PALACKÝ UNIVERSITY, OLOMOUC

Faculty of Science

Department of Physical Chemistry

Mgr. Saltuk Mustafa EYRİLMEZ

Doctoral Dissertation

Quantum Mechanical Investigation of Non-Covalent Interactions in Host-Guest and Protein-Ligand Complexes

Supervisor

Prof. Ing. Pavel Hobza, DrSc., dr. h. c., FRSC

Czech Advanced Technology and Research Institute, Palacký University Olomouc

Institute of Organic Chemistry and Biochemistry of the Czech Academy of Sciences, Prague

Olomouc 2021

UNIVERZITA PALACKÉHO V OLOMOUCI

Přírodovědecká Fakulta

Katedra Fyzikální Chemie

Mgr. Saltuk Mustafa EYRİLMEZ

Disertační práce

Quantum Mechanical Investigation of Non-Covalent Interactions in Host-Guest and Protein-Ligand Complexes

Školitel

Prof. Ing. Pavel Hobza, DrSc., dr. h. c., FRSC

Český institut výzkumu a pokročilých technologií, Univerzity Palackého v Olomouci

Ústav organické chemie a biochemie, Akademie věd České republiky, v.v.i., Praha

Olomouc 2021

Declaration of Authorship

I declare that I have worked out this dissertation titled, "QUANTUM MECHANICAL INVESTIGATION OF NON-COVALENT INTERACTIONS IN HOST-GUEST AND PROTEIN-LIGAND COMPLEXES" by myself using the cited references and it has not been submitted elsewhere for any academic degree.

Prague, 15th May, 2021 Saltuk Mustafa Eyrilmez

Dedicated to my son, Taner…

Acknowledgement

I am grateful to Prof. Pavel Hobza not only for being a good consultant in the field of science, but also for being a fair leader and a devoted guide in every field of real life.

I would like kindly to thank to Mrs. Helena Černá. Her helps made our life a lot easier.

I am thankful to my colleagues Dr. Jan Řezáč, Dr. Martin Lepšík, Dr. Jindřich Fanfrlík, Dr. Adam Pecina for their constant support and encouragement.

I would also like to thank Dr. Cemal Köprülüoğlu with my sincere wishes for his trust in what I will do here and his first steps.

I am grateful to Prof. Michal Otyepka, Sylva Kaděrková and Dana Gronychová for their amazing support.

I am grateful to my family for all their support that they have given me in becoming a scientist.

Finally, I thank my wife Gizem Oyman Eyrilmez for her unlimited help to make my dreams come true.

Contents

Tables

Figures

List of Abbreviations

AM1	Austin Model 1
B21	*closo,closo*-$[B_{21}H_{18}]^-$
CADD	Computer-Aided Drug Design
CC	Coupled Cluster
CD	Cyclodextrin
COSMO	Conductor-Like Screening Model
D3H4X	Dispersion, Hydrogen and Halogen Bonding Correction
DFT	Density Functional Theory
DFTB	Density Functional Based Tight Binding
HB	Hydrogen Bond
HF	Hartree-Fock
HTS	High-Throughput Screening
L	Ligand
LBDD	Ligand-Based Drug Design
MM	Molecular Mechanics
MP	Møller–Plesset
P	Protein
PCM	Polarizable Continuum Model
PDB	Protein Data Bank
P-L	Protein-Ligand Complex
PM6	Parametrized Method 6
PM7	Parametrized Method 7
QM	Quantum Mechanics
SBDD	Structure-Based Drug Design
SBVS	Structure-Based Virtual Screening
SF	Scoring Function
SMD	Solvation Model Based on Density
SQM	Semiempirical Quantum Mechanics
vDW	van der Waals
VS	Virtual Screening

Abstract

The abundance and variety of noncovalent interactions shape us and the life surrounding us. All the natural processes evolved due to existence of these effects. Understanding the basics of interactions is a key factor to manipulate them rationally. However, a detailed decomposition of these weak forces is not an easy task.

Computational chemistry is a multidisciplinary branch of science, born from the combination of scientific data collected systematically for centuries and cutting-edge technology. It allows us to analyze, model and even predict the properties of chemical systems. The complex nature of non-covalent interactions does not depend on their quantity in molecular complexes. In general, case-dependent differences in a very small fraction of the electronic structure are key to the binding with a high specificity. However, the weakness of these interactions makes them extremely difficult to observe. Accurate descriptions of non-covalent interactions require demanding QM methods. On the other hand, linear-scaling SQM methods in combination with implicit solvent model and successful corrections for non-covalent interactions enabled us to evaluate properties in protein-ligand systems.

Structure generation and validation is the most critical step for all physics-based CADD approaches. We used molecular dynamics to systematically generate and scan set of geometries for host-guest systems. The results of this approach are highly input geometry dependent for the systems such as protein-ligand complexes that include huge number of structural degrees of freedom. Extensive docking calculations can provide a near-native binding mode more efficiently than running long MD simulations with different setups. We evaluated the performance of SQM/COSMO scoring function capabilities on sampling and ranking studies on many different P-L complexes from diverse set of targets.

Finally, we built an efficient virtual screening pipeline which is capable of filtering out redundant poses and shrinks the database to an affordable size for further SQM/COSMO scoring calculations.

CHAPTER 1

INTRODUCTION

The acceleration of technological evolution in the beginning of 21st century has profoundly changed the way of how we live today. Electronic devices have become universally accessible. Their storage capacity and speed have been increasing exponentially, creating an astonishing amount of information which is shared on a worldwide scale. The use of supercomputers helped to broaden the horizons of scientists who investigate mechanisms of life from the smallest and most fundamental components. Simulations with suitable parameters have provided useful insights. In computational chemistry, we can run calculations to obtain information such as molecular geometries, reaction rates, interaction energies, physicochemical properties, and spectra [1]. These capabilities of computational chemistry have become indispensable in computer-aided drug design (CADD) branch of pharmaceutical industry where researchers design, evaluate and improve properties of drug candidates. CADD can go along two directions for modelling candidate molecules: ligand-based drug design (LBDD) and structure-based drug design (SBDD). While LBDD uses only the information of binding small molecules, SBDD uses the three-dimensional structure information for both candidate molecules and their biomolecular targets, mostly proteins. The main goal of SBDD is to obtain a better binder in terms of affinity and specificity [2]. A common drawback of most SBDD methods is an inaccurate description of noncovalent interactions which play a major role in recognition and binding. Their accurate description can be obtained with a high accuracy by using advanced quantum mechanical (QM) methods. But their scalability with system size

sets the limit to tens of atoms. Protein-ligand complexes, however, are composed of thousands of atoms. One solution to this conundrum is to use linear-scaling semiempirical QM (SQM) methods which are applicable to systems of up to 10,000 atoms. Nevertheless, their accuracy had to be increased by combining them with empirical corrections for noncovalent interactions. Such a tool enabled us to successfully evaluate binding energies of huge molecules with high accuracy [3].

1.1 Quantum Chemistry

Quantum mechanics (QM) theory has revolutionized theoretical description of molecules and gave rise to the field of quantum chemistry. QM methods are either based on describing the electron distribution using wave-functions and solving the Schrödinger equation (*ab initio* methods, such as Hartree-Fock (HF), Coupled cluster (CC), Møller–Plesset (MP)) or by describing the electron density using Density Functional Theory (DFT methods) [4]. Semiempirical QM (SQM) methods use some experimental values (parameters) to replace calculation of complicated integrals which would otherwise have to be evaluated [5]. The most frequently used SQM methods such as AM1 [6], PM6 [7] and PM7 [8] are approximations of the HF theory. Density-functional tight-binding (DF-TB) [9] is an SQM method, which is based on DFT, and became popular in recent years [10].

Larger systems (more than hundreds of atoms) can be calculated with DFT methods and huge ones (thousands of atoms) with SQM. These vacuum calculations are frequently supplemented with methods for an implicit treatment of solvent, such as PCM [11], COSMO [12], or SMD [13], [14]. This is much more efficient than having to treat solvent explicitly.

Another way to increase efficiency of calculations in large systems is using hybrid quantum mechanical/molecular mechanical (QM/MM) methods. The main idea of this approach is based on the evaluation of quantum effects which are localized in a smaller

part of system as in active sites of protein-ligand complexes. The rest of the system act as an embedding surrounding and described by molecular mechanics (MM) methods [15].

MM methods describe molecules using classical mechanics. Atoms are treated as spheres of a mass and a charge and are attached together via springs (bonds). The missing description of electrons is included via a set of parameters – so called force field. Several successful protein force fields have been developed over decades. These are AMBER (Assisted Model Building and Energy Refinement) [16], CHARMM (Chemistry at HARvard Macromolecular Mechanics) [17], GROMOS (Groningen Molecular Simulation System) [18] and OPLS (Optimized Potentials for Liquid Simulations) [19]. MM methods are suitable for investigations of motions and structural evaluation of huge systems, such as biomolecules, surrounded by thousands of explicit water molecules (TIP3P [20] or SCP/E [21]) over time in molecular dynamics (MD) calculations [1]. MM implicit solvation methods can also be used (GB [22], PB [23]).

1.2 Noncovalent Interactions

Noncovalent interactions govern the majority of biological processes on Earth. The most important ones are ionic bonds (charge···charge), hydrogen bonds and London dispersion interactions (Table 1.1).

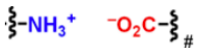
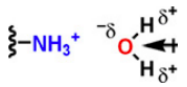
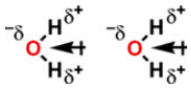
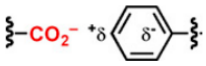
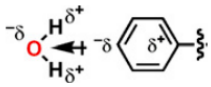
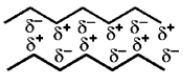
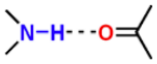
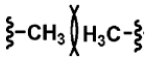
Ionic bonds are caused by the attraction between atoms of opposite charge. These interactions act over long ranges (typically nanometers). The major contribution to the binding comes from the electrostatic interactions. An ion pair between two amino acid side chains of a protein is called a salt bridge. They can occur between the carboxylate groups (e.g. from the side chains of Asp of Glu) and the amino (Lys or N-terminus) or guanidinium (Arg) moieties [24].

Hydrogen bond (HB) is schematically described as X-H⋯Y where the dots denote the noncovalent bonding. X-H represent the HB donor in which X is more electronegative than H and Y represents the acceptor which can be an atom, an anion, a molecule or a fragment of a molecule. Y serves as an electron-rich region such as a lone electron pair or π electron density [25]. The strength of HB can vary from 1 kcal/mol in vacuum (e.g. C-H⋯π, around 1-1.5 kcal/mol) [26] to few kcal/mol (e.g. N-H⋯O, O-H⋯O, 5-7 kcal/mol) [27]. The strongest HB can be seen in F-H⋯F- interactions (39 kcal/mol) due to extreme electronegativity of F atoms [28].

London dispersion interactions (previously called van der Waals; vDW) act on atoms or molecules due to induced dipole-induced dipole dispersion forces. These interactions are effective in very short range (tenths of nanometers) [29].

Some other examples for special type of noncovalent bonds are dihydrogen bonds, π…π interactions, halogen bonds, dative bonds [23]. Most of the noncovalent interactions are well described at MM level. But special cases which represent some property originating from purely quantum nature, require ad hoc corrections (e.g. halogen bonds [30]) or use of QM methods.

Table 1.1 Schematical representation of noncovalent interactions and their dependencies by distance. The table is taken from R. R. Knowles and E. N. Jacobsen, PNAS, 2010, Vol. 107, no. 48, 20678-20685

Noncovalent interaction		Energy dependence on distance
Charge–charge		$1/r$
Charge–dipole		$1/r^2$
Dipole–dipole		$1/r^3$
Charge-induced dipole		$1/r^4$
Dipole-induced dipole		$1/r^5$
Dispersion		$1/r^6$
H-bond		Complicated $\sim 1/r^2$
Steric repulsion		$1/r^{12}$

Even though individually any of these interactions (Table 1.1) are much weaker than a single covalent bond, combination of many noncovalent interactions provides the stability of complexes. Moreover, dissociation is possible easier than breaking a covalent bond, a trait important for molecular biology [31].

