JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

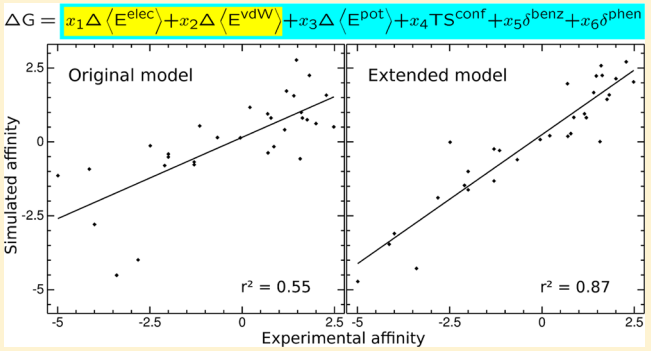
Hands-off Linear Interaction Energy Approach to Binding Mode and Affinity Estimation of Estrogens

Vedat Durmaz,*[,†] Sebastian Schmidt,[‡] Peggy Sabri,[†] Christian Piechotta,[‡] and Marcus Weber[†]

[†]Department of Numerical Analysis and Modelling, ZIB Zuse Institute Berlin, 14195 Berlin, Germany

[‡]Department of Analytical Chemistry, BAM Federal Institute for Materials Research and Testing, 12205 Berlin, Germany

ABSTRACT: With this work we target the development of a predictictive model for the identification of small molecules which bind to the estrogen receptor alpha and, thus, may act as endocrine disruptors. We propose a combined thermodynamic approach for the estimation of preferential binding modes along with corresponding free energy differences using a linear interaction energy (LIE) ansatz. The LIE model is extended by a Monte Carlo approach for the computation of conformational entropies as recently developed by our group. Incorporating the entropy contribution substantially increased the correlation with experimental affinity values. Both squared coefficients for the fitted data as well as the more meaningful leave-one-out cross-validation of predicted energies were elevated up to $r_{Fit}^2 = 0.87$ and $q_{LOO}^2 = 0.82$, respectively. All calculations have been performed on a set of 31 highly diverse ligands regarding their structural properties and affinities to the estrogen receptor alpha. Comparison of predicted ligand orientations with crystallographic data retrieved from the Protein database pdb.org revealed remarkable binding mode predictions.



INTRODUCTION

Up to now and inspite of the vast (parallel) computing power available nowadays, the prediction of binding modes and affinities for host−guest systems remains a time-consuming and highly nontrivial challenge. In order to achieve reliable affinities, several computational tasks need to be carried out of which the complexity increases drastically with the number of atoms.[1] Upon binding to a target, the ligand molecule must (as well as the target) adopt a proper conformation. In molecular modeling, this task is denoted as docking and usually followed by an estimation of the binding affinity or, respectively, the dissociation constant K_d which is, in terms of thermodynamics, considered as the ratio of the components' concentrations in two end states of a host−guest system at chemical equilibrium: the bound and the unbound state. The affinity K_d is associated with the difference ΔG of Gibbs free energy of binding which in turn is composed of enthalpic (ΔH) and entropic (ΔS) contributions

$$\Delta G = \Delta H - T\Delta S = RT\ln(K_d) \quad (1)$$

where Δ denotes the difference between the bound and the unbound state at chemical equilibrium $G = G^{bound} - G^{unbound}$ and T and R represent the temperature and the universal gas constant, respectively. Conventional docking tools for example provide both generation and thermodynamic quantification of binding modes. Two major strategies are known for the first step. The Autodock[2] software for example randomly suggests and optimizes translations and rotations as well as conformations of a ligand related to a target molecule using genetic algorithms. Other tools such as Dock[3,4] or FlexX[5] match pharmacophore features with complementary properties of the target's active site. In the next step, the complex conformations are valuated either by force field- or knowledge-based or empirical scoring functions.[6,7]

However, critical evaluations of results provided by docking tools attest them success in creating binding modes and, therefore, suitability for quick virtual screening of large databases in reasonable time, but neither in reliably identifying the natural one nor in correct affinity estimation.[8−10] In our opinion, a docking tool would do well to perform a systematic scanning of the space of orientations for the sake of overcoming rotational barriers, instead of relying on random poses or global minimization routines only especially, since during a subsequent ordinary MD simulation, essential rotations of guest molecules are very unlikely. In order to overcome this obstacle at a maximum level, we decided to gear to symmetric properties of the highest order platonic solid, namely the icosahedron resulting in 60 uniformly distributed binding modes. Applying this brute force approach followed by an energy minimization step at the active site, we will most likely get very close to the *true* binding mode.

Subsequent selection of the preferential binding mode on the basis of classical MD interaction energies appears quite evident

Received: July 5, 2013
Published: September 7, 2013

dx.doi.org/10.1021/ci400392p | *J. Chem. Inf. Model.* 2013, 53, 2681−2688

since we intend to apply the linear interaction energy[11] (LIE) method on short MD trajectories afterward for binding affinity prediction. Common methodology for this task ranges from extensive free energy perturbation[12] (FEP) and thermodynamic integration[13] (TI) approaches using a large set of MD simulations up to much faster but less accurate quantitative structure−activity relationship[14,15] (QSAR) strategies. Besides their high computational cost, TI and FEP cannot be applied to highly diverse compounds.[16,17] The LIE method was introduced by Åqvist as an alternative MD-based technique. It copes with thermodynamic principles and allows the computation of ΔG using averaged energy differences of the ligand's interaction (van-der-Waals E^{vdW} and electronic E^{elec}) with its surrounding from MD data of bound and unbound simulations:

$$\Delta G^{comp} = \alpha\Delta\langle E^{vdW}\rangle + \beta\Delta\langle E^{elec}\rangle \quad (2)$$

