Supporting Information for
‘One Size Does Not Fit All: The Limits of Structure-Based Models in Drug Discovery’

Gregory A. Ross\textsuperscript{1}, Garrett M. Morris\textsuperscript{2}, and Philip C. Biggin\textsuperscript{1}

\textsuperscript{1}Structural Bioinformatics and Computational Biochemistry, Department of Biochemistry, University of Oxford, South Parks Road, Oxford, OX1 3QU, United Kingdom
\textsuperscript{2}Crysalin Ltd., Cherwell Innovation Centre, 77 Heyford Park, Upper Heyford, OX2 55HD, United Kingdom

1 Summary

Structure-based models are an integral part of the drug discovery process. Although empirical models, such as scoring functions, have been developing over the last three decades,\textsuperscript{1} their performance still remains poor.\textsuperscript{2} By utilizing statistical learning theory\textsuperscript{3} and information theory,\textsuperscript{4} we show that even the most idealized structure-based model is significantly limited in its accuracy.

In Section 2, we provide additional background on using the cross entropy as a loss function by showing its relationship to the mean squared error. In Section 3, we prove the three main inequalities that underly fundamental sources of error for empirical scoring functions. The first (Section 3.1), is an upper bound to scoring function bias showing that arbitrarily large errors are possible if one samples the protein-ligand complexes in training and test sets in a different manner. The second and third inequalities show that while the scoring function which is optimal over a diverse range of protein-ligand complexes has the lowest bias of any other model (Section 3.2), its minimum error is on average larger than the minimum error of a protein-targeted scoring function (Section 3.3). The inability of a generalized scoring function to achieve the lowest error on protein-specific data sets has important implications for virtual screens. In Section 4 and 5 we detail the scoring function we used to test our theoretical predictions, list the complexes used in the single-protein data sets (Table S1), compare our diverse model to other popular models (Table S2) and show the minimum and maximum errors of the free parameter space surveyed in Figure 3 of the main text (Figure S1 and Table S3).

2 The cross entropy of normal distributions

2.1 Relationship to mean squared error

Our analysis utilizes cross entropy to quantify the error in scoring function regression. Similar information theoretic approaches to statistical learning can be found in references 5 and 6. The cross entropy $C(Y|X)$ for binding affinities $y$ given bound structures $x$ is given by

$$C(Y|X) = - \int \int p(x, y) \ln q(y|x) \, dx \, dy,$$

where $p(x, y)$ is the true PDF of the data set of binding affinities and structures and $q(y|x)$ is our model PDF of the conditional relationship between $x$ and $y$. The optimal scoring function for a data set sampled from $p(x, y)$ is encoded in the conditional distribution $p(y|x)$. For example, a well known result in statistical learning theory\textsuperscript{3} is that the scoring function that has the minimum mean squared error (MMSE), denoted $f_{\mu}(x)$, is given by the conditional average

$$f_{\mu}(x) = \int y \, p(y|x) \, dy. \quad (S1)$$

If the model $q(y|x)$ is a normal distribution with variance $\sigma_q^2$ and conditional average $f_q(x)$ so that
the cross entropy is equivalent to the mean squared error (MSE) of the scoring function $f_q(x)$ when applied to data sampled from $p(x, y)$. Inserting Equation S2 into $C(Y|X)$, we have

$$C(Y|X) = \frac{1}{2\sigma_q^2} \iint p(x, y)(y - f_q(x))^2 \, dx \, dy + \frac{1}{2} \ln 2\pi \sigma_q^2.$$  

The integral is the MSE of the scoring function $f_q(x)$. Note that the variance $\sigma_q^2$ does not depend on $f_q(x)$. Therefore, the scoring function that minimizes the cross entropy also minimizes the MSE for normally distributed errors. In a similar manner, one can also show that if $q(y|x)$ is chosen to be a Laplace distribution, the scoring function that minimizes the cross entropy also minimizes the mean absolute error.  