Reliable description of all types of noncovalent interactions is a prerequisite for a trustworthy characterization of the binding in molecular complexes. A bottom-up approach starting from small molecule databases [32], [33], parametrizing SQM

methods and then increase the size of the systems is a strategy to accomplish this hard task [34], [35], [36].

1.3 Solvation and Hydrophobic effect

Since most of chemical reactions occur in solution, consideration of solvent effects is essential. Solvent molecules interact both, with the solutes and with other solvent molecules via the noncovalent interactions discussed earlier (H-bonding in case of water) [24]. Water is the most abundant solvent in living systems. A biochemical reaction such as a formation of a protein-ligand complex (P-L) occurs in water environment. Both partners need to be partially desolvated to make the binding interface. In some case, the binding is partly mediated by water networks [37].

Water solvent entropy is behind another driving force of binding – the hydrophobic effect. This is connected with ordering of water molecules around nonpolar solutes, lowering unfavorably their entropy. Thus, the nonpolar groups/ molecules are forced to come close, increasing entropy favorably. The contributions of hydrophobic effect to protein folding, membrane formation and receptor-ligand binding are essential [24].

1.4 Molecular Complexes

Molecular complexes are structures held together by noncovalent interactions. In some cases, interaction may occur via multiple binding sites with different characteristics thus increasing structural organization [38]. By using this structural information, we can design host molecules specific for their guests or new drug molecules specific for their biomolecular targets.

1.4.1 Host-Guest Complexes

In supramolecular chemistry, hosts are larger organic molecules (such as cyclodextrins) that can specifically bind a smaller molecule or ion, called a guest. These associations are called host-guest complexes. Their binding occurs via non-covalent interactions [38] (see Figure 1.1.).

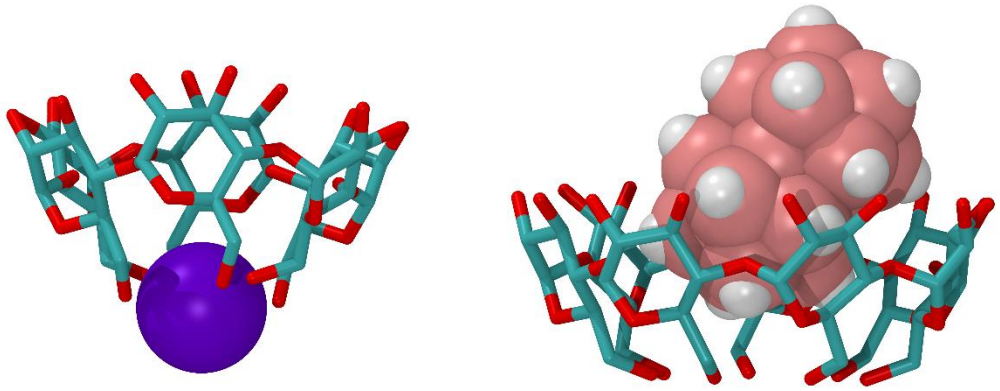


Figure 1.1 Host-guest complex structures of an α-cyclodextrin···K^+ and a β-cyclodextrin··· $[B_{21}H_{18}]^-$

The specificity and affinity of host-guest interactions can be tuned. Due to their adjustable properties, host-guest chemistry has been extensively studied in fields of molecular recognition, biosensors, analytical separation and purification, catalysis and drug development [39]. Host molecules can be used not only to recognize and bind to specific guest molecules but also to keep them encapsulated for specific purposes. For this reason, bioavailability of drug molecules can be improved by using host molecules such as cucurbiturils [40] and cyclodextrins [41]. Experimental studies in vacuum are good model systems for calculations of interactions, where only deformation energy of the host may be difficult. In solution, the affinities can be tuned by the number of water molecules which will be expelled upon guest binding.

1.4.2 Protein-Ligand Complexes

Proteins are one of the most abundant polymers found in all living cells. They are composed of 20 types of amino acids, which differ in charge, polarity, size as well as linearity, cyclicity or aromaticity of their side chains. These distinct chemical properties can lead a huge variety of different protein structures. Proteins play roles, e.g. in signal transmission, recognition or catalysis.

Proteins are highly specialized molecules with a specific three-dimensional structure to perform a unique function. They also give specific responses to environmental changes by the help of small variations in the noncovalent interaction pattern. These interactions play an essential role for proteins to gain and preserve their functional forms, recognition and binding to their ligands and often allow conformational changes upon binding (Figure 1.2). Binding of a ligand to a binding site of a protein should also satisfy the complementarities in ligand size, shape, charge distribution and hydrophilic or hydrophobic characters [37].

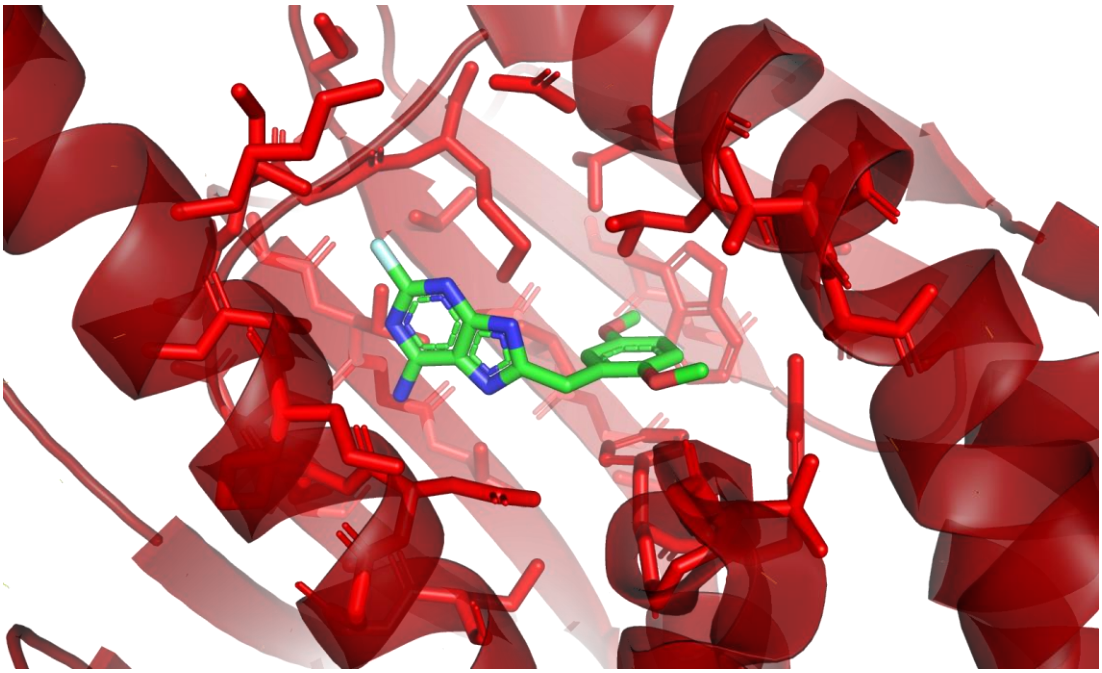


Figure 1.2 Binding of a ligand (green sticks) in the active site of HSP90 protein (red cartoon and sticks; PDB code: 1UYG).

1.4.3 Recognition and Binding

The strength of molecular association between a protein (P) and ligand (L) is quantified using binding affinity. It is an equilibrium between the unbound states of protein and ligand and bound state of protein-ligand (P-L) complex (Figure 1.3) characterized by an equilibrium association constant K_a. The relation between the Gibbs free energy of binding (ΔG) and the equilibrium constant (K_a) is,

$$\Delta G = -RT \ln K_a \quad (1)$$

where R is the gas constant (8.315 J/K/mol) and T is the absolute temperature.

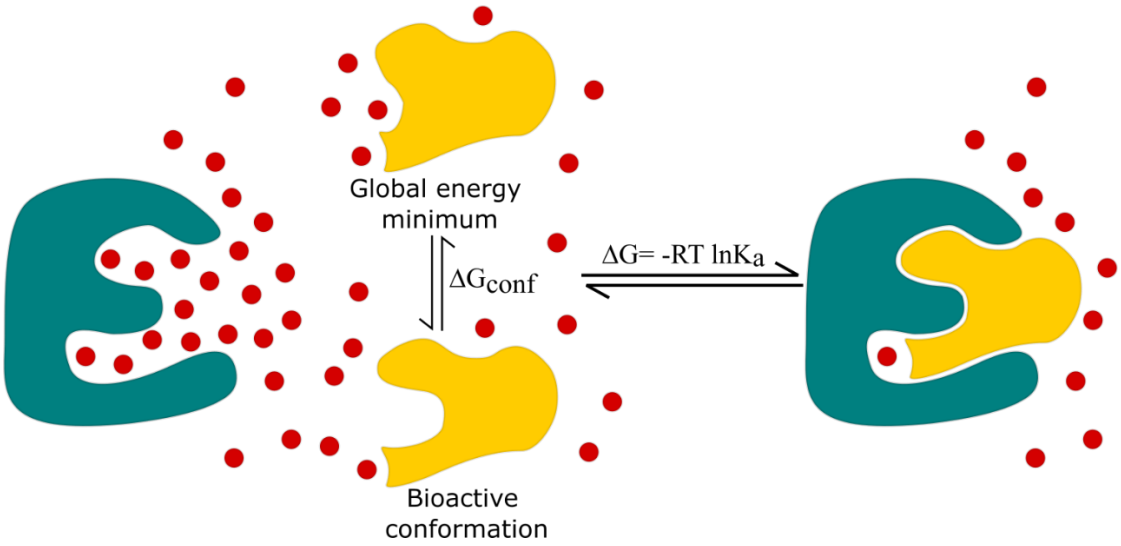


Figure 1.3 Free energy difference of binding. (*Adopted from Textbook of Drug Design and Discovery Fifth edition (2016), p.17)*

A shift of equilibrium through the formation of bound complex as illustrated in the Figure 1.3 results in a higher affinity. In this case, K value becomes more positive and ΔG more negative. In medicinal chemistry the affinity is given either by inhibition constant (K_i) or the half maximal inhibitory concentration (IC_{50}). Since $K_a = 1/K_i$ the equation 1 can be written as

$$\Delta G = RT \ln K_i \tag{2}$$

ΔG has enthalpic (ΔH) and entropic (ΔS) components:

$$\Delta G = \Delta H - T\Delta S \tag{3}$$

In some cases, IC_{50} can be used instead of K_i values, in which case the IC_{50} values are converted to the inhibition constant K_i by the Cheng-Prusoff equation:

$$K_i = \frac{IC_{50}}{\left(1 + \frac{[L]}{K_D}\right)} \tag{4}$$

where [L] is the concentration of the ligand used in the assay, K_D is the affinity of the ligand for the receptor.

IC_{50} value is related to activity and K_i is related to binding. While K_i is independent of ligand and its concentration, IC_{50} values are concentration dependent. This property of K_i makes comparisons of different assays possible [42].

1.5 Computer-Aided Drug Design

With the rapid development of computational facilities and efficiency, CADD has become an important tool in drug discovery process [43]. Besides that, the exponential growth of protein crystal structures deposited to the Protein Data Bank (PDB) has expanded the potential of SBDD. The number of crystal structures has reached ~180,000 today (Figure 1.4.). The major tools of SBDD are docking and scoring.

