Predicting the Sites and Energies of Non-Covalent Intermolecular Interactions Using Local Properties

Ahmed El Kerdawy. [a] *Christian R. Wick,*[a] *Matthias Hennemann* [a,b] *and Timothy Clark*[a,b,c,*]

[a] Computer-Chemie-Centrum, Friedrich-Alexander-Universität Erlangen-Nürnberg, Nägelsbachstraβe 25, 91052 Erlangen, Germany.

[b] Interdisciplinary Center for Molecular Materials, Friedrich-Alexander-Universität Erlangen-Nürnberg, Nägelsbachstraβe 49, 91052 Erlangen, Germany.

[c] Centre for Molecular Design, University of Portsmouth, Mercantile House, Portsmouth PO1 2EG, United Kingdom.

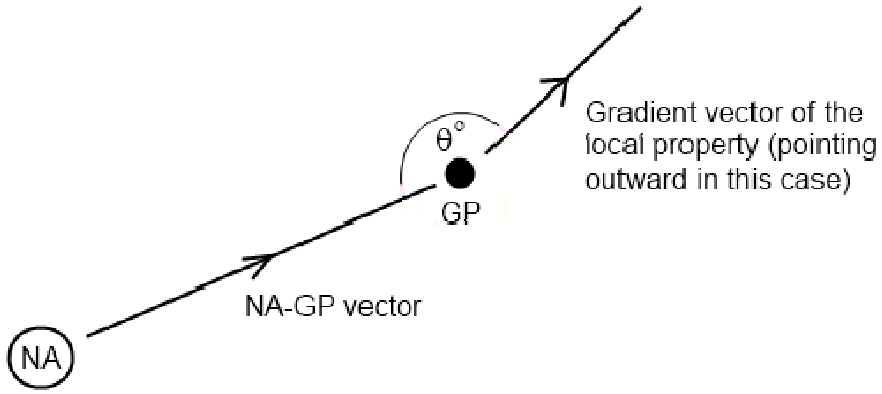
Supporting Information

Contents:

S1. Grid construction and descriptors calculation:

The output of VAMP was used as input for ParaSurf'10^{TM} to generate a grid around the molecule with a 4Å margin from the positions of the atoms in each direction and including all points for which the electron density is lower than 10^{-2} eÅ^{-3}. The spacing of the grid is 1Å (the default in ParaSurf^{TM}) and to calculate the local properties at each grid point using multipole electrostatics. All grid points inside the vdW volume of the central molecules were removed. At each grid point, the magnitude of the gradient vector for each local property and the cosine of the angle between it and the vector from the nearest atom in the central molecule to the grid point (cosθ) were calculated. Cosθ represents the direction of the gradient vector, as shown in Figure S1. We thus obtained the MEP, local ionization energy, local electron affinity, local polarizability, electron density and the magnitude and the direction (represented by cosθ) of the gradient for each of such local properties at each grid point. The distance between the grid point and nearest atom in the CM was also calculated and used as a descriptor.

Figure S1 Schematic representation for the gradient vector of the local property and its direction



NA nearest atom in the CM. **GP** Grid point.

S2. Interacting complex geometrical optimization and energy calculation:

Calculations of stabilization energies require higher levels of theory compared with geometry optimizations since the stabilization energy is more sensitive to an accurate description than the structure itself. To describe these non-covalent interactions accurately, high-level quantum mechanical calculations that reproduce a large portion of correlation energy are required. The most accurate quantum chemical method to obtain reliable geometries and energies of non-covalent complexes is the coupled-cluster expansion with single and double excitations augmented by non-iterative corrections for triple excitations CCSD(T) which is usually used as a replacement for the far more expensive CCSDT.[1-5] However, its computational cost is very high, which rules it out for the large number of calculations required in the current work, Møller-Plesset perturbation theory (MP2) gives geometries and energies not very different from those of CCSD(T) [1-4,6] and is much less demanding than CCSD(T) but remains more expensive than DFT calculations. Until recently, DFT was considered unsuitable for the study of the non-covalent complexes because it fails to describe the dispersion component of the non-covalent interactions because non-local correlation is not included in the local DFT energy. DFT therefore works well in cases of H-bonding and charge-transfer interactions but is not suitable for interactions in which dispersion energy is important.[1-3] However, many approaches are now available to overcome this drawback.[7,8] Of these approaches, the best known is an empirical approach based on a force-field-like terms, where the empirical expression for the dispersion energy is applied and the total energy E_{DFT-D} is constructed as a sum of DFT and empirical dispersion correction.[1,7,8] Here, we have used two of such DFT-D approaches that can reproduce the different non-covalent energies. These are a) a generalized gradient approximation GAA density functional with additional long-range dispersion correction **B97-D**,[7] which gives good geometrical and energetic descriptions of the non-covalent complexes because the short range part of the functional has been adjusted to the presence of the long-range correction and double-counting

effects are avoided. This functional gives a very balanced description of saturated vs. aromatic complexes and also a better description of hydrogen-bonded complexes than most of DFT approaches [7] and b) a long-range corrected hybrid density functional with damped atom-atom dispersion correction **ωB97X-D**, which is recommended for applications where non-covalent interactions are expected to be significant.[8]

The choice of basis set is critical for a reasonably reliable description of any structural type of non-covalent complexes. Because of the strong dependence of stabilization energy on the basis set size, it is important to perform the calculation with as large a basis set as possible. One small, reliable basis set is Dunning's aug-cc-pVDZ basis set, although aug-cc-pVTZ is even more reliable. [3] Because of the large number of calculations to be performed, calculations with both functionals used the more economical double-zeta **aug-cc-pVDZ** basis set.[9-12] Before selecting the level of theory for our calculations, we performed a pilot experiment in which we randomly selected 452 points to represent different interactions between different central molecules and probes and first performed constrained geometrical optimizations using the aug-cc-pVDZ basis set followed by single point energy calculations on the optimized structure with aug-cc-pVTZ. This approach is much more expensive than the one finally selected, but gives an RMSD between the two interaction energies of only 0.17 kcal mol^{-1}. The maximum absolute difference obtained using the B97-D functional was 1.06 kcal mol^{-1}, so we selected the less computationally expensive double-zeta basis set.