### 2.2 Relationship to scoring function regret

If one applies a scoring function that is optimal for a data set of structures and affinities sampled from $p(x, y)$ to a data set sampled from a different distribution $p_{\beta}(x, y)$, then the bias incurred by this model is given by

$$D(p_{\alpha}(y|x)||p_{\beta}(y|x)) = \iint p_{\alpha}(x, y) \ln \frac{p_{\alpha}(y|x)}{p_{\beta}(y|x)} \, dx \, dy.$$  

This bias represents a fundamental source of error for scoring functions which are trained and applied to data sets that have different sampling probabilities. In the main text (in the section “The optimization of scoring functions”), a scenario was considered where the conditional distributions $p_{\alpha}(y|x)$ and $p_{\beta}(y|x)$ were treated as normal distributions to help conceptualize this form of error. This section gives more background to that example and proves Equation 13 from the main text. The following analysis is treated more generally in reference 7. Let $p_{\alpha}(y|x)$ and $p_{\beta}(y|x)$ be normal distributions with variances $\sigma_{\alpha}^2$ and $\sigma_{\beta}^2$ and optimal scoring functions $f_{\alpha}(x)$ and $f_{\beta}(x)$ respectively. We investigate the case when the variances are similar in magnitude and we set $\sigma_{\beta} = \sigma_{\alpha} + \epsilon$ where $|\epsilon| \ll |\sigma_{\alpha}|$. The relative entropy becomes

$$D(p_{\alpha}(y|x)||p_{\beta}(y|x)) = \frac{1}{2} \ln \left( \frac{\sigma_{\alpha}^2 + 2\epsilon \sigma_{\alpha} + \epsilon^2}{\sigma_{\alpha}^2} \right) + \frac{1}{2(\sigma_{\alpha}^2 + 2\epsilon \sigma_{\alpha} + \epsilon^2)} \iint p_{\alpha}(x, y) (y - f_{\beta}(x))^2 \, dx \, dy$$

$$- \frac{1}{2\sigma_{\alpha}^2} \iint p_{\alpha}(x, y) (y - f_{\alpha}(x))^2 \, dx \, dy$$

$$\approx \frac{1}{2\sigma_{\alpha}^2} \iint p_{\alpha}(x, y) (y - f_{\beta}(x))^2 \, dx \, dy - \frac{1}{2\sigma_{\alpha}^2} \iint p_{\alpha}(x, y) (y - f_{\alpha}(x))^2 \, dx \, dy$$

$$= \frac{1}{2\sigma_{\alpha}^2} (\text{MSE}_{\beta, \alpha} - \text{MMSE}_{\alpha}),$$  

where the second line follows from the relative size of $\epsilon$ to $\sigma_{\alpha}$ and MMSE$_{\alpha}$ is the minimum mean squared error achievable on $p_{\alpha}(x, y)$ and MSE$_{\beta, \alpha}$ is the mean squared error of $f_{\beta}(x)$ applied to $p_{\alpha}(x, y)$. Since the relative entropy is always positive, Equation S4 means that the MSE of a modeled scoring function is always higher than that of the true scoring function. The difference between MSE$_{\beta, \alpha}$ and MMSE$_{\alpha}$ is the regret of model for normally distributed errors. As well as the regret, the relative entropy is also proportional to the mean squared distance between the true and estimated scoring functions. To show this, we start from Equation S4 so that we have
For the second inequality, we utilize the chain rule for relative entropy, \( \frac{1}{2\sigma^2} \int \int p_\alpha(x,y) \left[ (y-f_\beta(x))^2 - (y-f_\alpha(x))^2 \right] \, dx \, dy \)
\( = \frac{1}{2\sigma^2} \int p_\alpha(x) \left[ y^2 - 2yf_\beta(x) + f_\beta(x)^2 - y^2 + 2yf_\alpha(x) - f_\alpha(x)^2 \right] \, dx \)
\( = \frac{1}{2\sigma^2} \int p_\alpha(x) \left[ -2f_\beta(x) \int p(y|x) y \, dy + f_\beta(x)^2 + 2f_\alpha(x) \int p(y|x) y \, dy - f_\alpha(x)^2 \right] \, dx \)
\( = \frac{1}{2\sigma^2} \int p_\alpha(x) \left[ -2f_\beta(x)f_\alpha(x) + f_\beta(x)^2 + 2f_\alpha(x)^2 - f_\alpha(x)^2 \right] \, dx \)
\( = \frac{1}{2\sigma^2} \int p_\alpha(x) \left( f_\alpha(x) - f_\beta(x) \right)^2 \, dx, \) \tag{S5}