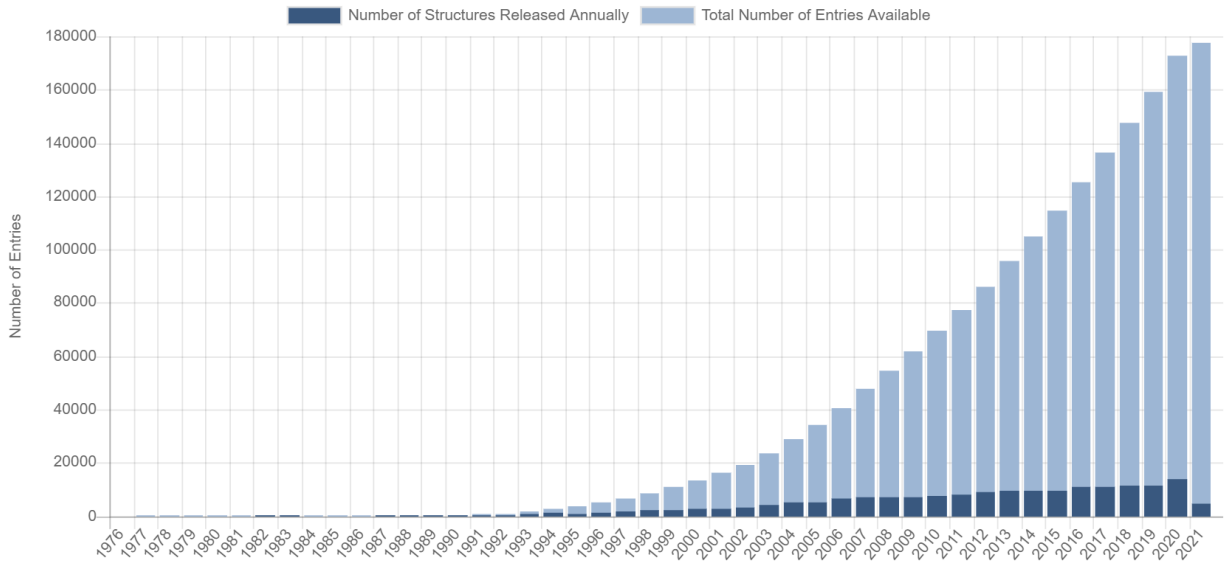


Figure 1.4 Overall growth of released structures per year (rcsb.org, Access date: May 13, 2021)

1.5.1 Ligand Docking

Molecular docking aims at identifying the native structures of P-L complexes using computations. A large number of docking programs and web services have been developed [44], such as AutoDock [45], DOCK [46], GOLD [47], Glide [48], AutoDock Vina [49], SMINA [50], PLANTS [51] etc.

Docking protocols produces P-L complex structures. A docking software consists of a search algorithm for generation of P-L complexes and a scoring function. A successful protocol will provide more realistic poses of ligands in an active cavity of a protein.

Search algorithms can explore the binding in three different ways. The simplest is called rigid docking where the ligands are limited with translational and rotational degrees of freedom. Flexible docking explores different positions by adding a conformational freedom to the ligands. The third way is to extend the conformational search space by considering the protein flexibility which is called induced fit docking [52].

1.5.2 Scoring Functions

Scoring functions are used for estimation of noncovalent interactions in a given P-L complex structure by using mathematical approximations. It is the most important component of a molecular docking for the binding pose prediction process [53]. Thus they are mainly responsible for the success or failure of a docking software [54].

Scoring functions can be divided into empirical, knowledge-based and physics-based.

Empirical scoring functions estimate the binding free energy by using a set of parameters which were generated from protein-ligand complexes with known affinities. These parameters are used to describe the interaction as components made

of hydrogen bonding, ionic bonding, non-polar interactions, desolvation and entropic terms which are multiplied by weight constants [53]. Glide Score [55] and DOCK6 [56] are examples for empirical scoring functions.

Knowledge-based scoring functions calculate the affinity by using energy potentials defined for atom or chemical group pairs. Score is given as a sum of each individual interactions [57].

Physics-based scoring functions mostly use MM methods for non-covalent interactions (sum of electrostatic and dispersion interactions) combined with implicit solvation free energy term. Change of internal energy of the ligand (deformation energy) is added to produce the final score [52]. Docking software programs such as DOCK [46], GOLD [58], and AutoDock [59] have some differences in the treatment of hydrogen bonds. The common drawback of MM-based methods is their inherent lack of description of QM effects, such as charge transfer, polarization or σ-hole. QM calculations provide accurate description of these effects but are computationally demanding [60]. SQM-based scoring functions which were introduced by Kenneth Merz group [61] were more cost-effective than QM but had some accuracy issues. SQM-based scoring function showed superior performance over MM in the case of metalloprotein [61], [62]. However corrections were needed for inaccurate descriptions of hydrogen bonding and dispersion interactions [63], [64]. We developed these in our laboratory and resulting PM6-D3H4X method is fast and provides accurate description of all types of non-covalent interactions without need for any specific parametrization. The PM6-COSMO SF was successfully used for hundreds of P-L complexes [65], [66].

1.5.3 Structure-Based Virtual Screening

Drug discovery process was based on random searching and empirical observations until 1980s. This process was improved by high-throughput screening (HTS) which

allows for automated screening of thousands of compounds against a target (a protein or a cellular assay) in a very short time [52]. Same strategy also started to be applied as virtual screening (VS) after the successful applications CADD studies. VS became a necessary tool for assisting the drug development processes. Structure-based virtual screening (SBVS) is a technique which predicts the affinity of the ligand molecules against a target with a known 3D structure by forming complex structure [67]. Although a combination of molecular docking followed by reliable scoring approach sounds as a good idea, a universal solution for the challenges regarding generating correct binding positions or accurate scoring has not yet achieved.

CHAPTER 2

PROJECTS

2.1 Host-Guest Complexes

Understanding host-guest interactions is an important step towards building our knowledge of noncovalent interactions. Besides their practical use mentioned in Chapter 1.4.1, host-guest complexes serve as great templates for computational chemistry by having challenging chemical properties comparing to their dimensions. Host molecules can vary by size which can affect their response on binding to guest molecules in steric, conformational and electronic manners. We have shown the gas phase interactions of *closo,closo*-$[B_{21}H_{18}]^-$ (B21) with macrocyclic α-, β- and γ-cyclodextrin (CD) host molecules with the existence of two K^+ counterions. (See, Publication A).

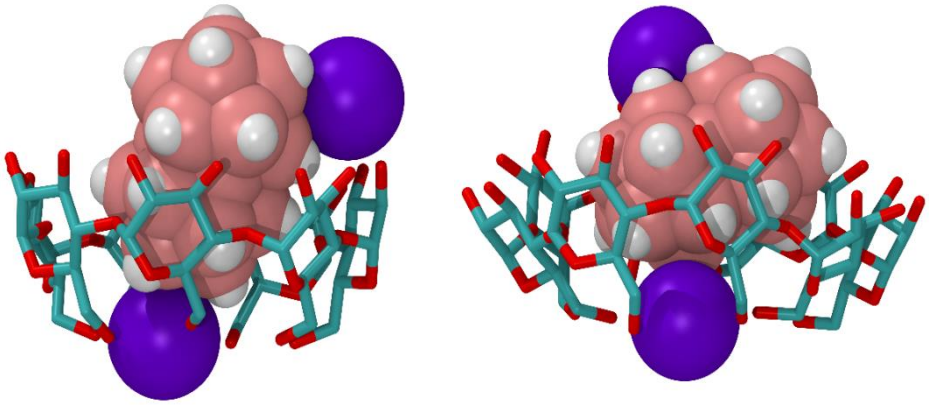


Figure 2.1 The most stable complexes of $[\beta\text{-CD} + B21 + 2K]^+$ (left) and $[\gamma\text{ -CD} + B21 + 2K]^+$. *Reprinted from Publication A.*

After the initial energy scans, it is shown that β- and γ-CD hosts can accommodate B21 guest (Figure 2.1).

Interaction energies were computed according to the eqn (5)

$$\Delta E = E_{(total\ complex)} - E_{(host+K^+)} - E_{(B21+K^+)} \quad (5)$$

Even the γ-CD showed more deformation upon complex formation, overall interactions for both [β-CD + B21 + 2K]$^+$ and [γ -CD + B21 + 2K]$^+$ were almost identical (-51.8 and -51.1 kcal/mol, respectively).

2.2 Protein-Ligand Complexes

Physics-based approaches in SBDD field require well refined three-dimensional structures of protein-ligand complexes. The performance of methods is evaluated under two criteria as sampling power and ranking power. Sampling power measures the ability of picking correct binding mode within a set of conformations of a ligand in the active cavity. Ranking power term indicates the success rate of ordering predicted affinities of different compounds versus their experimental affinities. If the applicability of the method is suitable for processing large databases, screening power becomes an important parameter for evaluation of enrichment in VS studies.

2.2.1 Sampling Power

Docking/scoring is one of the most frequently used tool in SBDD. While docking algorithms search for poses, SFs rank them by their predicted affinities. Ideally, a successful docking/scoring method provide the native binding pose as best binder. Most of the functions fail for finding or accurately scoring the native binding poses because of heavy approximations or missing parameters. Therefore, we developed and

tested an SQM based scoring function which can calculate noncovalent interactions with a high accuracy.

Since QM based methods requires high quality structures for interaction energy calculations, we applied strict criteria for the selection of crystal structures (See the Method, Publication B).

Based on the selections, we ended up with 17 diverse set of protein structures. The SQM/COSMO SFs performed better than all other classic SFs with a significantly lower hard false positive (HFP) rates per target (see Publication B, Figure 1B) and in total (Figure 2.2).

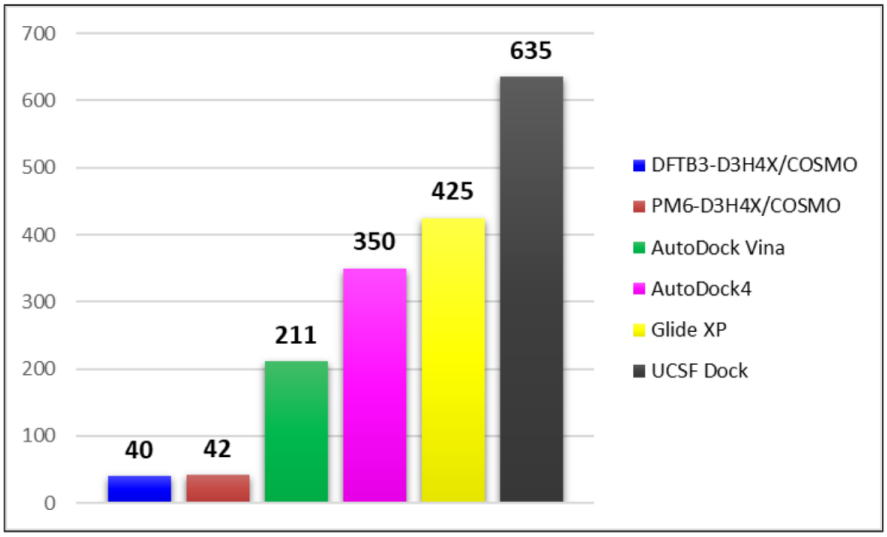


Figure 2.2 Number of total HFPs for six scoring functions. (*Reprinted from Publication B, Figure 1A)*

2.2.2 Ranking Power

Success of the ranking power of a SF can be measured by correlation of experimental binding affinities to computationally predicted affinities. Basically, application of the methods is the same as in sampling studies. But in the ranking case

(Publication C), an SF should additionally be able to distinguish the stronger binders from the weaker ones. For this task, we used a database of 10 carbonic anhydrase II (CAII) inhibitors. Having a Zn^{2+} in the active site of the protein was the main challenge for all the SFs. SQM/COSMO outperformed all other SFs and showed better correlation (R^2) and predictive index (PI) (detailed in the Publication C) performances (0.77 and 0.92, respectively).

2.2.3 Screening Power

Similar to the ranking step, the main aim of the VS studies is to prioritize the active molecules from the inactive ones. But in this case, instead of having few inhibitors, we must deal with huge libraries consisting of at least few thousands of compounds. Processing huge libraries require fully automated consistent and specific preparation protocols for each scoring function. Furthermore, resource and time management or each SF becomes a necessary step.

In our virtual screening study (Publication D), we selected a database consisting of 4541 inhibitors and decoys prepared for HSP90 protein from DUD-E (a Database of Useful Decoys-Enhanced). We applied virtual screening by using 9 different standardly used scoring functions along with our MM (based on AMBER/GB), SQM_1 (SQM/COSMO scoring applied on AMBER forcefield optimized geometries) and SQM_2 (SQM/COSMO scoring applied on geometries generated by restrained optimization protocol) scoring functions. Application of virtual screening protocols and result evaluation steps are detailed in the publication.