S3. Previous trials of non-covalent interaction models:

Many techniques for estimating non-covalent interaction strengths have been published in the last four decades, often for the most common and important type of interaction, the hydrogen bond. These techniques can be as simple as being based on simple indicator variables, such as the number of the H-bond donors and acceptors in the molecule.[13-15] The disadvantage of such methods is that they cannot differentiate between the different H-bond donor and acceptor strengths

because they neglect the effect of the remainder of the molecule on the H-bonding ability. Other methods are based on experimentally measured properties such as the difference between octanol-water and cyclohexane-water partition coefficients [16] or solvatochromic parameters (derived from spectroscopic data),[17-19] but such methods depend on experimentally derived properties. On the other hand various methods based on theoretically calculated properties such as LUMO and HOMO energies and atomic charges,[20,21] self-atom polarizability and superdelocalizability,[22, 23] molecular electrostatic potential MEP [24] or combination of these properties [25] and electrostatic potential $V_\alpha(r)$ at distance r [26] have been proposed. However, such approaches treat the H-bond as an isolated interaction and do not take the influence of the molecular environment into account. They are also often focused on limited group of compounds, so that they cannot be used as universal tools for predicting H-bond strengths. Many of such methods also only consider electrostatic parameters and neglect other non-covalent interaction steric and dispersion components. In recent years, because of the increased computational power and the increasing awareness that computational methods are capable of accurate geometry and energy calculations for non-covalent interactions,[27-31] attention has been given to using the quantum mechanical (QM) calculations to predict H-bond strength. In this context, Schwöbel et al.[32, 33] used *ab initio* and DFT based local molecular parameters for modeling and predicting the H-donor and H-acceptor strengths in some organic compounds. Another approach used the calculated interaction energy of the two interacting systems as a measure for the H-bond strength[34,35] based on the hypothesis that the hydrogen bond free energy is linearly related to the enthalpy.[19,36,37] This approach gives a better correlation between the calculated energies and the experimental hydrogen-binding constant than the other approaches.[34] It also takes all the different components of the non-covalent interaction including the steric and the dispersion components into consideration and considers the molecular environment, not just the isolated H-bond forming atoms. The major drawback of such an approach is the high demand in computational resources and time. This problem can be solved easily once a model for calculating the interaction energy is generated and then

the procedures can be extended to any other molecule. The key point is to choose descriptors for this model that can describe the different non-covalent interaction components to be able to predict the interaction energy as accurately as possible.

S4. Multiple linear regression (MLR) procedures:

The RapidMiner 5.0.008 software [38,39] was used with its default multiple linear regression (MLR) parameters, which use the Akaike criterion for model selection using the M5 prime method of feature selection during regression and delete colinear features during regression if exit.

We used a 10-fold cross-validation strategy for the MLR models. For each model, the data set was split into ten equal subsets. For each MLR, nine different subsets were selected as training sets and the tenth was used as the test set, yielding ten training sets of 90% of the points, and their corresponding test sets consisting of 10% of the points. Training and test sets were chosen such that each point appears in the test set for only one MLR. Then, ten separate MLRs were constructed. The cross-validated result for any given point is then the prediction by the MLR for which it appeared in the test set. This should therefore give a worst case prediction for the given point in this model.

S5. Removing points with secondary clashes:

This was done by calculating the distance between all the probe atoms, except the interacting atom on the grid, and all the CM atoms except the nearest at all points, if the distance is less than the sum of vdW radii of any two atoms, the point was removed.

S6. Plots of different models results (cross-validated):