The coefficients α and β need to be determined empirically for each binding site using a trainig set of ligands with known binding affinities. Extension of the LIE model either by entropic or structural descriptors has been investigated.[10,18] All in all, results gained from LIE applications turned out to correlate well with experimental affinities, especially compared to scoring functions.[18−20] For all these reasons, we decided on the extendable LIE model for the binding affinity estimation of a diverse set of natural and synthetic ligands collected from various data sets related to the estrogen receptor alpha (ERα, see Figure 1) by way of example. In the face of high estrogenic activities of several synthetic compounds, the risk of endocrine disruption by xenoestrogens has been elucidated in the early eighties already.[21] Therefore, this target system is undergoing many investigations regarding the prediction of binding affinities. On the basis of several host−guest systems, van Lipzig and co-workers achieved excellent squared coefficients of correlation around 0.9 ± 0.04 for ERα using a LIE model refined with respect to the number of hydroxy groups of 19 ligands.[19] In advance, four ligand orientations had been chosen manually inspired by crystallographic data. High squared coefficients q_{LOO}^2 of the leave-one-out cross-validation up to 0.71 were achieved with pure 3D-QSAR methods investigating ERα and xenoestrogens with initial conformations selected by comparison with known binding modes.[14] Using a QSAR model along with partial least-squares regression on affinity prediction on ERα, Wang and co-workers achieved a correlation at $r_{Train}^2 = 0.92$ and $r_{Test}^2 = 0.84$ for the training and test set, respectively, but poor cross-validation with $q_{LOO}^2 = 0.43$, which seems somewhat surprising in comparison with the test set's correlation.[22] Many other models for the estimation of host−guest affinities do either yield poor cross-validation values or include the manual selection of an initial binding mode.[23,24]

In our opinion, an ideal predictive model should abstain from any preliminary information about the ligand's orientation and in particular, avoid a manual or random choice of some favorable pose. Using a suitable thermodynamic model of a certain active site, the majority of all compounds coming into consideration and representing a wide range of affinities and structural properties should be covered by the same parameter and training set. In the light of these technical issues, we are going to describe in the following an as simple as effective strategy for the estimation of binding modes and affinities by searching the preferential binding mode systematically as described above and applying an appropriate extension of the LIE model.

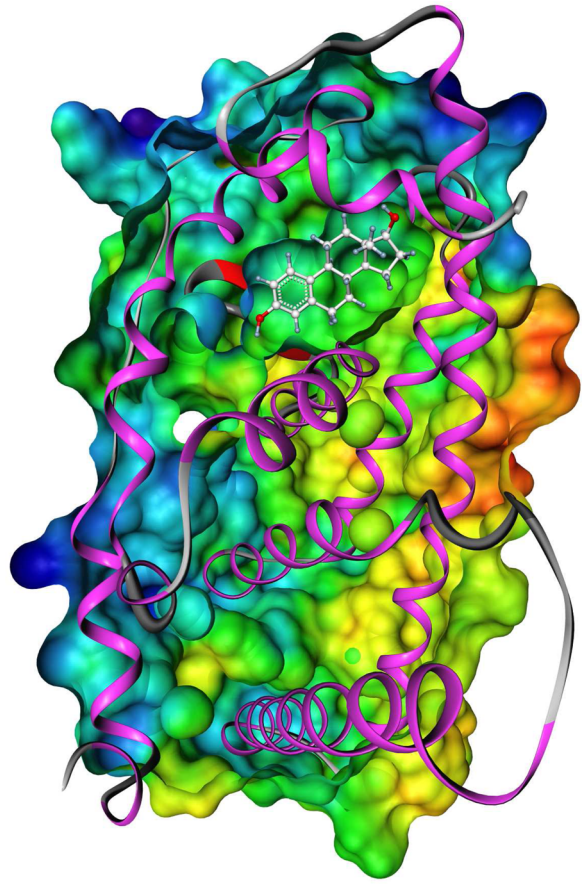


Figure 1. X-ray structure of the estrogen receptor α in complex with the natural binder 17β-estradiol (white carbon scaffold) represented by its secondary structure and clipped electrostatic solvent excluding surface.

METHODS

Data Set Preparation. The protein data bank[25] (PDB) entry 1GWR[26] in complex with the natural binder 17β-estradiol (E_2) was chosen as a target model. Missing loops had only been encountered at a large spatial distance to the binding site and were completed with the aid of PDB entry 3ERD after backbone alignment. From the PDB file of 1GWR, the second monomer of ERα (chain B), both ligands and all water molecules were eliminated before force field parametrization.

The data set of possible estrogens including diverse scaffolds and associated with affinities spread over 10^7 magnitudes was collected quasi-randomly from three different sources: twelve compounds from Kuiper's set,[27] ten compounds as published by Blair et al.,[28] and all eight molecules from unpublished binding assay studies performed by the Federal Institute for Materials Research and Testing in Germany as described in the experimental section below. On the basis of IC_{50} values from competitive binding assay analysis, the relative binding affinity (RBA) of any compound X can be expressed as $RBA(X) = 100 IC_{50}(E_2) IC_{50}(X)^{-1}$.

Experimental EC50 Determination. MCF-7 cells were obtained as third passage deposities in liquid nitrogen. After defrosting the first time, cells were cultivated and passaged three times: 25 aliquots of the sixth passage were frozen agiain until usage. Before usage, aliquots of the sixth passage were defrosted the second time and passages two times, thus experiments were conducted with the eighth passage. Thus, for

all main experiments, a certein consistence and homogeneity of cells are given and experiments are comparable. For passages 3−6 and 7−8, growing medium and experimental medium were used, respectively. They differ in their compositions: growing medium includes phenol red and bovine serum, which contains steroids. After, the solution was distributed in 200 μL aliquots on four 96 well plates and incubated under brood-conditions for 4 days. Pre-experiments were conducted for preliminary range finding: (i) concentration range for weak agonists was set from 5×10^{-4} to 5×10^{-15} M; and (ii) concentration range for strong agonists was set from 5×10^{-8} to 5×10^{-15} M. The positive control consisted of 17β-estradiol. The negative control consists of 10 nm fulvestrant, an antagonist inhibiting groth of MCF-7 cells. After preliminary range finding, stock solutions were diluted on deep well plates in order to measure the estimated EC_{50} value (from preliminary range finding) and at least 3 values between EC_{10} and EC_{90}. Additionaly, four values were prepared in equidistant manners above EC_{90} and below EC_{10} to get eleven different concentrations. The extinction was measured at 550 with a 630 nm reference signal. After a curve fitting using the Hill slope, RBA values were computed in analogy to the previous section using EC_{50} instead of IC_{50} values.