MM, SQM_1 and SQM_2 scoring functions are pure physics-based scoring functions. They treat the scoring using the same interaction energy calculation formula without any weight on any of the terms. Interestingly for MM, all our SFs outperformed other conventional scoring functions in early and overall enrichment

comparisons. This shows us that the interaction energy calculation successfully included major contributions more accurately.

Another interesting case is seen when we switch from MM to SQM_1. Even though we got a higher early enrichment for SQM_1 scoring, MM performed better in overall enrichment. Simply, this was due to incompatibility of geometry generation and energy evaluation methods. We fixed the issue by forcing the AMBER optimization protocols to use restraints generated from SQM/COSMO optimized isolated ligand. This solution brings us the almost best possible overall enrichment. It also emphasizes a very tiny but extremely important detail: accurate definition of the geometry is the first and most important key.

CHAPTER 3

CONCLUSION

In quest for a universal method for accurate descriptions of non-covalent interactions in host-guest or protein-ligand complexes we need to evaluate every possible scenario which may affect the results. In this aspect, computational chemistry is not only a branch to give us answers, but also produces some questions to be answered experimentally. Multidisciplinary collaborations of different fields with computational chemistry leads us to find more efficient ways to understand structure-activity relationship further.

In this thesis, we first dealt with host-guest molecule interactions (Publication A). Understanding of these interactions are important for determining physicochemical behavior of the boron-cage structure in an organic cavity and the respond of the guest molecule. Also, one of the most important subjects in this study was the introduction of K^+ ions to the calculations. While they were contributing to the structural stability, they were also greatly increasing the computational demand because of increased degrees of freedom.

In the following project (Publication B), we evaluated the sampling power performance of two SQM based SFs versus other SFs on a diverse set of protein-ligand complexes. This study also shown us the generality of the SQM based SFs.

Next, (Publication C) we presented the ranking power of SQM/COSMO scoring function on a challenging set of 10 inhibitors binding through Zn^{2+} of carbonic

anhydrase II protein. While we were getting fairly good results from SQM/COSMO scoring function, there were no correlation from other standard scoring functions.

Encouraging results enabled us to evaluate the screening power of SQM/COSMO scoring function on a large database made of active and decoy compounds (Publication D). Indications from the initial tests showed that an application of SQM/COSMO scoring function would only be possible by systematically eliminating the redundant structures obtained from extensive docking calculations. By this way, we achieved an enrichment value close to perfect case by calculating only 1.5% of the generated screening database, at SQM/COSMO level.

As a conclusion, SQM/COSMO provides the best compromise between the computational cost and accuracy of describing all types of non-covalent interactions.

BIBLIOGRAPHY

[1] E. G. Lewars, *Computational chemistry: Introduction to the theory and applications of molecular and quantum mechanics: Third Edition 2016*. 2016.

[2] C. J. Krusemark, *Drug Design: Structure- and Ligand-Based Approaches* . Edited by Kenneth M. Merz, Jr., Dagmar Ringe, and Charles H. Reynolds. Cambridge and New York: Cambridge University Press. Jun. 2012.

[3] J. Řezáč and P. Hobza, "Benchmark Calculations of Interaction Energies in Noncovalent Complexes and Their Applications," *Chemical Reviews*, vol. 116, no. 9. American Chemical Society, pp. 5038–5071, May 11, 2016, doi: 10.1021/acs.chemrev.5b00526.

[4] I. N. Levine, *Quantum Chemistry Pearson advanced chemistry series*. 2014.

[5] E. G. Lewars, *Computational Chemistry*, Second Edi. 2011.

[6] M. J. S. Dewar, E. G. Zoebisch, E. F. Healy, and J. J. P. Stewart, "AM1: A New General Purpose Quantum Mechanical Molecular Model1," *J. Am. Chem. Soc.*, vol. 107, no. 13, pp. 3902–3909, 1985, doi: 10.1021/ja00299a024.

[7] J. J. P. Stewart, "Optimization of parameters for semiempirical methods V: Modification of NDDO approximations and application to 70 elements," *J. Mol. Model.*, vol. 13, no. 12, pp. 1173–1213, Dec. 2007, doi: 10.1007/s00894-007-0233-4.

[8] J. J. P. Stewart, "Optimization of parameters for semiempirical methods VI: More modifications to the NDDO approximations and re-optimization of parameters," *J. Mol. Model.*, vol. 19, no. 1, pp. 1–32, Jan. 2013, doi: 10.1007/s00894-012-1667-x.

[9] M. Gaus, A. Goez, and M. Elstner, "Parametrization and Benchmark of

DFTB3 for Organic Molecules," *J. Chem. Theory Comput.*, vol. 9, no. 1, pp. 338–354, Jan. 2013, doi: 10.1021/ct300849w.

[10] A. S. Christensen, T. Kubař, Q. Cui, and M. Elstner, "Semiempirical Quantum Mechanical Methods for Noncovalent Interactions for Chemical and Biochemical Applications," *Chemical Reviews*, vol. 116, no. 9. American Chemical Society, pp. 5301–5337, May 11, 2016, doi: 10.1021/acs.chemrev.5b00584.

[11] J. Tomasi, B. Mennucci, and R. Cammi, "Quantum mechanical continuum solvation models," *Chemical Reviews*, vol. 105, no. 8. American Chemical Society, pp. 2999–3093, 2005, doi: 10.1021/cr9904009.

[12] A. Klamt and G. Schüürmann, "COSMO: a new approach to dielectric screening in solvents with explicit expressions for the screening energy and its gradient," *J. Chem. Soc., Perkin Trans. 2*, no. 5, pp. 799–805, 1993, doi: 10.1039/P29930000799.

[13] A. V. Marenich, C. J. Cramer, and D. G. Truhlar, "Universal solvation model based on solute electron density and on a continuum model of the solvent defined by the bulk dielectric constant and atomic surface tensions," *J. Phys. Chem. B*, vol. 113, no. 18, pp. 6378–6396, May 2009, doi: 10.1021/jp810292n.

[14] C. J. Cramer, *Essentials of Computational Chemistry second edition*, Second Edi. 2004.

[15] E. Brunk and U. Rothlisberger, "Mixed Quantum Mechanical/Molecular Mechanical Molecular Dynamics Simulations of Biological Systems in Ground and Electronically Excited States," *Chemical Reviews*, vol. 115, no. 12. American Chemical Society, pp. 6217–6263, Jun. 24, 2015, doi: 10.1021/cr500628b.

[16] D. A. Case *et al.*, "The Amber biomolecular simulation programs," *Journal of Computational Chemistry*, vol. 26, no. 16. NIH Public Access, pp. 1668–1688, Dec. 2005, doi: 10.1002/jcc.20290.

[17] B. R. Brooks *et al.*, "CHARMM: The biomolecular simulation program," *J. Comput. Chem.*, vol. 30, no. 10, pp. 1545–1614, Jul. 2009, doi: 10.1002/jcc.21287.

[18] W. F. Van Gunsteren and H. J. C. Berendsen, "The GROMOS Software for (Bio)Molecular Simulation GROMOS87 Groningen Molecular Simulation (GROMOS) Library Manual."

[19] W. L. Jorgensen and J. Tirado-Rives, "The OPLS Potential Functions for Proteins. Energy Minimizations for Crystals of Cyclic Peptides and Crambin," *J. Am. Chem. Soc.*, vol. 110, no. 6, pp. 1657–1666, 1988, doi: 10.1021/ja00214a001.

[20] W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, and M. L. Klein, "Comparison of simple potential functions for simulating liquid water," *J. Chem. Phys.*, vol. 79, no. 2, pp. 926–935, Aug. 1983, doi: 10.1063/1.445869.

[21] H. J. C. Berendsen, J. R. Grigera, and T. P. Straatsma, "The missing term in effective pair potentials," *J. Phys. Chem.*, vol. 91, no. 24, pp. 6269–6271, 1987, doi: 10.1021/j100308a038.

[22] A. Onufriev, D. A. Case, and D. Bashford, "Effective Born radii in the generalized Born approximation: The importance of being perfect," *J. Comput. Chem.*, vol. 23, no. 14, pp. 1297–1304, Nov. 2002, doi: 10.1002/jcc.10126.

[23] F. Fogolari, A. Brigo, and H. Molinari, "The Poisson-Boltzmann equation for biomolecular electrostatics: A tool for structural biology," *Journal of*

Molecular Recognition, vol. 15, no. 6. John Wiley & Sons, Ltd, pp. 377–392, Nov. 01, 2002, doi: 10.1002/jmr.577.

[24] E. V. Anslyn and D. A. Dougherty, *Modern Physical Organic Chemistry*. 2006.

[25] E. Arunan *et al.*, "Definition of the hydrogen bond (IUPAC Recommendations 2011)," *Pure Appl. Chem.*, vol. 83, no. 8, pp. 1637–1641, Jul. 2011, doi: 10.1351/PAC-REC-10-01-02.

[26] T. Steiner, "Weak Hydrogen Bonds," in *Implications of Molecular and Materials Structure for New Technologies*, Springer Netherlands, 1999, pp. 185–196.

[27] D. A. Dixon, K. D. Dobbs, and J. J. Valentini, "Amide-water and amide-amide hydrogen bond strengths," *J. Phys. Chem.*, vol. 98, no. 51, pp. 13435–13439, 1994, doi: 10.1021/j100102a001.

[28] J. Emsley, "Very strong hydrogen bonding," *Chemical Society Reviews*, vol. 9, no. 1. The Royal Society of Chemistry, pp. 91–124, Jan. 01, 1980, doi: 10.1039/CS9800900091.

[29] V. I. Minkin, "Glossary of terms used in theoretical organic chemistry (IUPAC Recommendations 1999)," *Pure Appl. Chem.*, vol. 71, no. 10, pp. 1919–1981, 1999, doi: 10.1351/pac199971101919.

[30] M. Kolář and P. Hobza, "On extension of the current biomolecular empirical force field for the description of halogen bonds," *J. Chem. Theory Comput.*, vol. 8, no. 4, pp. 1325–1333, Apr. 2012, doi: 10.1021/ct2008389.

[31] B. Fenderson, *Molecular Biology of the Cell*,*5th Edition*, 2008.

[32] "CCCBDB introduction navigation." https://cccbdb.nist.gov/ (accessed May 14, 2021).

[33] "NCIAtlas." http://www.nciatlas.org/ (accessed May 14, 2021).

[34] L. P. Yang *et al.*, "A supramolecular system that strictly follows the binding mechanism of conformational selection," *Nat. Commun.*, vol. 11, no. 1, pp. 1–9, Dec. 2020, doi: 10.1038/s41467-020-16534-9.

[35] S. M. Eyrilmez *et al.*, "Binary twinned-icosahedral [B21H18]- interacts with cyclodextrins as a precedent for its complexation with other organic motifs," *Phys. Chem. Chem. Phys.*, vol. 19, no. 19, pp. 11748–11752, May 2017, doi: 10.1039/c7cp01074e.

[36] D. Sigwalt *et al.*, "Unraveling the Structure-Affinity Relationship between Cucurbit[n]urils (n = 7, 8) and Cationic Diamondoids," *J. Am. Chem. Soc.*, vol. 139, no. 8, pp. 3249–3258, Mar. 2017, doi: 10.1021/jacs.7b00056.

[37] D. L. Nelson and M. M. Cox, *Lehninger Principles of Biochemistry*. 2021.

[38] E. P. Kyba *et al.*, "Host-Guest Complexation. 1. Concept and Illustration," *J. Am. Chem. Soc.*, vol. 99, no. 8, pp. 2564–2571, 1977, doi: 10.1021/ja00450a026.

[39] P. Xu *et al.*, "Computation of host-guest binding free energies with a new quantum mechanics based mining minima algorithm," *J. Chem. Phys.*, vol. 154, no. 10, p. 104122, Mar. 2021, doi: 10.1063/5.0040759.

[40] D. Das, K. I. Assaf, and W. M. Nau, "Applications of Cucurbiturils in Medicinal Chemistry and Chemical Biology," *Frontiers in Chemistry*, vol. 7. Frontiers Media S.A., p. 619, Sep. 13, 2019, doi: 10.3389/fchem.2019.00619.