Figure S2 H_2O as H-donor (B97-D functional) model

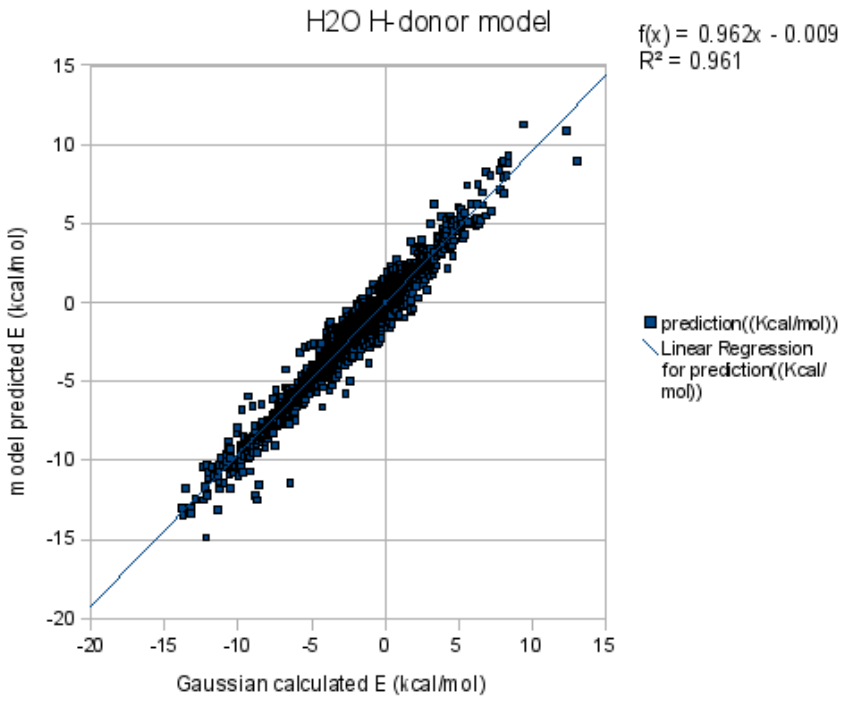


Figure S3 H_2O as H-acceptor (B97-D functional) model

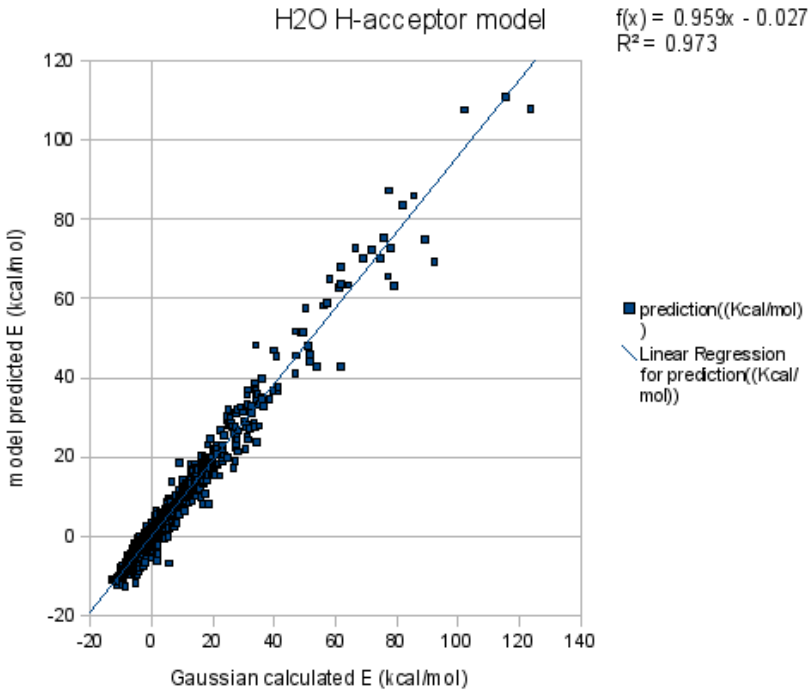


Figure S4 Formamide as H-donor (B97-D functional) model

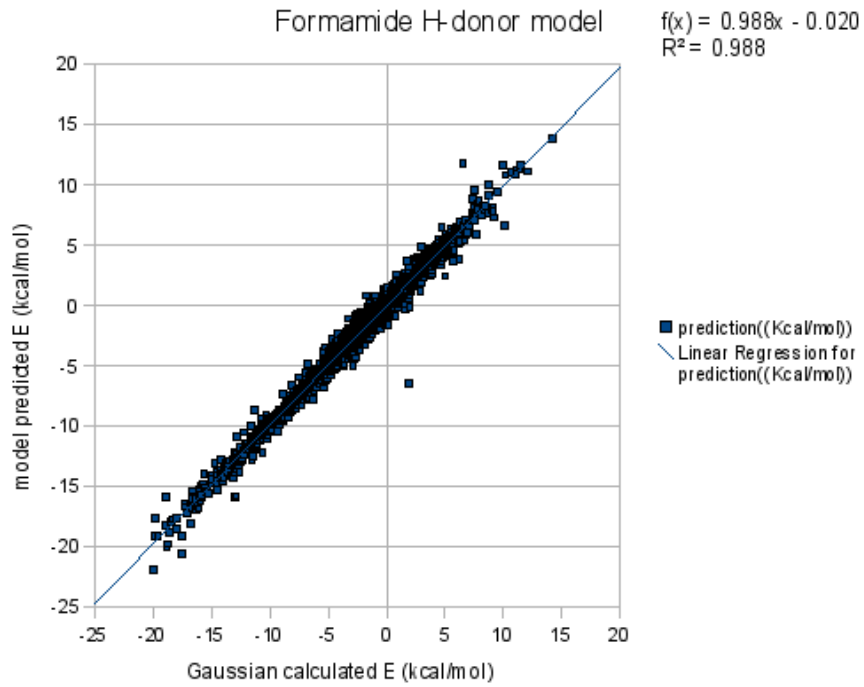


Figure S5 Formamide as H-acceptor (B97-D functional) model

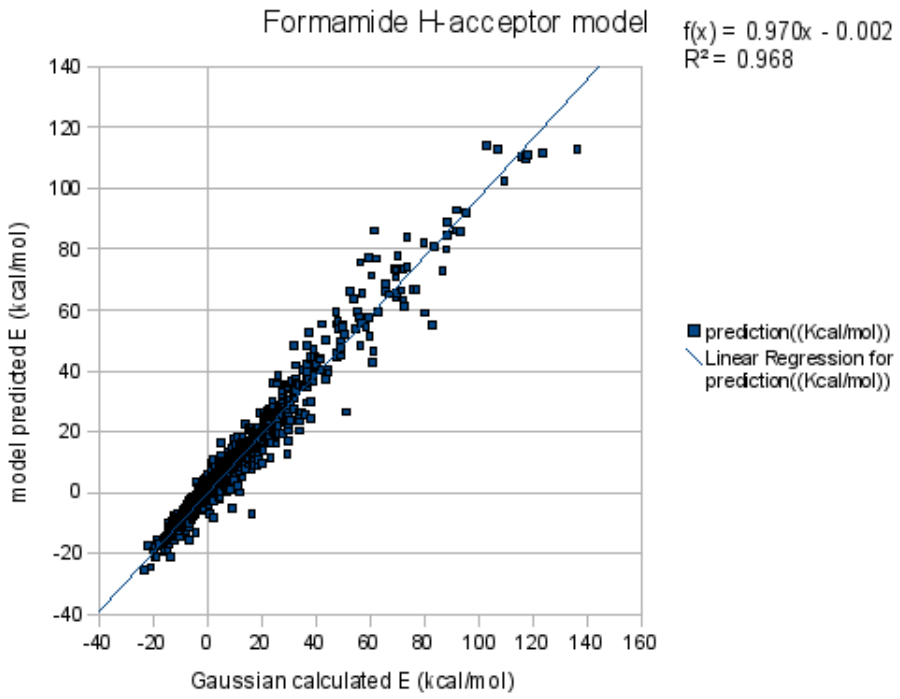


Figure S6 NH3 as H-donor (B97-D functional) model

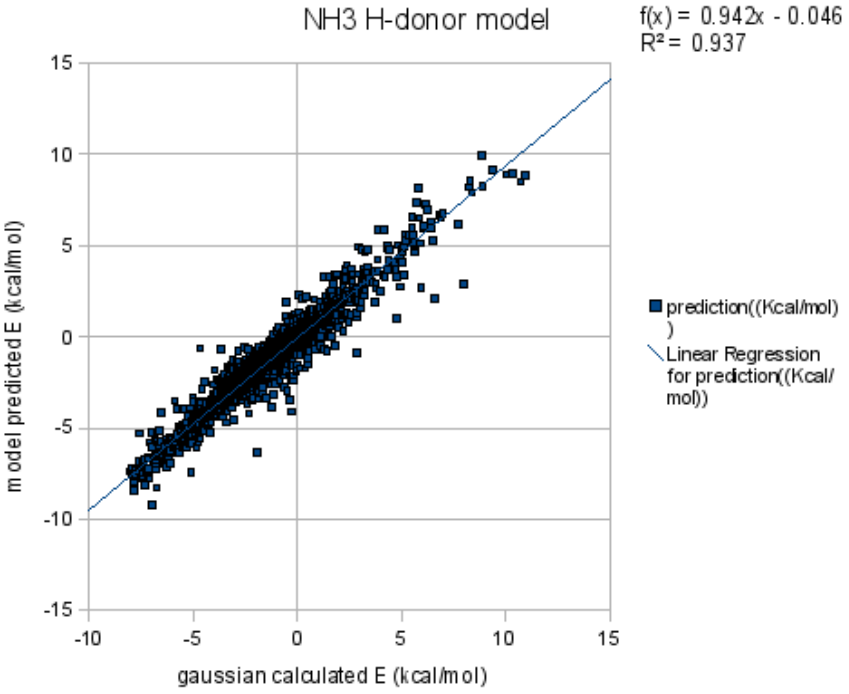


Figure S7 NH3 as H-acceptor (B97-D functional) model

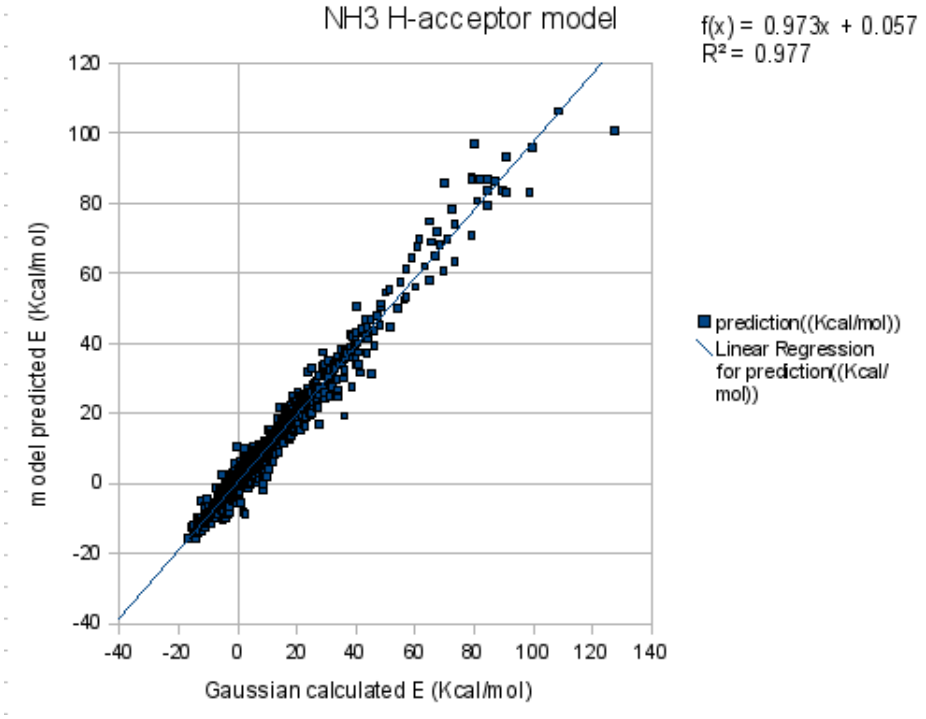


Figure S8 NH4 (B97-D functional) model

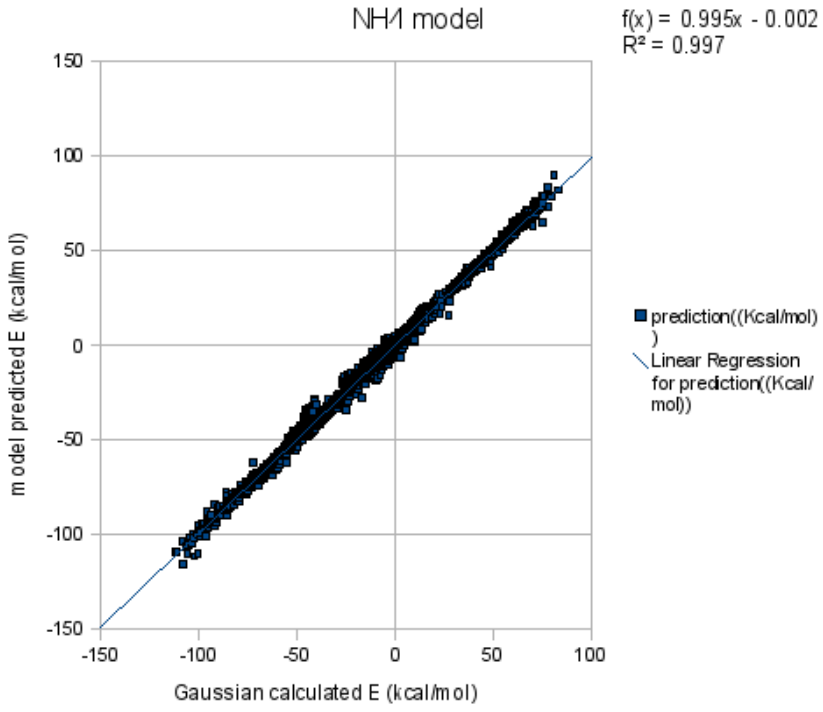


Figure S9 OH (B97-D functional) model

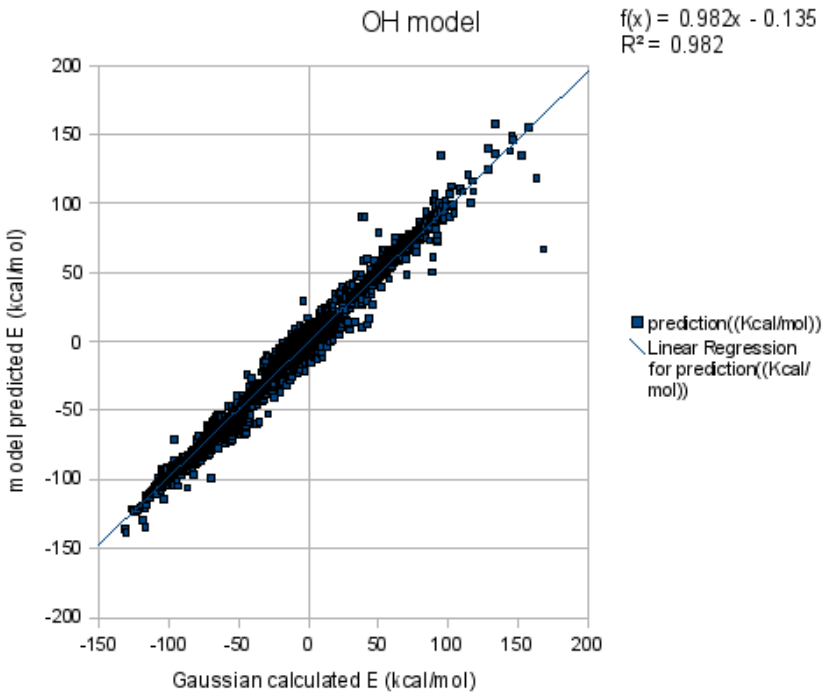


Figure S10 H2O as H-donor (ωB97X-D) model

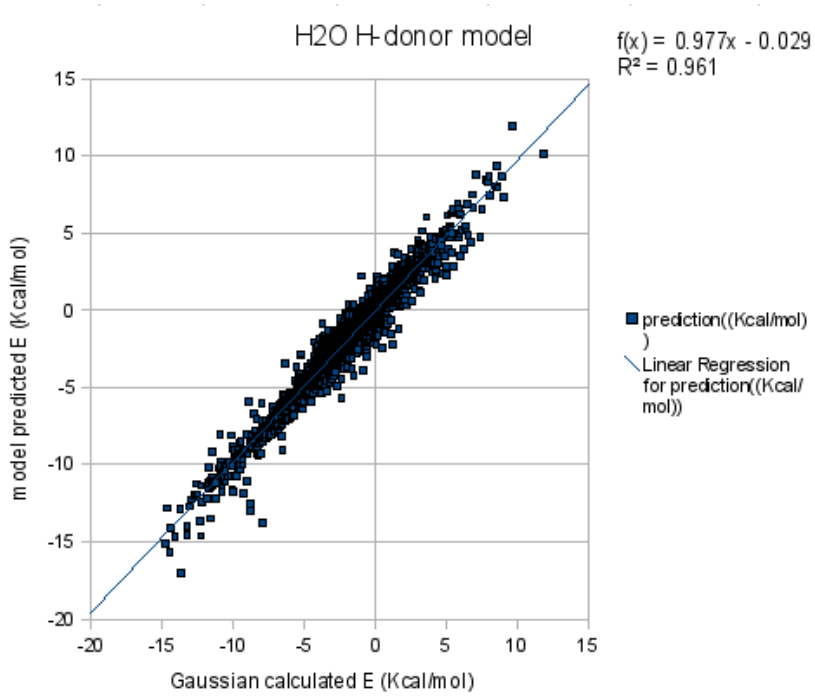


Figure S11 H2O as H-acceptor (ωB97X-D) model

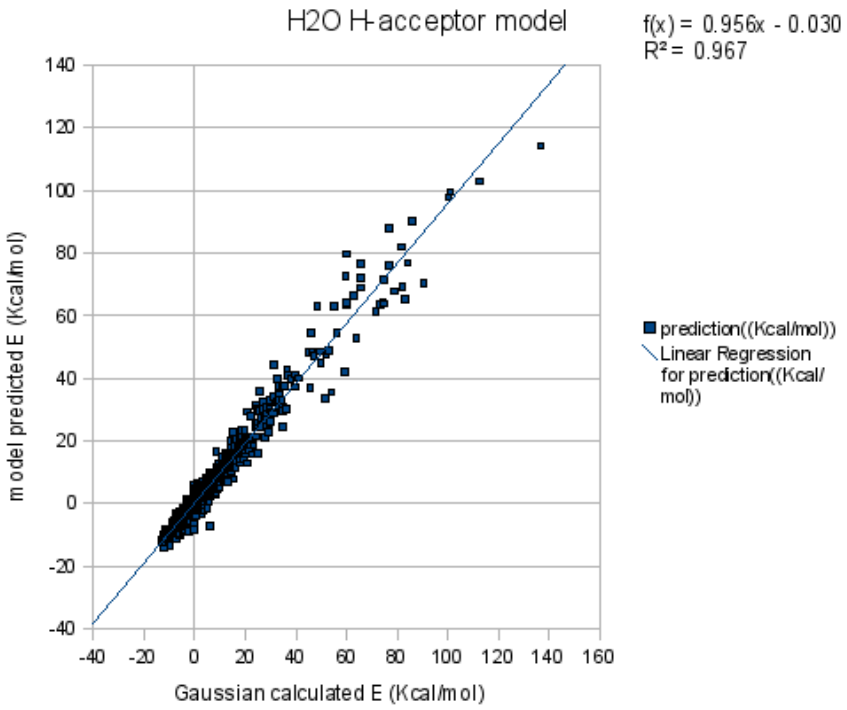


Figure S12 Formamide as H-donor (ωB97X-D) model

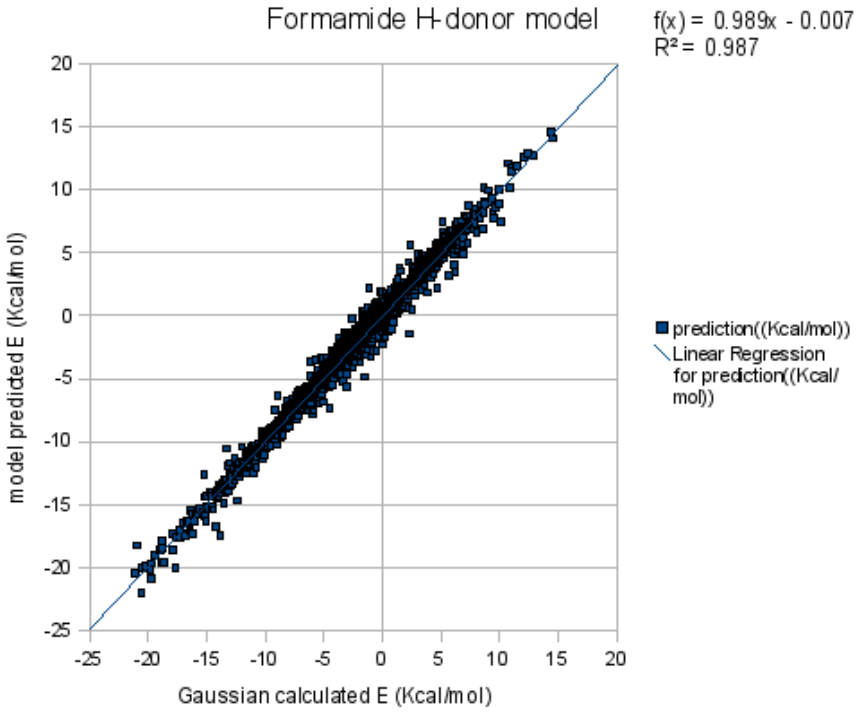


Figure S13 Formamide as H-acceptor (ωB97X-D) model

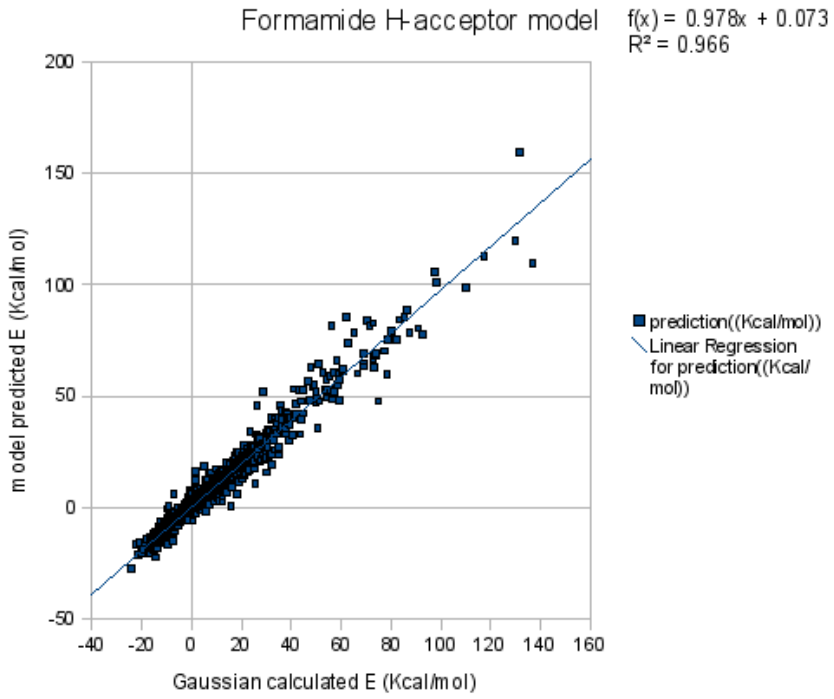


Figure S14 NH3 as H-donor (ωB97X-D) model