as stated in main text Equation 13. A graphical interpretation of this inspired the schematic diagram shown in Figure 1 of the main text, in which different data sets have their own optimal models that are embedded in a space that preserves the mean squared distance between them. While the above is symmetric with respect to the exchange of \( p_\alpha(y|x) \) and \( p_\beta(y|x) \), this is only true for special cases and is not true generally. \(^4\)

3 Proofs of main inequalities

3.1 Proof of Equation 9

Before proving Equation 9 from the main text, two inequalities must be stated. First, for non-negative numbers \( k_1, k_2, \ldots, k_n \) and \( l_1, l_2, \ldots, l_n \) the log-sum inequality is given by

\[
\left( \sum_{i=1}^{n} k_i \right) \ln \left( \frac{\sum_{i=1}^{n} k_i}{\sum_{i=1}^{n} l_i} \right) \leq \sum_{i=1}^{n} k_i \ln \frac{k_i}{l_i}, \quad \text{(S6)}
\]

with equality if and only if \( k_i/l_i \) is equal to a constant value. \(^4\)

For the second inequality, we utilize the chain rule for relative entropy, \(^4\) given by

\[
D(p_\alpha(x)||p_\beta(x)) + D(p_\alpha(y|x)||p_\beta(y|x)) = D(p_\alpha(x,y)||p_\beta(x,y)) \quad \text{(S7)}
\]

The non-negativity of \( D(p_\alpha(x)||p_\beta(x)) \) implies that

\[
D(p_\alpha(y|x)||p_\beta(y|x)) \leq D(p_\alpha(x,y)||p_\beta(x,y)). \quad \text{(S8)}
\]

Using the above inequalities,
\[
D(p_\omega(y|x)||p_\beta(y|x)) \leq \iint p_\alpha(x,y) \ln \frac{p_\alpha(x,y)}{p_\beta(x,y)} \, dx \, dy
\]
\[
= \iint \left( \sum_\theta p(x,y|\theta) \alpha(\theta) \right) \ln \frac{\sum_\theta p(x,y|\theta) \alpha(\theta)}{\sum_\theta p(x,y|\theta) \beta(\theta)} \, dx \, dy
\]
\[
\leq \iint \sum_\theta p(x,y|\theta) \alpha(\theta) \ln \frac{p(x,y|\theta) \alpha(\theta)}{p(x,y|\theta) \beta(\theta)} \, dx \, dy
\]
\[
= \iint \sum_\theta p(x,y|\theta) \alpha(\theta) \ln \frac{\alpha(\eta)}{\beta(\eta)} \, dx \, dy
\]
\[
= \sum_\theta \left( \iint p(x,y|\theta) \, dx \, dy \right) \alpha(\theta) \ln \frac{\alpha(\eta)}{\beta(\eta)}
\]
\[
= D(\alpha(\eta)||\beta(\eta)),
\]
where the last line follows from the discrete definition of relative entropy using natural logarithms.\(^4\)