[41] H. Bai *et al.*, "Cyclodextrin-based host-guest complexes loaded with regorafenib for colorectal cancer treatment," *Nat. Commun.*, vol. 12, no. 1, pp. 1–18, Dec. 2021, doi: 10.1038/s41467-021-21071-0.

[42] H. J. Smith and H. J. Williams, "Textbook of Drug Design and Discovery," *Textb. Drug Des. Discov.*, 2016, doi: 10.1201/b12381.

[43] M. Zheng, X. Liu, Y. Xu, H. Li, C. Luo, and H. Jiang, "Computational methods for drug design and discovery: Focus on China," *Trends in Pharmacological Sciences*, vol. 34, no. 10. Elsevier, pp. 549–559, Oct. 2013, doi: 10.1016/j.tips.2013.08.004.

[44] "Directory of in silico Drug Design tools." https://click2drug.org/ (accessed May 14, 2021).

[45] "AutoDock — AutoDock." http://autodock.scripps.edu/ (accessed May 14, 2021).

[46] "UCSF DOCK." http://dock.compbio.ucsf.edu/ (accessed May 14, 2021).

[47] "GOLD - Protein Ligand Docking Software - The Cambridge Crystallographic Data Centre (CCDC)." https://www.ccdc.cam.ac.uk/solutions/csd-discovery/components/gold/ (accessed May 14, 2021).

[48] "Glide | Schrödinger." https://www.schrodinger.com/products/glide (accessed May 14, 2021).

[49] O. Trott and A. J. Olson, "AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading," *J. Comput. Chem.*, vol. 31, no. 2, p. NA-NA, Jan. 2009, doi: 10.1002/jcc.21334.

[50] D. R. Koes, M. P. Baumgartner, and C. J. Camacho, "Lessons Learned in Empirical Scoring with smina from the CSAR 2011 Benchmarking Exercise," *J. Chem. Inf. Model.*, vol. 53, no. 8, pp. 1893–1904, Aug. 2013, doi: 10.1021/ci300604z.

[51] O. Korb, T. Stützle, and T. E. Exner, "Empirical scoring functions for advanced Protein-Ligand docking with PLANTS," *J. Chem. Inf. Model.*, 2009, doi: 10.1021/ci800298z.

[52] E. H. B. Maia, L. C. Assis, T. A. de Oliveira, A. M. da Silva, and A. G. Taranto, "Structure-Based Virtual Screening: From Classical to Artificial Intelligence," *Frontiers in Chemistry*, vol. 8. Frontiers Media S.A., p. 343, Apr. 28, 2020, doi: 10.3389/fchem.2020.00343.

[53] S. Y. Huang, S. Z. Grinter, and X. Zou, "Scoring functions and their evaluation methods for protein-ligand docking: Recent advances and future directions," *Phys. Chem. Chem. Phys.*, vol. 12, no. 40, pp. 12899–12908, Oct. 2010, doi: 10.1039/c0cp00151a.

[54] T. Ten Brink and T. E. Exner, "Influence of protonation, tautomeric, and stereoisomeric states on protein-ligand docking results," *J. Chem. Inf. Model.*, vol. 49, no. 6, pp. 1535–1546, Jun. 2009, doi: 10.1021/ci800420z.

[55] R. A. Friesner *et al.*, "Glide: A New Approach for Rapid, Accurate Docking and Scoring. 1. Method and Assessment of Docking Accuracy," *J. Med. Chem.*, vol. 47, no. 7, pp. 1739–1749, Mar. 2004, doi: 10.1021/jm0306430.

[56] W. J. Allen *et al.*, "DOCK 6: Impact of new features and current docking performance," *J. Comput. Chem.*, vol. 36, no. 15, pp. 1132–1156, Jun. 2015, doi: 10.1002/jcc.23905.

[57] J. Dittrich, D. Schmidt, C. Pfleger, and H. Gohlke, "Converging a Knowledge-Based Scoring Function: DrugScore 2018," *J. Chem. Inf. Model.*, vol. 59, no. 1, pp. 509–521, Jan. 2019, doi: 10.1021/acs.jcim.8b00582.

[58] G. Jones, P. Willett, R. C. Glen, A. R. Leach, and R. Taylor, "Development and validation of a genetic algorithm for flexible docking," *J. Mol. Biol.*, vol. 267, no. 3, pp. 727–748, 1997, doi: 10.1006/jmbi.1996.0897.

[59] G. M. Morris *et al.*, "AutoDock Version 4.2 Automated Docking of Flexible Ligands to Flexible Receptors," 2014, Accessed: Jun. 26, 2017. [Online]. Available: http://autodock.scripps.edu/faqs-help/manual/autodock-4-2-user-

guide/AutoDock4.2.6_UserGuide.pdf.

[60] M. Xu and M. A. Lill, "Induced fit docking, and the use of QM/MM methods in docking," *Drug Discovery Today: Technologies*, vol. 10, no. 3. Elsevier, pp. e411–e418, Sep. 01, 2013, doi: 10.1016/j.ddtec.2013.02.003.

[61] K. Raha and K. M. Merz, "A Quantum Mechanics-Based Scoring Function: Study of Zinc Ion-Mediated Ligand Binding," *J. Am. Chem. Soc.*, vol. 126, no. 4, pp. 1020–1021, Feb. 2004, doi: 10.1021/ja038496i.

[62] K. Raha and K. M. Merz, "Large-Scale validation of a quantum mechanics based scoring function: Predicting the binding affinity and the binding mode of a diverse set of protein-ligand complexes," *J. Med. Chem.*, vol. 48, no. 14, pp. 4558–4575, Jul. 2005, doi: 10.1021/jm048973n.

[63] P. Mikulskis, S. Genheden, K. Wichmann, and U. Ryde, "A semiempirical approach to ligand-binding affinities: Dependence on the hamiltonian and corrections," *J. Comput. Chem.*, vol. 33, no. 12, pp. 1179–1189, May 2012, doi: 10.1002/jcc.22949.

[64] H. S. Muddana and M. K. Gilson, "Calculation of host-guest binding affinities using a quantum-mechanical energy model," *J. Chem. Theory Comput.*, vol. 8, no. 6, pp. 2023–2033, Jun. 2012, doi: 10.1021/ct3002738.

[65] M. Lepšík, J. Řezáč, M. Kolář, A. Pecina, P. Hobza, and J. Fanfrlík, "The Semiempirical Quantum Mechanical Scoring Function for In Silico Drug Design," *Chempluschem*, vol. 78, no. 9, pp. 921–931, Sep. 2013, doi: 10.1002/cplu.201300199.

[66] A. Pecina *et al.*, "SQM/COSMO Scoring Function: Reliable Quantum-Mechanical Tool for Sampling and Ranking in Structure-Based Drug Design," *Chempluschem*, vol. 85, no. 11, pp. 2362–2371, Nov. 2020, doi: 10.1002/cplu.202000120.

[67] S. Liu *et al.*, “Practical Model Selection for Prospective Virtual Screening,” *Journal of Chemical Information and Modeling*, vol. 59, no. 1. American Chemical Society, pp. 282–293, Jan. 28, 2019, doi: 10.1021/acs.jcim.8b00363.

List of Publications

Included in the Thesis

A. Eyrilmez, S. M., Bernhardt, E., Dávalos, J. Z., Lepšík, M., Hobza, P., Assaf, K. I., Nau, W. M., Holub, J., Oliva-Enrich, J. M., Fanfrlík, J., & Hnyk, D. (2017). Binary twinned-icosahedral $[B_{21}H_{18}]^-$ interacts with cyclodextrins as a precedent for its complexation with other organic motifs. *Physical Chemistry Chemical Physics, 19*(19), 11748–11752.

B. Ajani, H., Pecina, A., Eyrilmez, S. M., Fanfrlík, J., Haldar, S., Řezáč, J., Hobza, P., & Lepšík, M. (2017). Superior Performance of the SQM/COSMO Scoring Functions in Native Pose Recognition of Diverse Protein-Ligand Complexes in Cognate Docking. *ACS Omega*, *2*(7), 4022–4029.

C. Pecina, A., Brynda, J., Vrzal, L., Gnanasekaran, R., Hořejší, M., Eyrilmez, S. M., Řezáč, J., Lepšík, M., Řezáčová, P., Hobza, P., Majer, P., Veverka, V., & Fanfrlík, J. (2018). Ranking Power of the SQM/COSMO Scoring Function on Carbonic Anhydrase II-Inhibitor Complexes. *ChemPhysChem*, *19*(7), 873–879.

D. Eyrilmez, S. M., Köprülüoğlu, C., Řezáč, J., & Hobza, P. (2019). Impressive Enrichment of Semiempirical Quantum Mechanics-Based Scoring Function: HSP90 Protein with 4541 Inhibitors and Decoys. *ChemPhysChem*, *20*(21), 2759–2766.

Not Included in the Thesis

1. Sedlak, R., Eyrilmez, S. M., Hobza, P., & Nachtigallova, D. (2018). The role of the σ-holes in stability of non-bonded chalcogenide···benzene interactions: the ground and excited states. *Physical Chemistry Chemical Physics*, *20*(1), 299–306.
2. Honda, D. E., Martins, J. B. L., Ventura, M. M., Eyrilmez, S. M., Lepšík, M., Hobza, P., Pecina, A., & de Freitas, S. M. (2018). Interface Interactions of the Bowman-Birk Inhibitor BTCI in a Ternary Complex with Trypsin and Chymotrypsin Evaluated by Semiempirical Quantum Mechanical Calculations. *European Journal of Organic Chemistry*, *2018*(37), 5203–5211.
3. Pecina, A., Eyrilmez, S. M., Köprülüoğlu, C., Miriyala, V. M., Lepšík, M., Fanfrlík, J., Řezáč, J., & Hobza, P. (2020). SQM/COSMO Scoring Function: Reliable Quantum-Mechanical Tool for Sampling and Ranking in Structure-Based Drug Design. *ChemPlusChem*, *85*(11), 2362–2371.
4. Hajduch, J., Fabre, B., Klopp, B., Pohl, R., Buděšínský, M., Šolínová, V., Kašička, V., Köprülüoglu, C., Eyrilmez, S. M., Lepšík, M., Hobza, P., Mitrová, K., Lubos, M., Hernández, M. S. G., & Jiráček, J. (2021). Multipodal insulin mimetics built on adamantane or proline scaffolds. *Bioorganic Chemistry*, *107*, 104548.

Presentation of the Results

3rd Users' Conference of IT4Innovations (05-06 November 2019), Ostrava, Czech Republic, **Poster Presentation:** "On the Performance of Semiempirical Quantum Mechanics-Based Scoring Functions for Virtual Screening Applications"

(12th AFMC / AIMECS 2019) Asian Federation for Medicinal Chemistry (AFMC) 12th International Symposium "New Avenues for Design and Development of Translational Medicine" (08-11 September 2019), Istanbul, Turkey, **Oral Presentation:** "On the Performance of Semiempirical Quantum Mechanics-Based Scoring Functions for Virtual Screening Applications"

(Drug Discovery-2018) International Conference on Drug Discovery, Development and Lead Optimization (3-5 December 2018), San Francisco, USA, **Poster Presentation:** *"Quantum Mechanical Investigation of Non-Covalent Interactions in Protein-Ligand Complexes"*

(ECHEMINFO 2018) Training and Innovation Course in Drug Design (16 – 20 July 2018), Milano, ITALY, **Poster Presentation:** *"Quantum Mechanical Investigation of Non-Covalent Interactions In Protein-Ligand Complexes"*

(WATOC) 11th Triennial Congress of the World Association of Theoretical and Computational Chemistry (27 August – 1 September 2017), Munich, GERMANY, **Poster Presentation:** *"Interactions of binary twinned-icosahedral $[B_{21}H_{18}]^-$ with Cyclodextrins"*

(ICAMM) International Conference on Advanced Materials Modelling, Rennes, FRANCE (September 5-7, 2016), **Poster Presentation:** *"Adsorption of Small organic molecules on graphene-based surface"*

Attached Publications

Publication A

PCCP

COMMUNICATION

View Article Online
View Journal | View Issue

Cite this: *Phys. Chem. Chem. Phys.*, 2017, **19**, 11748

Received 17th February 2017,
Accepted 19th April 2017

DOI: 10.1039/c7cp01074e

rsc.li/pccp

Binary twinned-icosahedral $[B_{21}H_{18}]^-$ interacts with cyclodextrins as a precedent for its complexation with other organic motifs†

Saltuk M. Eyrilmez,[ab] Eduard Bernhardt,[c] Juan Z. Dávalos,[d] Martin Lepšík,[a] Pavel Hobza,[ae] Khaleel I. Assaf,[f] Werner M. Nau,[f] Josef Holub,[g] Josep M. Oliva-Enrich,*[d] Jindřich Fanfrlík*[a] and Drahomír Hnyk*[g]

The weakly coordinating binary macropolyhedral anion *closo,closo*-$[B_{21}H_{18}]^-$ (B21; D_{3h} symmetry) has been synthesized using a simplified strategy compared to that in the literature. While gas-phase complexes of B21 with β- and γ-cyclodextrin (CD) were detected using ESI FT-ICR spectrometric measurements, α-CD did not bind to the B21 guest. This spectroscopic evidence has been interpreted using quantum-chemical computations, showing that β- and γ-CD are able to interact with B21 due to their larger cavities, in contrast to the smaller α-CD. The hydridic B–H vectors of the B21 anion interact with K^+ counterions and, *via* dihydrogen bonding, also with the partially positively charged hydrogens of the CD sugar units in the modeled β- and γ-CD complexes. In summary, it has been shown by combined spectrometric/computational analysis that macropolyhedral boron hydride anions with two counterions can form stable complexes with β- and γ-CD in the gas phase, offering a new perspective for the future investigation of this remarkable anion in the areas of supramolecular and medicinal chemistries.