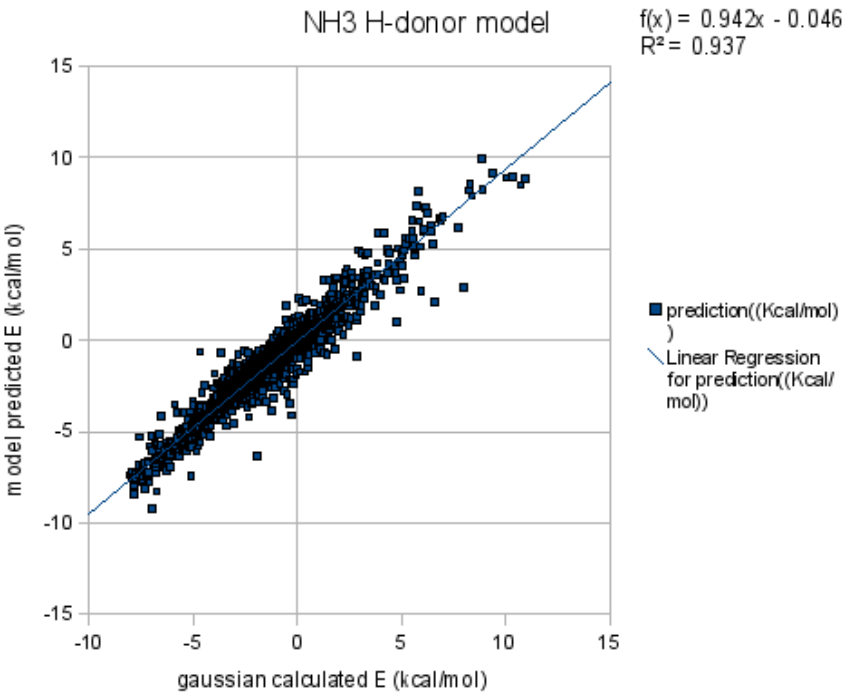


Figure S15 NH3 as H-acceptor (ωB97X-D) model

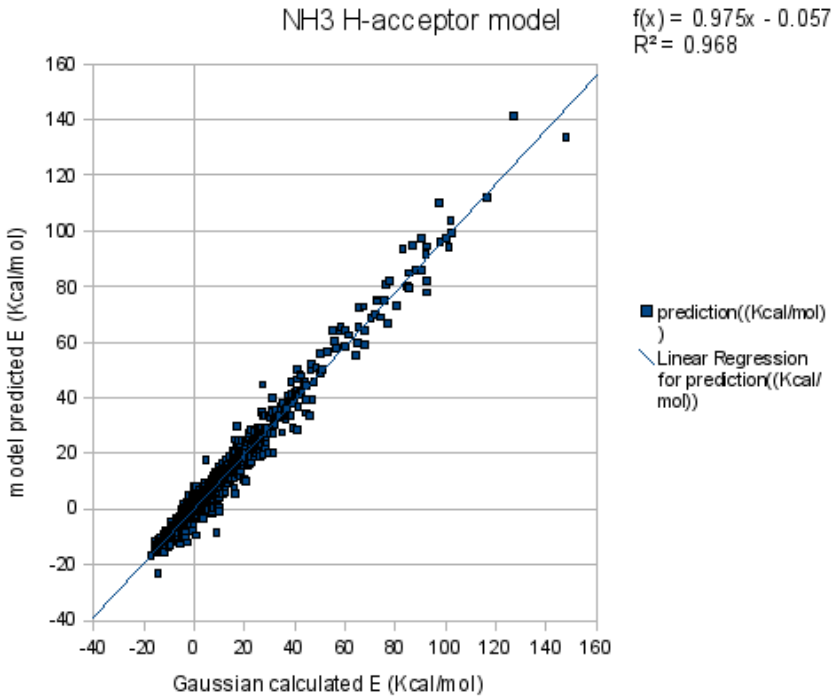


Figure S16 NH4 (ωB97X-D) model

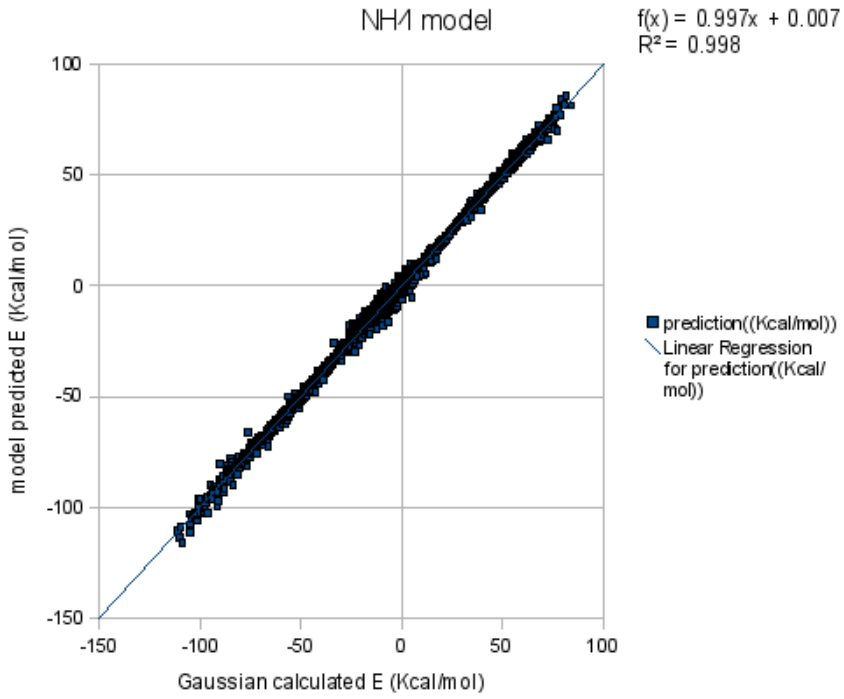
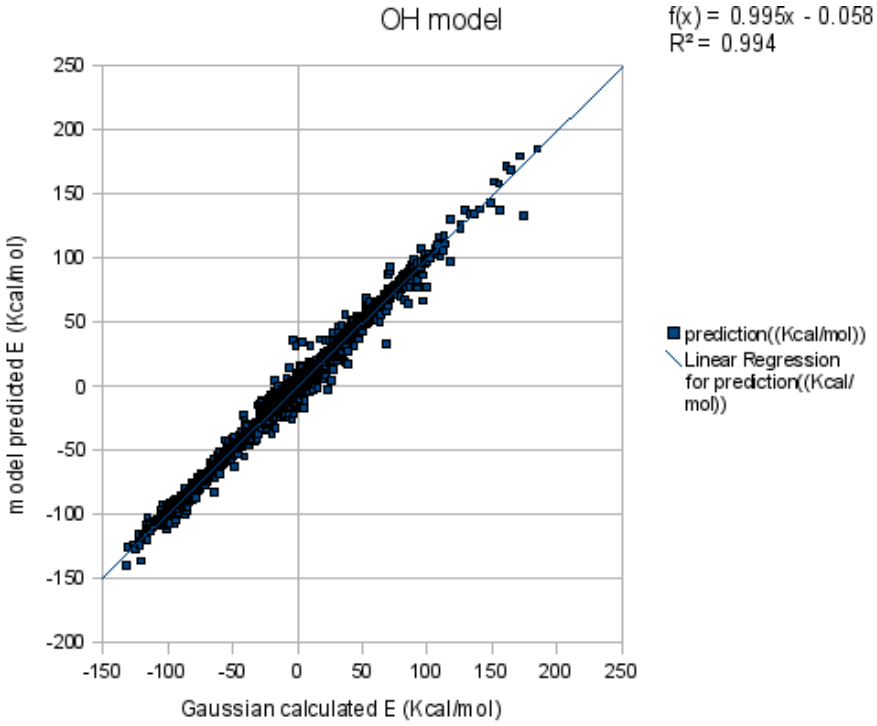


Figure S17 OH (ωB97X-D) model



S7. (un)substituted pyridines calculated and predicted interaction energies and their pK_{BHX}.

Compound	Gaussian calculated E (Kcal mol^{-1})	Model predicted E (Kcal mol^{-1})	pK_{BHX}
Pyridine	-6.12	-6.36	1.86
P-Chloropyridine	-5.67	-5.63	1.54
P-aminopyridine	-6.99	-6.47	2.56

References

1. Cerny, J.; Hobza, P. Non-covalent interactions in biomacromolecules. *Phys. Chem. Chem. Phys.* **2007,** *9*, (39), 5291-5303.
2. Daabkowska, I.; Jurecka, P.; Hobza, P. On geometries of stacked and H-bonded nucleic acid base pairs determined at various DFT, MP2, and CCSD(T) levels up to the CCSD(T)/complete basis set limit level. *J. Chem. Phys.* **2005,** *122*, (20).
3. Hobza, P.; Zahradnik, R.; Muller-Dethlefs, K. The world of non-covalent interactions: 2006. *Collect. Czech. Chem. Commun.* **2006,** *71*, (4), 443-531.