### 3.2 Proof of Equation 11

We define \(p_\omega(x,y) = \sum_i^N \omega_i p_i(x,y)\); the average over \(N\) different complex selection probabilities. Following this definition, we also have \(p_\omega(y|x) = \sum_i^N \omega_i p_i(x,y) / \sum_i^N \omega_i p_i(x)\). To prove Equation 11, we utilize a standard information theoretic manipulation\(^4\) in the second line. Starting from the left hand side of Equation 11 we have:

\[
\sum_i^N \omega_i D(p_i(y|x)||p_\omega(y|x)) = \sum_i^N \omega_i \iint p_i(x,y) \ln \frac{p_i(y|x)}{q(y|x)} \, dx \, dy
\]
\[
= \sum_i^N \omega_i \iint p_i(x,y) \ln \frac{p_i(y|x)}{p_\omega(y|x)} \, dx \, dy
\]
\[
= \sum_i^N \omega_i \iint p_i(x,y) \left[ \ln \frac{p_i(y|x)}{p_\omega(y|x)} \right] \, dx \, dy
\]
\[
= \sum_i^N \omega_i D(p_i(y|x)||p_\omega(y|x)) + \sum_i^N \omega_i \iint p_i(x,y) \ln \frac{p_\omega(y|x)}{q(y|x)} \, dx \, dy
\]
\[
= \sum_i^N \omega_i D(p_i(y|x)||p_\omega(y|x)) + \sum_i^N \omega_i \iint \left[ \sum_i^N \omega_i p_i(x,y) \right] \ln \frac{p_\omega(y|x)}{q(y|x)} \, dx \, dy
\]
\[
= \sum_i^N \omega_i D(p_i(y|x)||p_\omega(y|x)) + \sum_i^N \omega_i D(p_\omega(y|x)||p_\omega(y|x)) + D(p_\omega(y|x)||q(y|x))
\]
\[
\geq \sum_i^N \omega_i D(p_i(y|x)||p_\omega(y|x)),
\]
where the last line follows from the fact that \(D(p_\omega(y|x)||q(y|x)) \geq 0\).

### 3.3 Proof of Equation 12

To prove Equation 12, first we show that the average cross entropy of the average density, \(p_\omega(y|x)\), on each protein specific density, denoted \(C_i(Y|X)\), is equal to the conditional entropy, \(h_\omega(Y|X)\), of the average density:
\[
\sum_{i}^{N} \omega_i C_i(Y|X) = - \sum_{i}^{N} \omega_i \int \int p_i(x, y) \ln p_\omega(y|x) \, dx \, dy \\
= - \int \int \left[ \sum_{i}^{N} \omega_i p_i(x, y) \right] \ln p_\omega(y|x) \, dx \, dy \\
= - \int \int p_\omega(x, y) \ln p_\omega(y|x) \, dx \, dy \\
= h_\omega(Y|X). \quad (S10)
\]

As in Equation 2 of the main text, we also have

\[
\sum_{i}^{N} \omega_i C_i(Y|X) = \sum_{i}^{N} \omega_i \left[ h_i(Y|X) + D(p_i(y|x)||p_\omega(y|x)) \right] \\
= \sum_{i}^{N} \omega_i h_i(Y|X) + \sum_{i}^{N} \omega_i D(p_i(y|x)||p_\omega(y|x)). \quad (S11)
\]

The inequality stated in Equation 12 arises by considering the second term in Equation S11: the average relative entropy between a protein-specific density and the generalized density. As the relative entropy is non-negative, the average can be zero if and only if \( D(p_i(y|x)||p_\omega(y|x)) = 0 \) for all \( i = 1, 2, ..., N \). However, each \( p_i(y|x) \) is distinct. This means that at most only one \( i \) out of \( N \) can have \( D(p_i(y|x)||p_\omega(y|x)) = 0 \). The rest must have relative entropies greater than zero, so

\[
\sum_{i}^{N} \omega_i D(p_i(y|x)||p_\omega(y|x)) > 0. \quad (S12)
\]

Combining Equations S10, S11 and S12 we find

\[
h_\omega(Y|X) = \sum_{i}^{N} \omega_i h_i(Y|X) + \sum_{i}^{N} \omega_i D(p_i(y|x)||p_\omega(y|x)) \\
h_\omega(Y|X) > \sum_{i}^{N} \omega_i h_i(Y|X),
\]
as stated in the main text.