Introduction

The icosahedron is the most symmetrical way to arrange twelve atoms into a polyhedral cluster. It is the quintessential structural motif in boron hydride cluster chemistry, represented by the *closo*-$[B_{12}H_{12}]^{2-}$ dianion, which has I_h point-group symmetry. While parent boron hydrides have a tendency to fuse together, this happens not through a single shared boron atom, but rather requires at least one joint B–B vector.[1] In the case of the joining of two *closo*-$[B_{12}H_{12}]^{2-}$ cages, three vertices need to be shared.[2] On that basis, *closo,closo*-$[B_{21}H_{18}]^-$ (abbreviated as **B21** in this study) is formed by the oxidative coupling of two *closo*-$[B_{10}H_{10}]^{2-}$ clusters. The resulting *closo,closo*-$[B_{20}H_{18}]^{2-}$ macropolyhderal anion is isomerized, which is followed by the insertion of an additional boron vertex by heating with $BH_3{\cdot}NEt_3$.[3] The **B21** anion adopts overall D_{3h} symmetry, indicative of four symmetrically unique boron environments instead of one in *closo*-$[B_{12}H_{12}]^{2-}$ (Fig. 1).

Boron clusters form a number of unique types of noncovalent interactions,[4] of which dihydrogen bonding[5] and B–H···cation interactions are important for this study. Both interactions are based on the fact that boron-bound hydrogens are slightly negatively charged due to the lower electronegativity of boron as compared to hydrogen. This is evident from the calculated

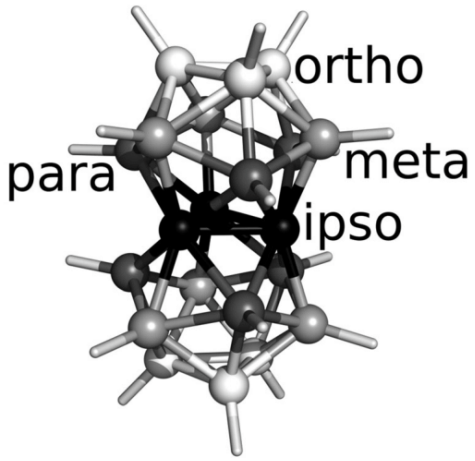


Fig. 1 A molecular diagram of *closo,closo*-$[B_{21}H_{18}]^-$ with D_{3h} symmetry that distinguishes between individual types of boron atoms.

[a] *Institute of Organic Chemistry and Biochemistry of the Czech Academy of Sciences, Flemingovo nam. 2, CZ-16610 Prague 6, Czech Republic. E-mail: fanfrlik@uochb.cas.cz*
[b] *Department of Physical Chemistry, Palacký University, CZ–77146 Olomouc, Czech Republic*
[c] *Bergische University Wuppertal, Gaussstrasse 20, D-42097 Wuppertal, Germany*
[d] *Instituto de Química-Física "Rocasolano", CSIC, ES-28006 Madrid, Spain. E-mail: j.m.oliva@iqfr.csic.es*
[e] *Regional Center of Advanced Technologies and Materials, Department of Physical Chemistry, Palacký University, CZ–77146 Olomouc, Czech Republic*
[f] *Department of Life Sciences and Chemistry, Jacobs University Bremen, Campus Ring 1, D-28759 Bremen, Germany*
[g] *Institute of Inorganic Chemistry of the Czech Academy of Sciences, v.v.i. CZ-25068 Husinec-Řež, Czech Republic. E-mail: hnyk@iic.cas.cz*

† Electronic supplementary information (ESI) available: Experimental and computational details, ESI FT-ICR spectrum of $KB_{21}H_{18}$ (Fig. S1) and binary complexes within overall $KB_{21}H_{18}$ + CD supramolecular complexes (Fig. S2–S5) as well as a table that summarizes interaction energies (Table S1). See DOI: 10.1039/c7cp01074e

electrostatic potential (ESP) or partial atomic charges obtained by the restrained fit to the electrostatic potential (RESP) methodology.[6] These two types of interaction (B–H···cation interactions[7] and dihydrogen bonding[8]) have been found to be crucial for the binding of boron-cage-containing inhibitors to protein receptors. Host–guest chemistry presents a broad field of supramolecular chemistry, that is based on the specific non-covalent recognition of inorganic ions or small-molecule organic guests by macrocyclic organic hosts. Typically, cationic or neutral guests are encapsulated into the cavity of neutral macrocyclic hosts. Cyclodextrin (CD) molecules, very well-known macrocyclic hosts, have three major forms differing in the number of glucose ring molecules: α-CD contains six, β-CD seven and γ-CD eight units. CDs are able to encapsulate in their cavities a wide range of hydrophobic organic guests; in contrast, only a few heteroborane-based guests have been reported.[9]

The complexation of boron cluster anions with hosts has been observed in solution in several examples. In each of the known complexes the anions have adopted the icosahedral structural motif.[10] To our knowledge, reported gas-phase complexes with the same cage architecture are exceptional.[11] The study mentioned in ref. 11*a* reports very strong intrinsic intermolecular interactions of *closo*-$[B_{12}X_{12}]^{2-}$ (X = H, F, Cl, Br and I) with several neutral organic receptors, where these dianionic halogenated *closo*-dodecaborates displayed selectivity for the large hosts with deep hydrophobic polarizable pockets, such as in the case of tetrathiafulvalene-based hosts or spherical cavities in the case of CDs. It is the *closo*-$[B_{12}F_{12}]^{2-}$ anion that strongly interacts with β-CD as reported in ref. 11*a*. The formation of these charged complexes was proven by means of electrospray ionization mass spectrometry (ESI-MS), which is a powerful tool to study the stoichiometry and interactions of supramolecular assemblies in the gas phase.[12–16] Postulated weak gas-phase basicities (GB) of these dianions served as an alternative explanation for the stability of these gas-phase complexes.

It is important to mention that (also due to its complicated synthesis[3]) no parent *macropolyhedral* borate has been found to interact with any organic molecule. We have therefore undertaken investigations aimed at testing the possibilities of the mutual interaction of purely organic and purely inorganic systems in the gas phase, the inorganic species being a unique joint-icosahedral boron hydride. The results are important for the understanding of macropolyhedral boron cluster affinity since this cluster is relatively inert to conventional substitution reactions, and because its structure differs from its geometrical building block, the *closo*-$[B_{12}H_{12}]^{2-}$ dianion.

Results and discussion

Simplified synthesis of B21

We based our synthesis on the synthetic procedure of **B21** reported in ref. 3. However, we have improved one step in this reaction pathway; namely the rearrangement of *trans*-$[B_{20}H_{18}]^{2-}$ upon protonation in anhydrous HF, which provides the face-shared *fac*-$[B_{20}H_{18}]^{2-}$. In order to avoid this time-consuming operation, we have proposed a simple step based on the reaction of the triethylammonium salt of *trans*-$[B_{20}H_{18}]^{2-}$ with $BF_3{\cdot}Et_2O$ in the presence of dioxane. Indeed, this yields the *fac*-$[B_{20}H_{18}]^{2-}$ isomer in the form of its trimethyl ammonium salt, which would otherwise be difficult to obtain, in 80% yield based on the starting *trans* isomer.

Mass spectrometry

Although several ESI detection conditions were examined by optimizing the corresponding FT-ICR parameters, the binary (**B21** + CD) complexes were not detected using mass spectrometry. In the negative mode of the ESI FT-ICR spectrum (Fig. S1 (ESI†); *m*/*z* range <300), we found isotopic mass distribution of a very high-intensity peak corresponding to the singly-charged anion **B21.** In the positive mode, *m*/*z* values higher than 1400 (for the β-CD case) or 1600 (for the γ-CD case) were identified, with mono-charged cationic complexes of the $[\beta\text{-CD} + K\mathbf{B21} + K]^+$ and $[\gamma\text{-CD} + K\mathbf{B21} + K]^+$ types being formed. Each of them is depicted with its corresponding isotopic mass distributions in Fig. 2.

Computational section

Electronic properties of B21. The hitherto unknown electronic properties of isolated **B21** were studied initially using QM methods. The computed electrostatic potential (ESP) surface of **B21** indicates that the negative charge is distributed over the whole molecule (see Fig. 3). Consequently, all BH vertices should possess similar

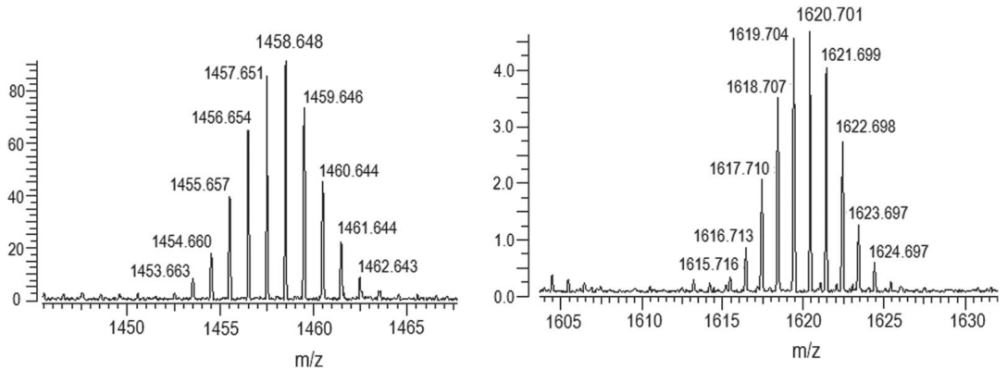


Fig. 2 The ESI FT-ICR spectra, in the positive mode, showing the isotopic mass distribution of cationic complexes formed by K**B21** with β-CD (left, the range of 1445 < *m*/*z* < 1470) and γ-CD (right, the range of 1600 < *m*/*z* < 1635).