4. Raub, S.; Marian, C. M. Quantum chemical investigation of hydrogen-bond strengths and partition into donor and acceptor contributions. *J.Comput. Chem.* **2007,** *28*, (9), 1503-1515.
5. Gauss, J.; Stanton, J. F. Analytic gradients for the coupled-cluster singles, doubles, and triples (CCSDT) model. *J. Chem.Phys.* **2002,** *116*, (5), 1773-1782.
6. Tsuzuki, S.; Uchimaru, T.; Matsumura, K.; Mikami, M.; Tanabe, K. Effects of basis set and electron correlation on the calculated interaction energies of hydrogen bonding complexes: MP2/cc-pV5Z calculations of H(2)O-MeOH, H(2)O-Me(2)O, H(2)O-H(2)CO, MeOH-MeOH, and HCOOH-HCOOH complexes. *J.Chem.Phys.* **1999,** *110*, (24), 11906-11910.
7. Grimme, S. Semiempirical GGA-type density functional constructed with a long-range dispersion correction. *J. Comput. Chem.* **2006,** *27*, (15), 1787-1799 and references therein.
8. Chai, J. D.; Head-Gordon, M. Long-range corrected hybrid density functionals with damped atom-atom dispersion corrections. *Phys. Chem. Chem. Phys.* **2008,** *10*, (44), 6615-6620 and references therein.