High-Temperature Hybrid Monte Carlo Simulation. Of each ligand, the global minimum was used as the initial structure for further simulations. Since MD simulations are known for trapping effects, the ligands underwent a high-temperature Hybrid Monte Carlo[29] simulation step in order to overcome energetic barriers.[30] After parametrization of all molecules according to the Merck molecular force field (mmff),[31] five Markov chains with 100.000 states each were generated applying the HMC method at an artificially high temperature of 1500 °C in order to efficiently sample the conformational space. Each HMC step included 30 MD steps with a 1.3 fs step size. Convergence of the simulation was checked according to Gelman and Rubin[32] on the basis of the five Markov chains. Afterward, all geometries were minimized with the conjugate gradient method,[33,34] and the lowest energy geometry of all was chosen as an estimate for a global minimum conformation. As described elsewhere,[35] this strategy seems to provide realistic global minima conformations for small molecules.

Multimode Hamiltonian Dynamics Simulation. For the estimation of affinities, the ERα target molecule was parametrized according to the amber99sb force field[36] and small molecules with the general Amber force field (GAFF)[37] using Antechamber from the AmberTools v1.4 package.[38] Charges were assigned with the am1bcc method[39,40] approximating restrained electrostatic potential (RESP) charges.[41] As described previously,[42] 60 initial orientations were generated of each ligand's global minimum conformation sharing the same geometric center which was superimposed with the geometric center of the crystallized E_2 molecule at the 1GWR binding site. Explicit water solvation was included using Amber's ffamber_tip3p model[43] which is part of the GAFF−Gromacs MD interface denoted as amber ports.[44] Using the Gromacs v.4.0.4 simulation package,[45−47] MD simulation was performed in three steps: Initially, the complex underwent 7000 steepest descent energy minimization steps if the maximum force had not ended up below 300 kJ mol^{-1} nm^{-1} before. During a subsequent 200 ps equilibration phase, all but solvent atoms and ions were restrained in their positions and the pressure was coupled weakly using Berendsen's algorithm.[48] Afterward, the whole system was simulated for 200 ps without position restraints but with constraints on all bonds according to the LINCS approach[49] allowing to set the discretized step size to 2 fs. In accordance with human physiology, the simulation temperature was coupled to 310 K by stochastically rescaling atomic velocities.[50] Interaction energies were computed on the basis of smooth particle mesh Ewald summation[51] for coulomb potentials with a cutoff at 10 Å and a van der Waals cutoff set to 14 Å.

RESULTS AND DISCUSSION

Optimal Binding Mode Identification. Instead of taking into account all $N = 60$ orientations per ligand i for affinity estimation using a Boltzmann-weighted sum E_{sum} of interaction energies $E^i(q_j)$ averaged over geometries q_j of n time frames

$$E_{sum} = -\frac{1}{\beta}\ln\left(\frac{1}{N}\sum_{i=1}^{N}\exp\left(-\frac{\beta}{n}\sum_{j=1}^{n}E^i(q_j)\right)\right) \quad (3)$$

we decided to omit all orientations but the one with the lowest mean interaction energy E_{opt}

$$E_{opt} = \min_{i\in[1,N]}\left(\frac{1}{n}\sum_{j=1}^{n}E^i(q_j)\right) \quad (4)$$

yielding exactly one preferential (most likely) binding pose. Here, β stands for the inverted product of the gas constant $R = 8.31447$ J mol^{-1} K^{-1} and temperature $T = 310$ K. Due to unrestrained equilibration purposes, the first 40 ps (25% of the trajectories) were skipped. Other publications[19,42] as well indicate a sufficient approximation of Boltzmann weights by focusing on the optimal orientation only. Since we want to predict exactly one binding mode anyway, eq 4 seems to be a convenient choice and in addition, slightly reduces the computational cost. Equating $E^i(q_j)$ with the Coulombic (electric) interaction energy only yielded by far the highest correlation as presented below. Table 1 and Figure 2 show six ligands investigated in this work, for which crystallographic data of the complexation with ERα was available in the PDB.

Table 1. Heavy Atom Root Mean Square Deviation of Computed Binding Modes from Respective Cocrystallized Estrogens Represented by PDB IDs

compound	PDB ID	rmsd
17-β-estradiol	1GWR	0.44
bisphenol A	3UU7	1.52
estriol	3Q95	0.51
estrone	3HM1	0.58
genistein	1X7R	1.22
(4-hydroxy-)tamoxifen	3ERT	2.04

Before rmsd computation, the protein backbones of each simulation's last time frame (green carbon scaffold in Figure 2) had been aligned to protein backbones of respective crystallized complexes (light blue scaffold). For all steroids as well as the phytoestrogen genistein, the native pose was successfully estimated as the preferential one. Due to the flexibility of ERα atoms during MD, slight deviations of these ligands at the binding site were to be expected, whereas bisphenol A and, in particular, tamoxifen show considerable deviations from the natural binding mode. Note that, in contrast to our simulations,