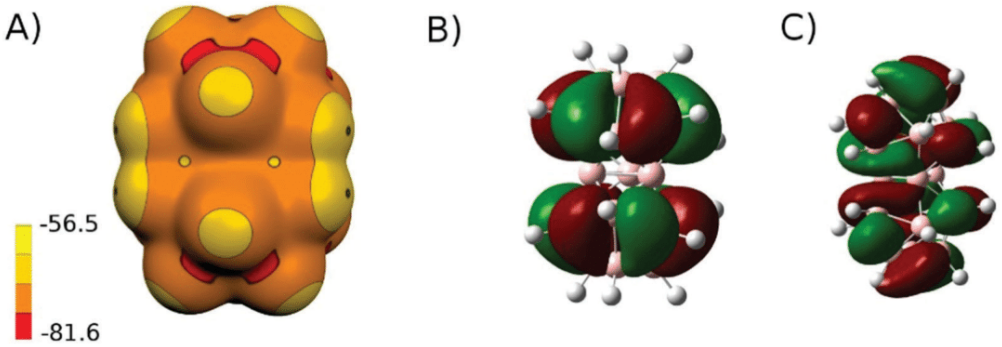


Fig. 3 The electrostatic potential (ESP) surface on 0.001 a.u. computed at the HF/6-31G* level (A). The ESP color range in kcal mol^{-1}. The HOMO (B) and LUMO (C) of **B21** were derived at the same level of theory.

chemical properties. As indicated by the HOMO of **B21**, attack by H^+ can occur close to each of the *ortho*, *meta*, and *para* BH vertices (*cf.* Fig. 1). Indeed, three structures of H**B21** differed in the positions of H^+ in relation to these three kinds of BH vertices; all were quite similar in energy. The *"meta"*-H**B21** isomer was about 2.3 kcal mol^{-1} less stable than *"para"*-H**B21**, *i.e.* its population at 295 K should be below 2%. On the other hand, the energetic difference between *"ortho"*-H**B21** and *"para"*-H**B21** was only 0.6 kcal mol^{-1}, with the structure in which the proton was close to the *para* boron atom computed as the most stable one. This would lead to a mixture containing 75% *"para"*-H**B21**. The structure with the extra H atom bonded to an *ipso*-boron atom was a first-order stationary point. From these calculations the gas-phase basicity[17] (GB) of **B21** was computed to be 233.1 kcal mol^{-1}, a value very close to the experimentally determined GB for histidine (232.9 kcal mol^{-1}).[18] The gas-phase acidity of water is reported to be 158.3 kcal mol^{-1}.[19] In order to compare GB of **B21** with that of other boron clusters forming stable complexes with CDs in the gas phase, we also computed GB values for *closo*-$[B_{12}H_{12}]^{2-}$ and *closo*-$[B_{12}F_{12}]^{2-}$. The obtained GB values were 355.5 and 313.7 kcal mol^{-1}, respectively. **B21** is, therefore, a considerably weaker base in the gas phase than the icosahedral boron clusters. The weak GB enables *closo*-$[B_{12}F_{12}]^{2-}$ to form stable binary dianionic complex with β-CD,[11a] although no structure of this complex has yet been reported.

Complexes. Initial energy scans were performed for the [α-CD + **B21**], [β-CD + **B21**] and [γ-CD + **B21**] binary complexes, revealing that α-CD could not encapsulate **B21** due to the small size of the host (the energy minimum was found at a distance of 5.5 Å; Fig. 4). As a consequence, α-CD was disregarded from further consideration. The [β-CD + **B21**] and [γ-CD + **B21**] complexes exhibited a fully encapsulated minimum (*z*-distances of 2.0 and 1.0 Å, respectively), which is consistent with experimentally observed bound complexes.

In order to understand the complex formation of [β-CD + **B21** + 2K^+] and [γ-CD + **B21** + 2K^+] complexes, we analyzed not only the total interaction energies of the quaternary complexes but also all other possible pairwise interactions that can occur within the studied complexes. The obtained interaction energies are summarized in Table S1 (ESI†).

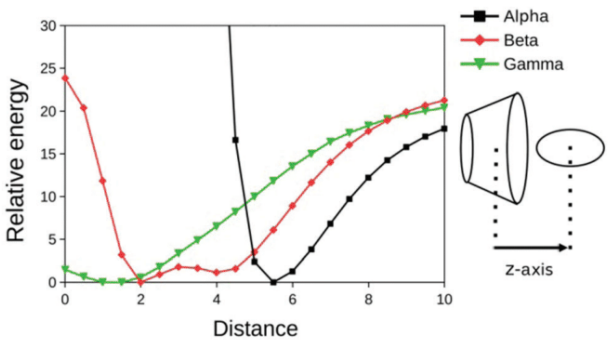


Fig. 4 DFT-D3/TPSS/TZVPP potential energy scans. Relative energy in kcal mol^{-1} and distance in Å.

*Complexes of **B21** with K^+.* The highly symmetrical structure of **B21** resulted in only four binding modes for K^+. Fig. S2 (ESI†) shows the three most favorable positions according to the computations performed. The K^+ ion interacts with four BH vertices of **B21** (two *meta* and two *para*), and the affinity of **B21** to a single K^+ ion is directly proportional to the number of donor hydrogens. When two K^+ ions interact with **B21**, the mutual positions of the K^+ ions are more important than the number of hydrogen donors. The most stable arrangement occurs when the K^+ ions are located on opposite sites, *i.e.* interacting with (a) three *ortho* BH vertices or (b) two *meta* and two *para* BH vertices (see Fig. S3, ESI†).

Complexes of β-, γ-CD with K^+. The most stable binding position of K^+ ion to the host molecules was dictated by the smaller openings of the CD molecules (see Fig. S4, ESI†). The K^+ ion caused significant ring deformations for both β- and γ-CD.

Complexes of β-, γ-CD with B21. The structures obtained show that the chance of the guest molecule penetrating the cavity is proportional to the host molecule size (Fig S5, ESI†). γ-CD with its larger cavity is a more favorable host than β-CD (Fig. S5, ESI†). The interaction energies for the [β-CD + **B21**] and [γ-CD + **B21**] complexes were computed to be −24.8 and −31.0 kcal mol^{-1}, respectively. Although, the interaction energies of the binary complexes are highly negative, and the GB of **B21** is very low, the

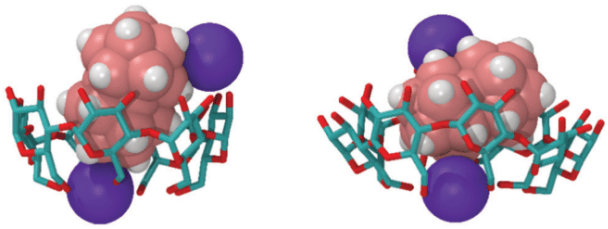


Fig. 5 The most stable computed structures of the (a) [β-CD + **B21** + 2K]$^+$ and (b) [γ-CD + **B21** + 2K]$^+$ complexes.

binary complexes were not experimentally observed under the conditions employed.

*Quaternary complexes of β-, γ-CD with **B21** and 2K$^+$.* The most stable [β-CD + K**B21** + K]$^+$ and [γ-CD + K**B21** + K]$^+$ complexes that are predicted by calculations are shown in Fig. 5. Armed with the knowledge of the pairwise interactions described above, we computed interaction energies according to eqn (1):

$$\Delta E = E_{(\text{total complex})} - E_{(\text{host}+\text{K}^+)} - E_{(\text{B21}+\text{K})} \quad (1)$$

The K$^+$ ions were placed in the small openings of the guest molecules and interacted with both O atoms of the host and H atoms of the guest molecules. They functioned as a bridge and reduced the host deformations.

It is quite apparent that the BH vertices of **B21** are of hydridic nature. The hydrogen atoms of **B21** form short contacts, *i.e.* less than 240 pm (the sum of the van der Waals radii of two hydrogens), with the partially positively charged hydrogens bonded to carbon or oxygen atoms of the sugar units. The [β-CD + **B21** + 2K]$^+$ complex exhibited six (prevailingly *meta* BH) vertices, with the distances ranging from 184 to 219 pm, whereas the [γ-CD + B21 + 2K]$^+$ complex had seven vertices (of all kinds) and the distances ranged from 198 to 237 pm. In both cases the shortest dihydrogen bond was a result of the participation of a polar hydroxyl group. **B21** penetrated the cavity of β-CD almost parallel to the *z* axis (see Fig. 5a) in the [β-CD + **B21** + 2K]$^+$ complex. Furthermore, the conformation of [β-CD + K]$^+$ in the [β-CD + **B21** + 2K]$^+$ complex is 27.2 kcal mol^{-1} less stable than the optimal geometry of isolated [β-CD + K]$^+$, which considerably affects the resulting interaction energy. In the [γ-CD + **B21** + 2K]$^+$ complex, on the other hand, **B21** binds γ-CD in a position perpendicular to the *z* axis (see Fig. 5b). The weaker interactions (*e.g.* longer dihydrogen bonds, see above) in the [γ-CD + **B21** + 2K]$^+$ complex were compensated by the smaller penalty for [γ-CD + K]$^+$ deformation (an energy penalty of 14.3 kcal mol^{-1}). Consequently, the computed total interaction energies of the [β-CD + **B21** + 2K]$^+$ and [γ-CD + **B21** + 2K]$^+$ complexes (as provided by eqn (1), *i.e.* K**B21** with [CD + K]$^+$) were nearly identical (−51.8 and −51.1 kcal mol^{-1}, respectively) despite differences in the **B21** binding modes to β- and γ-CDs in the quaternary complexes. Note also that outer interaction of **B21** with β-CD and γ-CD would have been disfavored since the contact surface area would be considerably reduced.

Conclusions

The synthesis of **B21** has been improved by simplifying the most complicated rearrangement in the synthetic procedure. This allowed **B21** to be synthesized more quickly and in higher yield than previously. **B21** was found to be inert to various attempts to obtain mono-substituted **B21**. With the exception of B_{ipso}–B_{ipso} vectors, all the remaining B–B separations contribute to the LUMO, which also features participation of the terminal hydrogens. It is possible that nucleophilic attacks (*e.g.* with OH$^-$ or halogenide anions) occur at these B–B–H sites and no geometrical preference can be determined from the LUMO. This might account for the fact that all synthetic efforts to prepare mono-substituted **B21** resulted in the mixtures of differently substituted derivatives of **B21**.

Despite the low chemical reactivity of **B21**, we observed gas-phase interactions of **B21** with β- and γ-CD. These interactions were examined by ESI FT–ICR spectrometric measurements. In contrast to both larger CDs, α-CD did not bind the anion, which was explained by its spatial requirements. The structures of both β- and γ-CD complexes were determined using QM calculations. Hydridic B–H vertices of the anion interact both with the partially positively charged hydrogens of the sugar units *via* dihydrogen bonding and with potassium counterions through B–H···cation interactions in the computed structures of the complexes. The observed interactions of the anion under investigation give hope to the tantalizing possibility of promising interactions with biomolecules. Having knowledge of these kinds of interactions is of great importance, in particular with the precedence of the ability of joint icosahedra to inhibit biologically relevant targets.[4]

Acknowledgements

This work was supported by research project RVO 61388963 of the Czech Academy of Sciences. We acknowledge the financial support of the Czech Science Foundation (SME, ML, JF, PH: P208/12/G016 and DH 15-0556775). This work was supported by the Ministry of Education, Youth and Sports from the Large Infrastructures for Research, Experimental Development and Innovations project "IT4Innovations National Supercomputing Center – LM2015070" as well as from project LO1305 (PH).

References

1 Such a common edge is associated with both isomers of $B_{18}H_{22}$ – see *e.g.* M. G. S. Londesborough, D. Hnyk, J. Bould, L. Serrano-Andrés, V. Sauri, J. M. Oliva, P. Kubát, T. Polívka and K. Lang, *Inorg. Chem.*, 2012, **51**, 1471–1479 and the references therein.

2 It is possible to join two $B_{12}H_{12}^{2-}$ icosahedra through four joint vertices like in the case of $B_{20}H_{16}$ – see D. Hnyk, J. Holub, T. Jelínek, J. Macháček and M. G. S. Londesborough, *Collect. Czech. Chem. Commun.*, 2010, **75**, 1115–1123 and the references therein.

