis highlight the importance of these interactions for achieving potency against P38α. Finally, interactions with the hinge toward the solvent channel of P38α were in comparison much less important for potency. As substitution toward solvent is typically aimed at improving inhibitor solubility, physical properties, and

selectivity (Fitzgerald, Patel et al. 2003), it is not surprising that the weights determined from potency alone are not high. However, inhibitors with solvent channel substituents that made hydrophobic contacts with Val30 did receive relatively high weights in our analysis.
In addition to being interpretable, we have demonstrated that with an optimized set of target-specific weights, weighted profiles are able to rank order compounds based on potency. The weighted SIFt scoring function could be used as a virtual screening tool for mining potent compounds from chemical databases. The first step of the virtual screening protocol would involve docking the inhibitors against the target protein and determining accurate poses based on a SIFt based filter as demonstrated by Deng et al. (Deng, Chuaqui et al. 2004). The weighted profile and the SIFts of the docked poses are now used to compute the w-SIFt score, which is used as a ranking criterion.

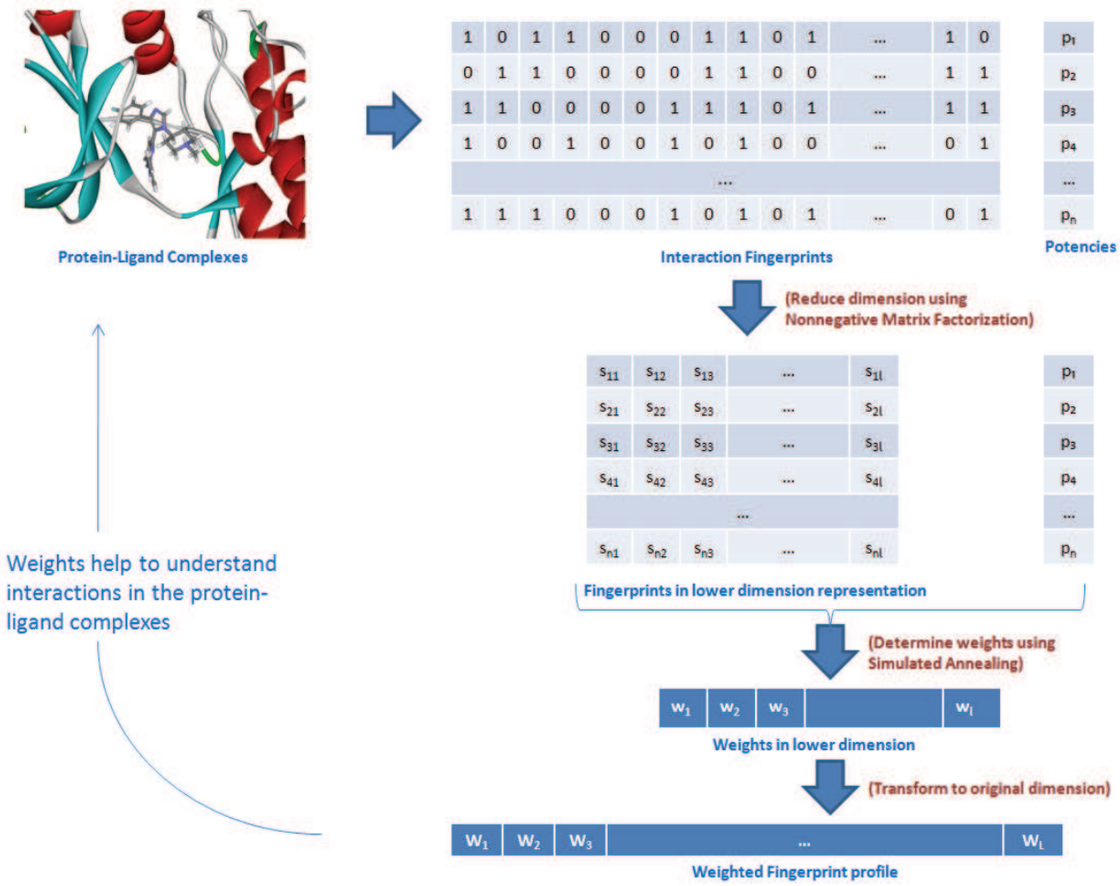



Fig. 9. Illustrative figure summarizing the full workflow involving determining SIFts from protein-ligand complexes, dimensionality reduction, weights determination, and interpretation of weights for better understanding of protein-ligand interactions.

Figure 9 shows a summary of the overall algorithm starting with the generation of SIFts from protein-ligand structures followed by the dimensionality reduction, and calculation of weights using simulated annealing, The weights so determined in turn help the understanding of the protein-ligand interactions which eventually will be useful for designing more efficient virtual screening algorithms to search for better binding ligands.

The concept of weighting the bits in SIFt can be extended to determine other criteria such as selectivity of a compound towards two targets. Rather than training the weights for learning experimental potency values, the weights now have to be trained for learning the relative potencies expressed as $\Delta(-Log(IC50))$ for example. The w-SIFt scoring function however suffers from the shortcoming that it is entirely based on assigning potency to protein-ligand binding interactions and does not include terms to delineate entropic contributions. There is however scope to combine the concept of weighting the interactions with other important ligand based terms like polar surface area, molecular weight, etc that also play a critical role in protein-ligand binding.

## 5. References

Glide. New York, Schrodinger Inc.

Omega. Santa Fe, NM, OpenEye Scientific Software.

Brewerton, S. C. (2008). "The use of protein-ligand interaction fingerprints in docking." Current Opinion in Drug Discovery & Development 11(11): 356-364.