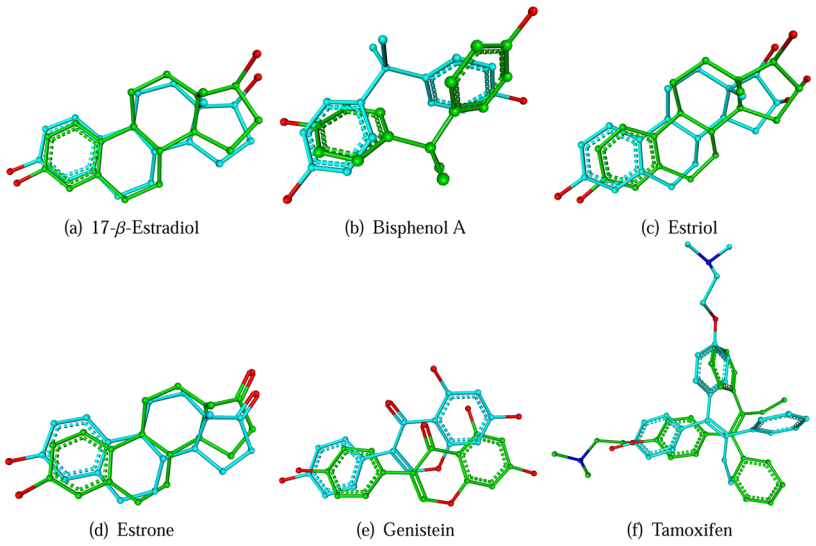


Figure 2. Alignment of predicted binding modes (green carbon scaffold) to respective cocrystallized (xeno)estrogens retrieved from the Protein Data Bank. Oxygens and nitrogens are colored red and dark blue, respectively.

the PDB entry 3ERT contains the 4-hydroxylated form of tamoxifen. For this reason, the comparison of prediction and crystal data is of limited suitability. However, chemical properties and groups such as benzene or hydroxy groups were matched well with the crystallographic positions. In the face of these results, it seems likely that the prediction of compounds with increasing number of rotational bonds will tend to result in wrong binding modes.

Monte Carlo Estimation of Conformational Entropies. Recently, Weber et al. presented a simple variance-based method for the qualitative approximation of conformational entropies from MD or Monte Carlo simulations,[52] also useful for free energy calculations.[53] Originally, it was designed to use internal (torsional) coordinates. Instead, we took external (kartesian) coordinates representing conformational, translational, and rotational entropies. In few words, the algorithm was implemented as follows: from a statistical ensemble of ligand i with $n^{(i)} = 50.000$ states, select $P = 10$ reference points (RP) with (nearly) mean-energy conformations. For each RP l, determine the fraction $n_l^{(i)}/n^{(i)}$ of ensemble states of which the root-mean-square deviation (rmsd) from the current RP is less than a predefined cutoff. In case of no variance in the coordinates, i. e. $n_l^{(i)} = n^{(i)}$, the entropy

$$S^i \approx -R\ln\left(\frac{1}{P}\sum_{l=1}^{P}\frac{n_l^{(i)}}{n^{(i)}}\right)$$

will equal zero; whereas, in case of maximal variance, when only the RP itself lies within the environment defined by the cutoff, the entropy becomes a maximal value depending on $n^{(i)}$. A careful choice of the cutoff is required in order to achieve useful entropies (larger than zero and less than the maximal value) for every RP of each ligand for comparison purposes. We had chosen the rmsd of the trajectory's multidimensional standard deviation of ligand coordinates, but a proper constant might have been suitable as well. For further details, refer to the original publications.[52,53]

Extended Linear Interaction Energy Model. Apart from mean van der Waals $\Delta\langle E^{\mathrm{vdW}}\rangle$ and Coulomb $\Delta\langle E^{\mathrm{elec}}\rangle$ interaction energy terms, our $m = 6$ parameter LIE model incorporated the $n = 31$ ligands' mean potential energy differences $\Delta\langle E^{\mathrm{pot}}\rangle$, the conformational entropies TS^{conf} of their unbound systems multiplied with temperature T as well as two QSAR-like structural descriptors δ^{benz} and $\delta^{\mathrm{phen}} \in \{0, 1\}$ indicating whether they contain (= 1) a benzene ring and a phenyl group, respectively, or not (= 0)

$$\Delta G^{\mathrm{comp}} = x_1\Delta\langle E^{\mathrm{elec}}\rangle + x_2\Delta\langle E^{\mathrm{vdW}}\rangle + x_3\Delta\langle E^{\mathrm{pot}}\rangle + x_4 TS^{\mathrm{conf}} + x_5\delta^{\mathrm{benz}} + x_6\delta^{\mathrm{phen}} \quad (5)$$

Here, Δ denotes the difference between the ligand bound and unbound to the target molecule. The interaction energies represented by the first two terms of eq 5 incorporate the ligand interaction with solvent molecules only in the unbound case and with target as well as solvent molecules in the bound state. A matrix $\mathbf{A} \in \mathbb{R}^{n\times m}$ was constructed coloumnwisely from the m descriptor vectors, after having had normalized each column by subtracting its mean and deviding it by its standard deviation. In addition, from the common logarithm of experimental RBAs, vector $y \in \mathbb{R}^n$ was built such that with the aid of the linear sytem of equations

$$\underbrace{\begin{pmatrix}\log(\mathrm{RBA}_1)\\ \vdots \\ \log(\mathrm{RBA}_n)\end{pmatrix}}_{y} \approx \underbrace{\begin{pmatrix}\Delta\langle \mathrm{E}_1^{\mathrm{elec}}\rangle & \Delta\langle \mathrm{E}_1^{\mathrm{vdW}}\rangle & \dots & \delta_1^{\mathrm{phen}}\\ \vdots & \vdots & \ddots & \vdots \\ \Delta\langle \mathrm{E}_n^{\mathrm{elec}}\rangle & \Delta\langle \mathrm{E}_n^{\mathrm{vdW}}\rangle & \dots & \delta_n^{\mathrm{phen}}\end{pmatrix}}_{A} \underbrace{\begin{pmatrix}x_1\\ \vdots \\ x_m\end{pmatrix}}_{x} \quad (6)$$

the parameter coefficients x could be determined by least-squares fitting, that is minimizing the squared deviation $\min_x\|y - Ax\|_2^2$ of simulated Ax from experimental data y via $x = (A^{\mathrm{T}}A)^{-1}A^{\mathrm{T}}y$. Table 2 shows mean values and standard deviations of the six model parameters along with coefficients x.