4 Scoring function development

In this section, we outline the scoring function we used to test our analytical predictions. As is typical, we approximated binding free energy, denoted \( \Delta G \), as a linear combination of protein-ligand interaction terms so that

\[
\Delta G = a_1 U_M + a_2 U_E + a_3 U_B + a_4 U_L + a_5 U_P + a_6 U_T, \quad (S13)
\]

where \( a_1 \) to \( a_6 \) regression coefficients and the \( U_x \) are protein-ligand interaction potentials described below. Unless otherwise stated, the distance cutoff for atomic interactions was 6 Å. The mathematical notation used in this section is distinct from the notation used previously.

We modeled van der Waals interactions with the Morse potential

\[
U_M = \sum_{i,j} \left[ 1 - \exp \left( - \alpha (r_{ij} - R_{ij}) \right) \right]^2, \quad (S14)
\]
where $r_{ij}$ is the distance between the atoms $i$ and $j$, $R_{ij}$ is the sum of their van der Waal radii and $\alpha$ is the parameter that sets the width of the potential. The van der Waals radii of the atoms were taken from reference 10. While previous empirical models have used a number of Gaussian terms to measure shape complementarity, we have sought to simplify regression by using a single shape-based term. While designed for bonded interactions, the Morse potential was chosen purely for its functional shape, which is similar to the Lennard-Jones 6-12 potential. Accordingly, we set $\alpha = \frac{6}{R_{ij}}$ to reproduce the scale of this interaction. However, unlike the Lennard-Jones 6-12 potential, $U_M$ tends to zero at a faster rate for an increasing $r$, thereby rewarding larger atomic distance less, and has a softer hard core potential. For an atomic pair, the above potential has a minimum of $-1$, indicating a ‘good’ fit between the two atoms.

Using the partial charges computed by AutoDockTools (ADT), we calculated the electrostatic potential energy, $U_E$, between pairs of polar atoms:

$$U_E = \sum_{i,j} \frac{q_i q_j}{r_{ij}}$$  \hspace{1cm} (S15)

where $q_i$ and $q_j$ are the partial charges on the atoms $i$ and $j$. Formal charges were used for metal ions. We used the same hydrogen bond potential as in X-score and AutoDock Vina. For a particular pair of atoms, the hydrogen bond strength, $B_{ij}$, is given by

$$B_{ij} = \begin{cases} 1 & \text{if } (r_{ij} - R_{Xij}) \leq -0.7 \\ -10(r_{ij} - R_{Xij})/7 & \text{if } -0.7 < (r_{ij} - R_{Xij}) \leq 0 \\ 0 & \text{if } (r_{ij} - R_{Xij}) > 0, \end{cases}$$  \hspace{1cm} (S16)

where $R_{Xij}$ is the sum of the X-Score van der Waals radii of the atoms. The total hydrogen bond score, $U_B$, is then given by

$$U_B = \sum_{i,j} B_{ij}.$$  \hspace{1cm} (S17)

Following the same convention and parameterization as X-Score, metal interactions were also scored using $U_B$. As AutoDock 4’s scoring function, we used a torsional score $U_T$, given by

$$U_T = N,$$  \hspace{1cm} (S18)

where $N$ is the number of rotatable bonds in the ligand. The rotatable bonds were first identified using ADT and adapted using our own scripts so that bonds between aromatic rings and carboxylic acids or diaminomethanes (such as in benzamidine) were not counted as rotatable. Bonds between aromatic rings were treated as rotatable, which is the default in ADT.