9. Dunning, T. H. Gaussian-Basis Sets for Use in Correlated Molecular Calculations .1. The Atoms Boron through Neon and Hydrogen. *J. Chem. Phys.* **1989,** *90*, (2), 1007-1023.
10. Kendall, R. A.; Dunning, T. H.; Harrison, R. J. Electron-Affinities of the 1st-Row Atoms Revisited - Systematic Basis-Sets and Wave-Functions. *J. Chem. Phys.* **1992,** *96*, (9), 6796-6806.
11. Woon, D. E.; Dunning, T. H. Gaussian-Basis Sets for Use in Correlated Molecular Calculations .3. The Atoms Aluminum through Argon. *J.Chem. Phys.* **1993,** *98*, (2), 1358-1371.
12. Davidson, E. R. Comment on Dunning's correlation-consistent basis sets - Comment. *Chem.Phys. Lett.* **1996,** *260*, (3-4), 514-518.
13. Fujita, T.; Nishioka, T.; Nakajima, M. Hydrogen-Bonding Parameter and Its Significance in Quantitative Structure-Activity Studies. *J. Med. Chem.* **1977,** *20*, (8), 1071-1081.
14. Charton, M.; Charton, B. I. Hyper-Conjugation as a Parameter in Correlation-Analysis. *J. Org. Chem.* **1982,** *47*, (1), 8-13.
15. Yang, G. Z.; Lien, E. J.; Guo, Z. R. Physical Factors Contributing to Hydrophobic Constant-Pi. *Quant. Struct.-Act. Relat.* **1986,** *5*, (1), 12-18.