The terms log(RBA) from the fitting process, respective RBA values, and the deviations from experimental log(RBA) are listed in Table 3 together with experimental and predicted data from the leave-one-out cross-validation described in the next section. Applying the presented model to this data, the squared coefficient of correlation of experimental versus fitted log(RBA) yields $r_{\mathrm{Fit}}^2 = 0.87$ and a mean deviation and rmsd of 0.8 and 1.1 kcal mol^{-1}, respectively. It should be noted that the logarithm of RBA (shifted affinities) does not yield physically meaningful

dx.doi.org/10.1021/ci400392p | J. Chem. Inf. Model. 2013, 53, 2681−2688

Table 2. Mean and Standard Deviation (std) of Parameter Vectors (with Units of Kilojoules per Mole for the Four Thermodynamic Quantities) Used for Normalization and Their Coefficients x (Units of Moles per Kilojoule for Thermodynamic Quantities) from Least Squares Fitting of Simulated to Experimental Data

descriptor	mean	std	x
$\Delta\langle E^{elec}\rangle$	21.42	49.75	−1.55
$\Delta\langle E^{vdW}\rangle$	67.98	16.84	−2.02
$\Delta\langle E^{pot}\rangle$	12.20	18.70	−0.36
TS^{conf}	12.02	2.05	0.68
δ^{benz}	0.56	0.28	−0.47
δ^{phen}	0.40	0.35	1.60

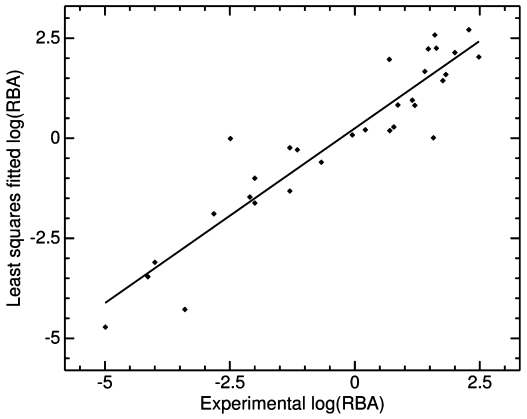


Figure 3. Correlation of least-sqaures fitted versus experimental logarithms of relative binding affinities of 31 chemical compounds to the estrogen receptor α using an empirical linear interaction energy-based approach.

free binding energies. The correlation coefficient itself is affected neither by shifting nor by the choice of the logarithmic basis. However, the (root) mean (square) deviation given in physical units does also not depend on RBA shifting, but on the logarithm's basis as well as on factors such as the gas constant R. In order to get useful deviations, we computed the natural logarithm of RBA values and multiplied them with RT = 0.002 kcal mol^{-1} K^{-1} × 310 K = 0.616 kcal mol^{-1}. The correlation is sketched in Figure 3. Considerably less correlation r_{Fit}^2 = 0.64 was achieved if having had neglected the two structural parameters, which seem to play an important role, especially, concerning the phenolic hydroxy group building several hydrogen bonds with ERα.

Table 3. Natural and Chemical Estrogens along with Relative Binding Affinities (RBA), log(RBA), and the Deviation of Fitted (Fit) and Predicted (LOO) from Experimental (Exp) log(RBA)

	relative binding affinity (RBA)			log(RBA)			deviation	
ligand	exp	fit	LOO	exp	fit	LOO	fit	LOO
17-β-estradiol (E_2)	100.0000	137.9022	142.1272	2.00	2.14	2.15	0.14	0.15
17-α-estradiol[a]	58.0000	27.6766	26.9553	1.76	1.44	1.43	−0.32	−0.33
17-α-ethinylestradiol[b]	190.0630	515.8976	594.7081	2.28	2.71	2.77	0.43	0.50
2-hydroxy-E_2[b]	29.5100	169.0191	286.3182	1.47	2.23	2.46	0.76	0.99
4-hydroxy-E_2[b]	66.0690	39.2186	36.0221	1.82	1.59	1.56	−0.23	−0.26
17-β-D-glucuronid-E_2[c]	0.0015	0.0128	0.0648	−2.82	−1.89	−1.19	0.93	1.64
3−17-disulfat-E_2[c]	0.0004	<0.0001	<0.0001	−3.40	−4.28	−4.74	−0.88	−1.35
3-β-D-glucuronid-E_2[c]	0.0079	0.0341	0.0552	−2.10	−1.47	−1.26	0.63	0.84
3-β-D-glucuronid-17-sulfat-E_2[c]	0.0001	0.0008	0.0027	−4.00	−3.10	−2.57	0.90	1.43
16-epiestriol[c]	4.9390	92.3863	217.7086	0.69	1.97	2.34	1.27	1.64
17-epiestriol[c]	39.8800	378.8939	566.2818	1.60	2.58	2.75	0.98	1.15
19-nortestosterone[a]	0.0100	0.1011	0.1665	−2.00	−1.00	−0.78	1.00	1.22
1-nonylphenol[b]	0.0032	0.9661	1.3895	−2.49	−0.01	0.14	2.48	2.64
5-androstenediol[a]	6.0000	1.9180	1.5759	0.78	0.28	0.20	−0.50	−0.58
5-α-dihydrotestosterone[a]	0.0500	0.0479	0.0474	−1.30	−1.32	−1.32	−0.02	−0.02
benzylbutylphthalate[c]	<0.0001	0.0004	0.0006	−4.14	−3.46	−3.24	0.68	0.90
β-zearalenol[a]	16.0000	6.5454	3.0178	1.20	0.82	0.48	−0.39	−0.72
bisphenol A[a]	0.0500	0.5820	0.7084	−1.30	−0.24	−0.15	1.07	1.15
clomifene[a]	25.0000	47.2091	60.9882	1.40	1.67	1.79	0.28	0.39
coumestrol[b]	0.8900	1.1900	1.2384	−0.05	0.08	0.09	0.13	0.14
di-*N*-butylphthalate[c]	<0.0001	<0.0001	<0.0001	−4.99	−4.72	−4.55	0.27	0.44
dienestrol[b]	37.1530	1.0191	0.8399	1.57	0.01	−0.08	−1.56	−1.65
estriol[a]	14.0000	8.9760	8.2113	1.15	0.95	0.91	−0.19	−0.23
estrone[b]	7.2400	6.7137	6.6900	0.86	0.83	0.83	−0.03	−0.03
genistein[a]	5.0000	1.5639	1.4727	0.70	0.19	0.17	−0.50	−0.53
hexestrol[b]	301.9976	107.4794	96.4767	2.48	2.03	1.98	−0.45	−0.50
methoxychlor[a]	0.0100	0.0241	0.0351	−2.00	−1.62	−1.45	0.38	0.55
moxestrol[a]	43.0000	178.3520	206.5474	1.63	2.25	2.32	0.62	0.68
norethindrone[a]	0.0700	0.5148	0.7479	−1.15	−0.29	−0.13	0.87	1.03
norethynodrel[b]	0.2137	0.2539	0.2612	−0.67	−0.60	−0.58	0.07	0.09
tamoxifen[b]	1.6218	1.6278	1.6313	0.21	0.21	0.21	0.00	0.00