Brunet, J. P., P. Tamayo, et al. (2004). "Metagenes and molecular pattern discovery using matrix factorization." Proc Natl Acad Sci 101: 4164-4169.

Chuaqui, C., Z. Deng, et al. (2005). "Interaction profiles of protein kinase-inhibitor complexes and their application to virtual screening." Journal of Medicinal Chemistry 48(1): 121-133.

Deng, Z., C. Chuaqui, et al. (2004). "Structural Interaction Fingerprint (SIFt): A novel method for analyzing three-dimensional protein-ligand binding interactions." Journal of Medicinal Chemistry 47(2): 337-344.

Devarajan, K. (2008). "Nonnegative Matrix Factorization: An Analytical and Interpretive Tool in Computational Biology." PLoS Comput Biol 4(7).

Fitzgerald, C. E., S. B. Patel, et al. (2003). "Structural basis for p38 alpha MAP kinase quinazolinone and pyridol-pyrimidine inhibitor specificity." Nature Structural Biology 10(9): 764-769.

Gao, Y. and G. Church (2005). "Improving molecular cancer class discovery through sparse non-negative matrix factorization." Bioinformatics 21: 3970-3975.

Hastie, T., R. Tibshirani, et al. (2003). The elements of statistical learning, Springer.

Kim, P. M. and B. Tidor (2003). "Subsystem identification through dimensionality reduction of large-scale gene expression data." Genome Res. 13: 1706-1718.

Kirkpatrick, S., C. D. Gelatt, et al. (1983). "Optimization by Simulated Annealing." Science 220(4598): 671-680.

Lee, D. D. and H. S. Seung (1999). "Learning the parts of objects by non-negative matrix factorization." Nature 401(6755): 788-791.

Lee, D. D. and H. S. Seung (2001). "Algorithms for Non-negative Matrix Factorization." Advanced in Neural Information Processing Systems 13: Proceedings of the 2000 Conference. MIT Press: 556-562.

Lyne, P. D. (2002). "Structure-based virtual screening: an overview." Drug Discovery Today 7(20): 1047-1055.

Lyne, P. D., P. W. Kenny, et al. (2004). "Identification of compounds with nanomolar binding affinity for checkpoint kinase-1 using knowledge-based virtual screening." Journal of Medicinal Chemistry 47(8): 1962-1968.

Nandigam, R. K., S. Kim, et al. (2009). "Position Specific Interaction Dependent Scoring Technique for Virtual Screening Based on Weighted Protein-Ligand Interaction Fingerprint Profiles." J. Chem. Inf. Model. 49(5): 1185-1192.

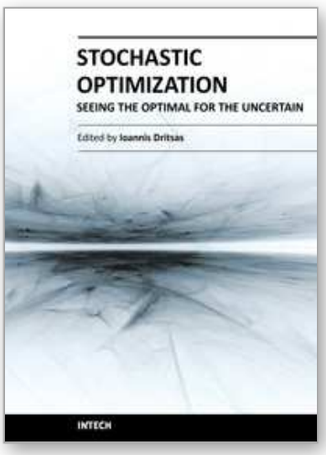
Pérez-Nueno VI, Rabal O, et al. (2009). "APIF: a new interaction fingerprint based on atom pairs and its application to virtual screening." J. Chem. Inf. Model. 49(5): 1245-1260.

Sato, T., T. Honma, et al. (2010). "Combining Machine Learning and Pharmacophore-Based Interaction Fingerprint for in Silico Screening." J. Chem. Inf. Model. 50(1): 170-185.

Singh, J., Z. Deng, et al. (2006). "Structural interaction fingerprints: A new approach to organizing, mining, analyzing, and designing protein-small molecule complexes." Chemical Biology & Drug Design 67(1): 5-12.

Taylor, R. D., P. J. Jewsbury, et al. (2002). "A review of protein-small molecule docking methods." Journal of Computer-Aided Molecular Design 16(3): 151-166.

Wallace, A. C., R. A. Laskowski, et al. (1995). "LIGPLOT - A program to generate schematic diagrams of protein-ligand interactions." Protein Engineering 8(2): 127-134.




**Stochastic Optimization - Seeing the Optimal for the Uncertain**
Edited by Dr. Ioannis Dritsas

ISBN 978-953-307-829-8
Hard cover, 476 pages
**Publisher** InTech
**Published online** 28, February, 2011
**Published in print edition** February, 2011

Stochastic Optimization Algorithms have become essential tools in solving a wide range of difficult and critical optimization problems. Such methods are able to find the optimum solution of a problem with uncertain elements or to algorithmically incorporate uncertainty to solve a deterministic problem. They even succeed in â€œfighting uncertainty with uncertaintyâ€. This book discusses theoretical aspects of many such algorithms and covers their application in various scientific fields.

**How to reference**
In order to correctly reference this scholarly work, feel free to copy and paste the following:

Ravi K. Nandigam and Sangtae Kim (2011). Understanding Protein-Ligand Interactions Using Simulated Annealing in Dimensionally Reduced Fingerprint Representation, Stochastic Optimization - Seeing the Optimal for the Uncertain, Dr. Ioannis Dritsas (Ed.), ISBN: 978-953-307-829-8, InTech, Available from: http://www.intechopen.com/books/stochastic-optimization-seeing-the-optimal-for-the-uncertain/understanding-protein-ligand-interactions-using-simulated-annealing-in-dimensionally-reduced-fingerp

INTECH
open science | open minds

**InTech Europe**
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

**InTech China**
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821

IntechOpen