Heuristic lipophilic and hydrophilic interaction terms, developed originally for the classification of water molecules in binding sites, were also used as descriptors for protein-ligand interactions. Their inclusion into the scoring function descriptor sets is due to their success in water molecule classification and was partly inspired by the Hint forcefield, which is uses calculated atomic logP data to account for a ligand’s chemical environment when bound to the protein. The lipophilic interaction term is simply a sum of carbon contacts weighted by their distance and is given by

$$U_L = \sum_{i,j} l_i \exp \left( -\frac{r_{ij}}{d_0} \right),$$  \hspace{1cm} (S19)

where $l$ is equal to one for carbon atoms and equal to zero for all other elements. The $d_0$ is a constant equal to 1 Å and is included to set the interaction range and ensure to non-dimensionality of the exponent. Similarly, The hydrophilic term is given by

$$U_P = \sum_{i,j} p_i \exp \left( -\frac{r_{ij}}{d_0} \right),$$  \hspace{1cm} (S20)
where $p$ is the hydration propensity of the atom. For protein atoms, these were taken from the study by Kuhn et al.\textsuperscript{15} For ligand atoms, the average values of the protein hydration propensities for each element were used. Halogens were given the same hydration propensity as the average carbon value and metal ions were given the same value as the highest protein atom propensity, due to their comparatively high solvation free energies.

Table S1: A table showing the PDB codes of the protein-ligand complexes used in the single-protein test sets. The proteins were chosen to be of pharmaceutical interests with a diverse range of functions. They were selected from the PDBbind data set and it was ensured they did not share any complexes with training set and diverse protein test set (data set A). Data set A consisted of a random selection of the CSAR data set. Additional complexes are shown in brackets for data sets B, C, E and G. These complexes were included to match the data sets in Cheng et al.’s study\textsuperscript{2} so that we could compare our diverse scoring function to others (comparison shown in Table S3). Unlike the rest of the single-protein data sets, the bracketed complexes are not present in the 2011 edition of refined PDBbind set, so for consistency, they were omitted from all other analyses.
5 Additional scoring function results

Table S2: Comparison of our scoring function to most and least accurate scoring functions (out of thirty three) tested in the study by Cheng et al.\textsuperscript{2} \textsuperscript{a} Spearman’s rank coefficient. \textsuperscript{b} Pearson’s correlation coefficient. \textsuperscript{c} Standard deviation in linear correlation (in \(\log K_d\) units). \textsuperscript{d} The fraction out of thirty three (in percent) that our scoring function has a higher Spearman’s rank coefficient than. The scoring functions with the highest \textsuperscript{e} and lowest \textsuperscript{f} ranking power in Cheng et al.’s study. These results show our scoring function is representative of a top performing function.

| Homo sapiens | Scoring function | \(\rho^a\) | \(R^b\) | SD\(^c\) | Better than\(^d\) |
|--------------|------------------|-------------|----------|-----------|------------------|
| HIV1 protease | Our model        | 0.29        | 0.29     | 1.6       | 79%              |
|              | X-Score::HPScore\(^e\) | 0.34        | 0.34     | 1.5       |                  |
|              | DS::LUDI1\(^f\)  | 0.01        | 0.01     | 1.6       |                  |
| Trypsin      | Our model        | 0.76        | 0.70     | 1.2       | 61%              |
|              | X-Score::HSScore\(^e\) | 0.82        | 0.82     | 1.0       |                  |
|              | GOLD::GoldScore\(^f\) | 0.05        | 0.15     | 1.8       |                  |

| Homo sapiens | Scoring function | \(\rho^a\) | \(R^b\) | SD\(^c\) | Better than\(^d\) |
|--------------|------------------|-------------|----------|-----------|------------------|
| Carbonic anhydrase II | Our model        | 0.41        | 0.65     | 1.0       | 39%              |
|                | DS::PLP2\(^e\)  | 0.77        | 0.80     | 0.8       |                  |
|                | GlideScore::XP\(^f\) | 0.10        | 0.05     | 1.4       |                  |
| Thrombin      | Our model        | 0.76        | 0.76     | 1.3       | 100%             |
|                | DS::PLP1\(^e\)  | 0.67        | 0.69     | 1.5       |                  |
|                | DS::PMF04\(^f\) | 0.02        | 0.12     | 1.9       |                  |