[a]Kuiper et al., 1997. [b]Blair et al., 2000. [c]BAM, 2010−2011.

dx.doi.org/10.1021/ci400392p | J. Chem. Inf. Model. 2013, 53, 2681−2688

Cross-Validation an Predictive Power. Since the ordinary coefficient of correlation of fitted simulation data of some training set has no impact when we arrive at the estimation of experimental values, the predictive power of the empiric model represented by eq 5 was evaluated using the leave-one-out cross-validation. For each ligand i of the entire set, the coefficients of the LIE model were trained on the basis of the remaining compounds analogously to the previous section and subsequently applied to the prediction of ligand i yielding a set of predicted log(RBA). These are listed in Table 3 along with respective RBA values and deviations from experimental log(RBA). The squared coefficient of correlation between cross-validated and experimental log(RBA) was computed as $q_{LOO}^2 = 0.82$ which is remarkable in light of the fully automatic approach.

Although some scientists regard the use of q_{LOO}^2 with suspicion when no further test set was subjected to the predictive model,[54] meeting a couple of criteria substantially increases its reliability. The size of the training set has been more than 5-fold of the descriptor space. In addition, the set consisted of highly diverse chemicals with affinities ranging over nearly 10^7 magnitudes, allowing for the evaluation of strongly differing compounds. Apart from high coefficients q_{LOO}^2 of cross-validation, the respective regression line should hardly diverge from the bisecting line regarding its slope and intercept as depicted in Figure 4, since the plotting axes represent the same entity log(RBA).[54] The most flexible ligand 1-nonylphenol was predicted worst (far too affine) indicating difficulties when estimating the conformation and binding mode of highly flexible compounds, in particular in combination with target-specific descriptors such as the phenyl group. Neglecting this ligand merely would have increased the cross-validated correlation up to $q_{LOO}^2 = 0.85$. Applying the original two-parameter LIE model incorporating van der Waals and electric interaction energies yielded $q_{LOO}^2 = 0.50$ only which increased substantially to $q_{LOO}^2 = 0.67$ and further to $q_{LOO}^2 = 0.74$ by simply adding the phenyl indicator and conformational entropies of the unbound ligand, respectively. At the end, there will always exist an upper bound considerably less than 1 for the coefficient of correlation, as long as experimental measurements are statistically error-prone.

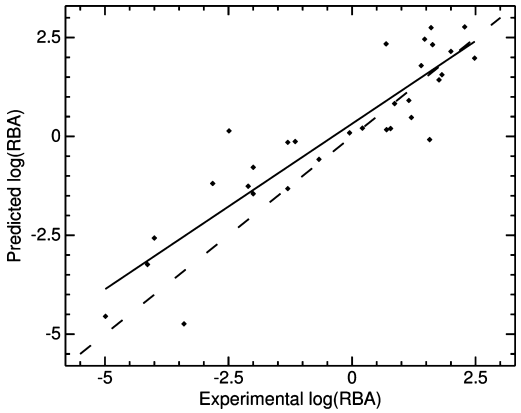


Figure 4. Correlation of predicted versus experimental logarithms of relative binding affinities of 31 chemical compounds to the estrogen receptor α using an empirical linear interaction energy-based approach.

CONCLUSION AND OUTLOOK

A promising hands-off approach to the prediction of binding modes and affinities for molecular host−guest systems has been presented. Without any need for manual operations, but given a binding site and an arbitrary set of coordinates only for a drug-sized compound, the presented strategy proposes one preferential binding mode agreeing very well with the X-ray structure of nearly all examples available at PDB. For ERα used exemplarily as target, a systematic scanning of the space of binding modes, which we strongly suggest for every molecular docking attempt, in combination with the automatic selection of one favorable pose due to coulomb interaction energies seems entirely sufficient. Using a set of 31 ligands with RBA from three different sources, an LIE model extended by a novel Monte Carlo-based method for conformational entropies, potential energy differences of the guest molecule and two structural descriptors as known from QSAR models yielded predicted RBAs correlating considerably with the experiment with $q_{LOO}^2 = 0.82$. In particular, the estimation strategy of conformational entropies as published recently and the phenyl indicator substantially increased the predictive power. Overfitting to the training set can be considered as minimal, since its size is 5-fold larger than the number of model parameters and due to highly divers molecular scaffolds associated with affinities ranging over 10^7 magnitudes.

AUTHOR INFORMATION

Corresponding Author

*E-mail: durmaz@zib.de.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

The BAM Federal Institute for Material Research and Testing is kindly acknowledged for its cooperation and financial support.

REFERENCES

(1) Griebel, M.; Knapek, S.; Zumbusch, G. *Numerical Simulation in Molecular Dynamics*; Springer: Heidelberg, Germany, 2007.