16. Seiler, P. Interconversion of Lipophilicities from Hydrocarbon-Water Systems into Octanol-Water System. *Eur. J. Med. Chem.* **1974,** *9*, (5), 473-479.
17. Kamlet, M. J.; Doherty, R. M.; Abboud, J. L. M.; Abraham, M. H.; Taft, R. W. Linear Solvation Energy Relationships .36. Molecular-Properties Governing Solubilities of Organic Nonelectrolytes in Water. *J. Pharm. Sci.* **1986,** *75*, (4), 338-349.
18. Kamlet, M. J.; Doherty, R. M.; Taft, R. W.; Abraham, M. H.; Veith, G. D.; Abraham, D. J. Solubility Properties in Polymers and Biological Media .8. An Analysis of the Factors That Influence Toxicities of Organic Nonelectrolytes to the Golden Orfe Fish (Leuciscus-Idus-Melanotus). *Environ. Sci. Technol.* **1987,** *21*, (2), 149-155.
19. Abraham, M. H.; Duce, P. P.; Prior, D. V.; Barratt, D. G.; Morris, J. J.; Taylor, P. J. Hydrogen-Bonding .9. Solute Proton Donor and Proton Acceptor Scales for Use in Drug Design. *J. Chem. Soc., Perkin Trans. 2* **1989**, (10), 1355-1375.
20. Wilson, L. Y.; Famini, G. R. Using Theoretical Descriptors in Quantitative Structure-Activity-Relationships - Some Toxicological Indexes. *J. Med. Chem.* **1991,** *34*, (5), 1668-1674.