Table S3: The highest and lowest values for the mean absolute error (in kcal/mol) on the test sets on the parameter space displayed in Figure 3 from the main text. These values determine the color ranges used in each grid.

| Data set | Lowest MAE (kcal/mol) | Highest MAE (kcal/mol) |
|----------|------------------------|------------------------|
| A        | 1.6                    | 2.3                    |
| B        | 1.7                    | 3.5                    |
| C        | 1.2                    | 1.4                    |
| D        | 1.4                    | 3.3                    |
| E        | 3.7                    | 5.2                    |
| F        | 1.3                    | 2.3                    |
| G        | 1.4                    | 2.8                    |
| H        | 1.0                    | 4.0                    |
| I        | 1.2                    | 2.2                    |

Figure S1: Grid of heatmaps showing the mean absolute error for our own scoring function when applied to data sets A-I (see Table 1). Each heatmap shows the absolute error of the model on the data set as the two free parameters of the scoring functions are varied (x and y-axis). This figure supplements Figure 3 of the main text, which shows the relative errors on each data set.
References

[1] Kuntz, I. D.; Blaney, J. M.; Oatley, S. J.; Langridge, R.; Ferrin, T. E. A geometric approach to macromolecule-ligand interactions. *J. Mol. Biol.* 1982, 161, 269–288.

[2] Cheng, T.; Li, X.; Li, Y.; Liu, Z.; Wang, R. Comparative Assessment of Scoring Functions on a Diverse Test Set. *J. Chem. Inf. Mod.* 2009, 49, 1079–1093.

[3] Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning*; Springer Series in Statistics; Springer: New York, NY, USA, 2009.

[4] Cover, T. M.; Thomas, J. A. *Elements of Information Theory*; Wiley-Interscience: New York, NY, USA, 2006.

[5] Burnham, K.; Anderson, D. *Model Selection and Multimodel Inference*; Springer: New York, NY, USA, 2002.

[6] Principe, J. C. *Information Theoretic Learning*; Springer: New York, NY, USA, 2010.

[7] Verdu, S. Mismatched Estimation and Relative Entropy. *IEEE Trans. Inf. Theory.* 2010, 56, 3712–3720.

[8] Bondi, A. van der Waals Volumes and Radii. *J. Phys. Chem.* 1964, 68, 441–451.

[9] Trott, O.; Olson, A. J. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J. Comp. Chem.* 2010, 31, 455–461.

[10] Morris, G. M.; Huey, R.; Lindstrom, W.; Sanner, M. F.; Belew, R. K.; Goodsell, D. S.; Olson, A. J. AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. *J. Comp. Chem.* 2009, 30, 2785–2791.

[11] Wang, R.; Lai, L.; Wang, S. Further development and validation of empirical scoring functions for structure-based binding affinity prediction. *J. Comput.-Aided. Mol. Des.* 2002, 16, 11–26.

[12] Huey, R.; Morris, G. M.; Olson, A. J.; Goodsell, D. S. A semiempirical free energy force field with charge-based desolvation. *J. Comp. Chem.* 2007, 28, 1145–1152.

[13] Ross, G. A.; Morris, G. M.; Biggin, P. C. Rapid and Accurate Prediction and Scoring of Water Molecules in Protein Binding Sites. *PLOS ONE* 2012, 7, e32036.

[14] Eugene Kellogg, G.; Abraham, D. J. Hydrophobicity: is LogPo/w more than the sum of its parts? *Eur. J. Med. Chem.* 2000, 35, 651–661.

[15] Kuhn, L. A.; Swanson, C. A.; Pique, M. E.; Tainer, J. A.; Getzoff, E. D. Atomic and residue hydrophilicity in the context of folded protein structures. *Proteins* 1995, 23, 536–547.