(2) Morris, G. M.; Goodsell, D. S.; Halliday, R. S.; Huey, R.; Hart, W. E.; Belew, R. K.; Olson, A. J. Automated Docking Using a Lamarckian Genetic Algorithm and an Empirical Binding Free Energy Function. *J. Comput. Chem.* **1998**, *19*, 1639−1662.

(3) 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.

(4) Moustakas, D. T.; Lang, P. T.; Pegg, S.; Pettersen, E.; Kuntz, I. D.; Brooijmans, N.; Rizzo, R. C. Development and validation of a modular, extensible docking program: DOCK 5. *J. Comput.-Aided Mol. Des.* **2006**, *20*, 601−619.

(5) Rarey, M.; Kramer, B.; Lengauer, T.; Klebe, G. A Fast Flexible Docking Method using an Incremental Construction Algorithm. *J. Mol. Biol.* **1996**, *261*, 470−489.

(6) Jain, A. N. Scoring functions for protein-ligand docking. *Curr. Protein Pept. Sci.* **2006**, *7*, 407−420.

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

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

(9) Michel, J.; Essex, J. W. Prediction of protein-ligand binding affinity by free energy simulations: assumptions, pitfalls and expectations. *J. Comput.-Aided Mol. Des.* **2010**, *24*, 639−658.
(10) Foloppe, N.; Hubbard, R. Towards predictive ligand design with free-energy based computational methods? *Curr. Med. Chem.* **2006**, *13*, 3583−608.
(11) Åqvist, J.; Medina, C.; Samuelsson, J. E. A new method for predicting binding affinity in computer-aided drug design. *Protein Eng.* **1994**, *7*, 385−391.
(12) Zwanzig, R. W. HighTemperature Equation of State by a Perturbation Method. I. Nonpolar Gases. *J. Chem. Phys.* **1954**, *22*, 1420−1426.
(13) Kirkwood, J. G. Statistical Mechanics of Fluid Mixtures. *J. Chem. Phys.* **1935**, *3*, 300−313.
(14) Shi, L. M.; Fang, H.; Tong, W.; Wu, J.; Perkins, R.; Blair, R. M.; Branham, W. S.; Dial, S. L.; Moland, C. L.; Sheehan, D. M. QSAR Models Using a Large Diverse Set of Estrogens. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 186−195.
(15) Tosco, P.; Balle, T. A 3D-QSAR-Driven Approach to Binding Mode and Affinity Prediction. *J. Chem. Inf. Model.* **2012**, *52*, 302−307.
(16) Kollman, P. Free Energy Calculations: Applications to Chemical and Biochemical Phenomena. *Chem. Rev.* **1993**, *93*, 2395−2417.
(17) Beveridge, D. L.; DiCapua, F. M. FREE ENERGY VIA MOLECULAR SIMULATION: Applications to Chemical and Biomolecular Systems. *Annu. Rev. Biophys. Biophys. Chem.* **1989**, *18*, 431−492.
(18) Nervall, M.; Hanspers, P.; Carlsson, J.; Boukharta, L.; Åqvist, J. Predicting Binding Modes from Free Energy Calculations. *J. Med. Chem.* **2008**, *51*, 2657−2667.
(19) van Lipzig, M. M. H.; ter Laak, A. M.; Jongejan, A.; Vermeulen, N. P. E.; Wamelink, M.; Geerke, D.; Meermann, J. H. N. Prediction of Ligand Binding Affinity and Orientation of Xenoestrogens to the Estrogen Receptor by Molecular Dynamics Simulations and the Linear Interaction Energy Method. *J. Med. Chem.* **2004**, *47*, 1018−1030.
(20) Stjernschantz, E.; Marelius, J.; Medina, C.; Jacobsson, M.; Vermeulen, N. P. E.; Oostenbrink, C. Are Automated Molecular Dynamics Simulations and Binding Free Energy Calculations Realistic Tools in Lead Optimization? An Evaluation of the Linear Interaction Energy (LIE) Method. *J. Chem. Inf. Model.* **2006**, *46*, 1972−1983.
(21) McLachlan, J. A.; Korach, K. S.; Newbold, R. R.; Degen, G. H. Diethylstilbestrol and other estrogens in the environment. *Fundam. Appl. Toxicol.* **1984**, *4*, 686−691.
(22) Wang, Z.; Li, Y.; Ai, C.; Wang, Y. In Silico Prediction of Estrogen Receptor Subtype Binding Affinity and Selectivity Using Statistical Methods and Molecular Docking with 2-Arylnaphthalenes and 2-Arylquinolines. *Int. J. Mol. Sci.* **2010**, *11*, 3434−3458.
(23) Khandelwal, A.; Balaz, S. Improved estimation of ligand-macromolecule binding affinities by Linear Response approach using a combination of multi-mode MD simulation and QM/MM Methods. *J. Comput.-Aided Mol. Des.* **2007**, *21*, 131−137.
(24) Vasanthanathan, P.; Olsen, L.; Jørgensen, F. S.; Vermeulen, N. P. E.; Oostenbrink, C. Computational Prediction of Binding Affinity for CYP1A2-Ligand Complexes Using Empirical Free Energy Calculations. *Drug Metab. Dispos.* **2010**, *38*, 1347−1354.
(25) Bergman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucelic Acids Res.* **2000**, *28*, 235−242.
(26) Warnmark, A.; Treuter, E.; Gustafsson, J.-A.; Hubbard, R. E.; Brzozowski, A. M.; Pike, A. C. W. Interaction of transcriptional intermediary factor 2 nuclear receptor box peptides with the coactivator binding site of estrogen receptor alpha. *J. Biol. Chem.* **2002**, *277*, 21862−21868.
(27) Kuiper, G. G. J. M.; Carlsson, B.; Grandien, K.; Enmark, E.; ggblad, J. H.; Nilsson, S.; Gustafsson, J.-Å. Comparison of the Ligand Binding Specificity and Transcript Tissue Distribution of Estrogen Receptors α and β. *Endocrinology* **1997**, *138*, 863−870.