21. Dearden, J. C.; Ghafourian, T. Investigation of calculated hydrogen bonding parameters for QSAR. In *QSAR and molecular modelling: consepts, computational tools and biological applications*, Sanz, F.; Giraldo, J.; Manaut, F., Eds. Prous Science Publishers: Barcelona, 1995; pp 117-119.
22. Dearden, J. C.; Cronin, M. T. D.; Wee, D. Prediction of hydrogen bond donor ability using new quantum chemical parameters. *J. Pharm. Pharmacol.* **1997,** *49*, (Suppl. 4), 110.
23. Gancia, E.; Montana, J. G.; Manallack, D. T. Theoretical hydrogen bonding parameters for drug design. *J. Mol. Graphics Modell.* **2001,** *19*, (3-4), 349-362.
24. Murray, J. S.; Politzer, P. Correlations between the Solvent Hydrogen-Bond-Donating Parameter Alpha and the Calculated Molecular-Surface Electrostatic Potential. *J. Org. Chem.* **1991,** *56*, (23), 6715-6717.
25. Hennemann, M.; Clark, T. A QSPR-approach to the estimation of the pK(HB) of six-membered nitrogen-heterocycles using quantum mechanically derived descriptors. *J. Mol. Model.* **2002,** *8*, (4), 95-101.
26. Kenny, P. W. Hydrogen Bonding, Electrostatic Potential, and Molecular Design. *J. Chem. Inf. Model.* **2009,** *49*, (5), 1234-1244.
27. Smallwood, C. J.; McAllister, M. A. Characterization of low-barrier hydrogen bonds .7. Relationship between strength and geometry of short-strong hydrogen bonds. The formic acid formate anion model system. An ab initio and DFT investigation. *J. Am. Chem. Soc.* **1997,** *119*, (46), 11277-11281.
28. Guerra, C. F.; Bickelhaupt, F. M.; Snijders, J. G.; Baerends, E. J. Hydrogen bonding in DNA base pairs: Reconciliation of theory and experiment. *J. Am. Chem. Soc.* **2000,** *122*, (17), 4117-4128.
29. Koch, W.; Holthausen, M. C. Hydrogen bonds and weakly bound systems. In *A chemist's guide to density functional theory*, 2nd ed.; Wiley-VCH Verlag GmbH & Co. KGaA: Weinheim, Germany, 2002; pp 213-235.
30. Morozov, A. V.; Kortemme, T.; Tsemekhman, K.; Baker, D. Close agreement between the orientation dependence of hydrogen bonds observed in protein structures and quantum mechanical calculations. *Proceedings of the National Academy of Sciences of the United States of America* **2004,** *101*, (18), 6946-6951.
31. Koné, M.; Illien, B.; Graton, J.; Laurence, C. B3LYP and MP2 calculations of the enthalpies of hydrogen-bonded complexes of methanol with neutral bases and anions: Comparison with experimental data. *J. Phys. Chem. A* **2005,** *109*, (51), 11907-11913.
32. Schwöbel, J.; Ebert, R. U.; Kuhne, R.; Schuurmann, G. Modeling the H Bond Donor Strength of -OH, -NH, and -CH Sites by Local Molecular Parameters. *J. Comput. Chem.* **2009,** *30*, (9), 1454-1464.
33. Schwöbel, J.; Ebert, R. U.; Kuhne, R.; Schuurmann, G., Prediction of the Intrinsic Hydrogen Bond Acceptor Strength of Organic Compounds by Local Molecular Parameters. *J. Chem. Inf. Model.* **2009,** *49*, (4), 956-962.
34. Hao, M. H. Theoretical calculation of hydrogen-bonding strength for drug molecules. *J. Chem. Theory Comput.* **2006,** *2*, (3), 863-872.
35. Nocker, M.; Handschuh, S.; Tautermann, C.; Liedl, K. R. Theoretical Prediction of Hydrogen Bond Strength for Use in Molecular Modeling. *J. Chem. Inf. Model.* **2009,** *49*, (9), 2067-2076.
36. Jencks, W. P. Binding energy, specificity, and enzymic catalysis: the circe effect. *Adv. Enzymol. Relat. Areas Mol. Biol.* **1975,** *43*, 219-410.
37. Page, M. I., Entropy Binding-Energy, and Enzymic Catalysis. *Angew. Chem., Int. Ed. Engl.* **1977,** *16*, (7), 449-459.
38. *RapidMiner 5.0.008*, Rapid- l GmbH: Dortmund, Germany, 2010.
39. Mierswa, I.; Wurst, M.; Klinkenberg, R.; Scholz, M.; Euler, T. YALE: rapid prototyping for complex data mining tasks. In *Proceedings of the 12th ACM SIGKDD international*

conference on Knowledge discovery and data mining, ACM: Philadelphia, PA, USA, 2006; pp 935-940.