(28) Blair, R. M.; Fang, H.; Branham, W. S.; Hass, B. S.; Dial, S. L.; Mol, C. L.; Tong, W.; Shi, L.; Perkins, R.; Sheehan, D. M. The Estrogen Receptor Relative Binding Affinities of 188 Natural and Xenochemicals: Structural Diversity of Ligands. *Toxicol. Sci.* **2000**, *54*, 138−153.
(29) Duane, S.; Kennedy, A. D.; Pendleton, B. J.; Roweth, D. Hybrid Monte Carlo. *Phys. Lett. B* **1987**, *195*, 216−222.
(30) Brass, A.; Pendleton, B. J.; Chen, Y.; Robson, B. Hybrid Monte Carlo Simulations Theory and Initial Comparison with Molecular Dynamics. *Biopolymers* **1993**, *33*, 1307−1315.
(31) Halgren, T. A. Merck Molecular Force Field: I-V. *J. Comput. Chem.* **1996**, *17*, 490−641.
(32) Gelman, A.; Rubin, D. Inference from Iterative Simulation using Multiple Sequences. *Statist. Sci.* **1992**, *7*, 457−511.
(33) Hestenes, M. R.; Stiefel, E. Methods of Conjugate Gradients for Solving Linear Systems. *J. Res. Nat. Bur. Stand.* **1952**, *49*, 409−436.
(34) Fletcher, R.; Reeves, C. M. Function minimization by conjugate gradients. *Comput. J.* **1964**, *7*, 149−154.
(35) Weber, M.; Becker, R.; Durmaz, V.; Köppen, R. Classical hybrid Monte-Carlo simulation of the interconversion of hexabromocyclododecane stereoisomers. *Mol. Simul.* **2008**, *34*, 727−736.
(36) Hornak, V.; Abel, R.; Okur, A.; Strockbine, B.; Roitberg, A.; Simmerling, C. Comparison of Multiple Amber Force Fields and Development of Improved Protein Backbone Parameters. *Proteins Struct. Funct. Bioinf.* **2006**, *65*, 712−725.
(37) Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. Development and Testing of a General Amber Force Field. *J. Comput. Chem.* **2004**, *25*, 1157−1174.
(38) Wang, J.; Wang, W.; Kollman, P. A.; Case, D. A. Automatic atom type and bond type perception in molecular mechanical calculations. *J. Mol. Graphics Modell.* **2006**, *25*, 247−260.
(39) Jakalian, A.; Bush, B. L.; Jack, D. B.; Bayly, C. I. Fast, Efficient Generation of High-Quality Atomic Charges. AM1-BCC Model: I. Method. *J. Comput. Chem.* **2000**, *21*, 132−146.
(40) Jakalian, A.; Jack, D. B.; Bayly, C. I. Fast, Efficient Generation of High-Quality Atomic Charges. AM1-BCC Model: II. Parameterization and Validation. *J. Comput. Chem.* **2002**, *23*, 1623−1641.
(41) Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Kollman, P. A. Application of RESP Charges To Calculate Conformational Energies, Hydrogen Bond Energies, and Free Energies of Solvation. *J. Am. Chem. Soc.* **1993**, *115*, 9620−9631.
(42) Durmaz, V.; Becker, R.; Weber, M. How to Simulate Affinities for Host-Guest Systems Lacking Binding Mode Information: Application in the Liquid Chromatographic Separation of Hexabromocyclododecane Stereoisomers. *J. Mol. Model.* **2012**, *18*, 2399−2408.
(43) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* **1983**, *79*, 926−935.
(44) Sorin, E. J.; Pande, V. S. Exploring the Helix-Coil Transition via All-Atom Equilibrium Ensemble Simulations. *Biophys. J.* **2005**, *88*, 2472−2493.
(45) Berendsen, H. J. C.; van der Spoel, D.; van Drunen, R. GROMACS: A message-passing parallel molecular dynamics implementation. *Comput. Phys. Commun.* **1995**, *91*, 43−56.
(46) van der Spoel, D.; Lindahl, E.; Hess, B.; Groenhof, G.; Mark, A. E.; Berendsen, H. J. C. GROMACS: Fast, Flexible, and Free. *J. Comput. Chem.* **2005**, *26*, 1701−1718.
(47) Hess, B.; Kutzner, C.; van der Spoel, D.; Lindahl, E. GROMACS 4: Algorithms for Highly Efficient, Load-Balanced, and Scalable Molecular Simulation. *J. Chem. Theory Comput.* **2008**, *4*, 435−447.
(48) Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; DiNola, A.; Haak, J. R. Molecular dynamics with coupling to an external bath. *J. Chem. Phys.* **1984**, *81*, 3684−3690.
(49) Hess, B.; Bekker, H.; Berendsen, H. J. C.; Fraaije, J. G. E. M. LINCS: A Linear Constraint Solver for Molecular Simulations. *J. Comput. Chem.* **1997**, *18*, 1463−1472.
(50) Bussi, G.; Donadio, D.; Parrinello, M. Canonical sampling through velocity rescaling. *J. Chem. Phys.* **2007**, *126*, 014101.
(51) Essmann, U.; Perera, L.; Berkowitz, M. L. A smooth particle mesh Ewald method. *J. Chem. Phys.* **1995**, *103*, 8577−8592.

(52) Weber, M.; Andrae, K. A simple method for the estimation of entropy differences. *MATCH Commun. Math. Comput. Chem.* **2010**, *63*, 319−332.
(53) Klimm, M.; Bujotzek, A.; Weber, M. Direct Reweighting Strategies in Conformation Dynamics. *MATCH Commun. Math. Comput. Chem.* **2011**, *65*, 333−346.
(54) Agatonovic-Kustrin, S.; Alexander, M.; Morton, D. W.; Turner, J. V. Beware of q^2! *J. Mol. Graphics Modell.* **2002**, *20*, 269−276.