3 E. Bernhardt, D. J. Brauer, M. Finze and H. Willner, *Angew. Chem., Int. Ed.*, 2007, **46**, 2927–2930.
4 R. Sedlak, J. Fanfrlík, A. Pecina, D. Hnyk, P. Hobza and M. Lepšík, *Boron – the Fifth Element, Challenges and Advances in Computational Chemistry and Physics*, ed. D. Hnyk and M. McKee, Springer, Heidelberg, New York, Dordrecht and London, 2015, ch. 9, vol. 20 and the references therein.
5 J. Fanfrlík, M. Lepšík, D. Hořínek, Z. Havlas and P. Hobza, *ChemPhysChem*, 2006, **7**, 1100–1105.
6 It was recently shown that the RESP methodology described in C. I. Bayly, P. Cieplak, W. D. Cornell and P. A. Kollman, *J. Phys. Chem.*, 1993, **97**, 10269–10280 represents a method of choice for heteroboranes. This is in contrast with NBO, which closely corresponds to the picture of localized bonds and lone pairs as basic units of molecular structure. This is not true for delocalized heteroboranes. More details can be found in ref. 5.
7 J. Fanfrlík, J. Brynda, J. Řezáč, P. Hobza and M. Lepšík, *J. Phys. Chem. B*, 2008, **112**, 15094–15102.
8 (*a*) M. Kožíšek, P. Cígler, M. Lepšík, J. Fanfrlík, P. Řezáčová, J. Brynda, J. Pokorná, J. Plešek, B. Grüner, K. Grantz Šašková, J. Václavíková, V. Král and J. Konvalinka, *J. Med. Chem.*, 2008, **51**, 4839–4843; (*b*) J. Brynda, P. Mader, V. Šícha, M. Fábry, K. Poncová, M. Bakardiev, B. Grünwe, P. Cígler and P. Řezáčová, *Angew. Chem., Int. Ed.*, 2013, **52**, 13760–13763; (*c*) A. Pecina, M. Lepšík, J. Řezáč, J. Brynda, P. Mader, P. Řezáčová, P. Hobza and J. Fanfrlík, *J. Phys. Chem. B*, 2013, **117**, 16096–16104.
9 K. I. Assaf, D. Gabel, W. Zimmermann and W. M. Nau, *Org. Biomol. Chem.*, 2016, **14**, 7702–7706.
10 (*a*) A family of *closo* dodecaborate anions of the type $[B_{12}X_{12}]^{2-}$ and $[B_{12}X_{11}Y]^{2-}$ (X = H, Cl, Br and I; Y = OH, SH, NH_3^+ and NR_3^+) was tackled, see K. I. Assaf, M. S. Ural, F. Pan, T. Georgiev, S. Simova, K. Rissanen, D. Gabel and W. M. Nau, *Angew. Chem., Int. Ed.*, 2015, **54**, 6852–6856; (*b*) Complexes of a neutral icosahedron with CD are reported in P. Neirynck, J. Schimer, P. Jonkheim, L.-G. Milroy, P. Cigler and L. Brunsveld, *J. Mater. Chem. B*, 2015, **3**, 539–545.
11 (*a*) J. Warneke, C. Jenne, J. Bernarding, V. A. Azov and M. Plaumann, *Chem. Commun.*, 2016, **52**, 6300–6303; (*b*) Interestingly, the same perfluoronated dodecaborate anion interacts in the gas-phase with all *cis* 1,2,3,4,5,6-hexafluorocyclohexane, see M. J. Lecours, R. A. Marta, V. Steinmetz, N. Keddie, E. Fillion, D. O'Hagan, T. B. McMahon and W. Scott Hopkins, *J. Phys. Chem. Lett.*, 2017, **8**, 109–113.
12 C. A. Schalley, *Mass Spectrom. Rev.*, 2001, **20**, 253–309.
13 J. Cernochov, P. Brann, M. Rouchal, P. Kulhánek, I. Kuritka and R. Vícha, *Chem. – Eur. J.*, 2012, **18**, 13633–13637.
14 J. W. Lee, S. W. Heo, S. J. C. Lee, J. Y. Ko, H. Kim and H. I. Kim, *J. Am. Soc. Mass Spectrom.*, 2013, **24**, 21–29.
15 G. Carroy, V. Lemaur, J. De Winter, L. Isaacs, E. De Pauw, J. Cornilc and P. Gerbaux, *Phys. Chem. Chem. Phys.*, 2016, **18**, 12557–12568.
16 T.-C. Lee, E. Kalenius, A. I. Lazar, K. I. Assaf, N. Kuhnert, C. H. Grün, J. Jänis, O. A. Scherman and W. M. Nau, *Nat. Chem.*, 2013, **5**, 376–382.
17 Note that the gas-phase acidity of 1-COOH-1,7-*closo*-$[C_2B_{10}H_{11}]$ is measured to be 316.7 kcal mol^{-1} (315.7 kJ mol^{-1} as computed at B3LYP/6-311++G(d,p)) and reported in J. Z. Dávalos, J. Gonzáles, R. Ramos, D. Hnyk, J. Holub, J. A. Santaballa, M. Canle-L. and J. M. Oliva, *J. Phys. Chem. A*, 2014, **118**, 2788 and the references therein. Note also that closo-$[H(B_{12}F_{12})]^{2-}$ act also as Brønsted acid, which precludes deprotonation of β-CD in the complex and the concept of weak GB as a driving force for the complex stability is justified, see C. Jenne, M. Kessler and J. Warneke, *Chem. – Eur. J.*, 2015, **21**, 5887–5891 and L. Lipping, I. Leito, I. Koppel, I. Krossing, D. Himmel and I. A. Koppel, *J. Phys. Chem. A*, 2015, **119**, 735–743.
18 G. Bouchoux, *Mass Spectrom. Rev.*, 2012, **31**, 391–435.
19 I. Leito, I. A. Koppel, P. Burk, S. Tamp, M. Kutsar, M. Mishima, J.-L. M. Abboud, J. Z. Dávalos, R. Herrero and R. Notario, *J. Phys. Chem. A*, 2010, **114**, 10694–10699.

Publication B

ACS OMEGA

Article

http://pubs.acs.org/journal/acsodf

Superior Performance of the SQM/COSMO Scoring Functions in Native Pose Recognition of Diverse Protein−Ligand Complexes in Cognate Docking

Haresh Ajani,[†,‡] Adam Pecina,[†] Saltuk M. Eyrilmez,[†,‡] Jindřich Fanfrlík,[†] Susanta Haldar,[†] Jan Řezáč,[†] Pavel Hobza,*[,†,§] and Martin Lepšík*[,†]

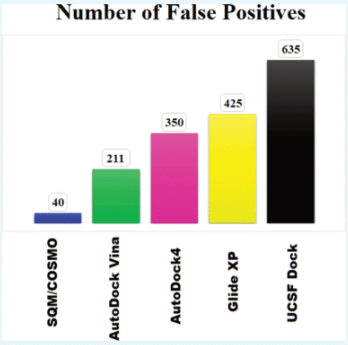
[†]Department of Computational Chemistry, Institute of Organic Chemistry and Biochemistry, Czech Academy of Sciences, v.v.i., Flemingovo nam. 2, 16610 Praha 6, Czech Republic

[‡]Department of Physical Chemistry, Palacký University, tř. 17. listopadu 1192/12, 77146 Olomouc, Czech Republic

[§]Department of Physical Chemistry, Regional Centre of Advanced Technologies and Materials, Palacký University, 77146 Olomouc, Czech Republic

Supporting Information

ABSTRACT: General and reliable description of structures and energetics in protein−ligand (PL) binding using the docking/scoring methodology has until now been elusive. We address this urgent deficiency of scoring functions (SFs) by the systematic development of corrected semiempirical quantum mechanical (SQM) methods, which correctly describe all types of noncovalent interactions and are fast enough to treat systems of thousands of atoms. Two most accurate SQM methods, PM6-D3H4X and SCC-DFTB3-D3H4X, are coupled with the conductor-like screening model (COSMO) implicit solvation model in so-called "SQM/COSMO" SFs and have shown unique recognition of native ligand poses in cognate docking in four challenging PL systems, including metalloprotein. Here, we apply the two SQM/COSMO SFs to 17 diverse PL complexes and compare their performance with four widely used classical SFs (Glide XP, AutoDock4, AutoDock Vina, and UCSF Dock). We observe superior performance of the SQM/COSMO SFs and identify challenging systems. This method, due to its generality, comparability across the chemical space, and lack of need for any system-specific parameters, gives promise of becoming, after comprehensive large-scale testing in the near future, a useful computational tool in structure-based drug design and serving as a reference method for the development of other SFs.



INTRODUCTION

In structure-based drug design, docking/scoring is a prime and well-established computational tool. Molecular docking generates ligand geometries bound to the protein (poses), whereas scoring using scoring functions (SFs) ranks them by the predicted affinity (score). Owing to the approximations embodied in docking/scoring methods for the sake of their acceleration, their accuracy has often been compromised.[1] Nevertheless, recent methodological advances made docking/scoring methods an indispensable tool in discovering new protein ligands.[2]

The "docking power" or "sampling power"[3,4] of a docking/scoring method is assessed by its ability to identify the native ligand pose (root-mean-square deviation (RMSD) from the crystal pose <2 Å) in protein−ligand (PL) complexes. Comprehensive testing across diverse PL complexes has shown that in up to 80% of PL complexes this task can be accomplished.[4−8] However, classical SFs had troubles with the identification of the native binding mode as the best-scoring pose (especially in the case of metalloproteins, halogenated ligands, inorganic ligands, etc.).[4] Thus, reliable identification of native PL poses within a diverse set of PL complexes using a single SF remains a challenging task.[3,4,9]

The four major approaches toward scoring are empirical,[10−12] knowledge-based,[8,13,14] statistical/machine learning,[15,16] and physics-based.[17,18] The first three approaches require a training set, and by use of parametrization and statistics, useful models can be obtained.[19] However, because these approaches are dependent on the training set, their predictive power is limited. In contrast, physics-based methods rely on a generally valid description of PL interactions. Traditionally, such approaches were limited to molecular mechanics (MM) methods and simplified variants thereof. Thus, these approaches were inherently limited by the underlying approximations, most importantly the implicit treatment of electrons.

A general solution to the problem of accurately calculating noncovalent interactions in PL systems is the use of quantum mechanics (QM).[20] With QM methods, phenomena of quantum origin, such as charge transfer, are described without further ad

Received: April 24, 2017
Accepted: July 18, 2017
Published: July 27, 2017

DOI: 10.1021/acsomega.7b00503
ACS Omega 2017, 2, 4022−4029

hoc parametrization. This is important for systems involving halogen bonding,[21,22] metalloprotein binding,[23] inorganic ligands,[24−26] or covalent bond formation.[27] But because of the high computational demands, QM calculations of sufficient quality (e.g., DFT-D3 level with a triple-ζ basis set) are limited to a few hundred atoms. This limitation can be overcome by use of fragmentation[28,29] or a QM/MM approach.[30−34] Another route is the use of semiempirical QM (SQM). The first QM-based SF was introduced by the Merz group.[35] They combined the Austin model 1 (AM1) SQM method with empirical dispersion (D) and implicit solvation (Poisson−Boltzmann (PB) model). Its validation on a large dataset of PL complexes showed its superior performance, especially for metalloprotein−ligand complexes.[36] Although this was an important pioneering step, the accuracy of the underlying methods, both for the vacuum part (AM1-D)[37] and for solvation (PB),[38] was not sufficient to yield quantitative results. More recently, Sulimov laboratory used the PM7 SQM method[39] in conjunction with the conductor-like screening model (COSMO) implicit solvent model[40] for identification of native ligand poses of 16 PL complexes in cognate docking.[41,42] They showed superior performance of their SQM/COSMO SFs over force-field-based scoring. We should note here that PM7 results for noncovalent interactions can be slightly improved by using the latest version of empirical corrections to the PM6 SQM method (PM6-D3H4X, see below).[43]

In our laboratory, we have been systematically developing empirical corrections to SQM methods to accurately treat an array of noncovalent interactions.[20] The latest version of empirical corrections for dispersion, hydrogen bonding, and halogen bonding yielded the PM6-D3H4X[37,44] method, which, coupled with the COSMO implicit solvent model,[40] forms the core of our SQM-based SF (eq 1).[45,46]

$$\text{score} = \Delta E_{\text{int}} + \Delta\Delta G_{\text{solv}} + \Delta G'^{\text{w}}{}_{\text{conf}}(P) + \Delta G'^{\text{w}}{}_{\text{conf}}(L) - T\Delta S_{\text{int}} \quad (1)$$

The score (an estimate of the PL binding free energy) is expressed as an unweighted sum of thermodynamic terms. It consists of the gas-phase PL interaction energy (ΔE_{int}), the change in solvation/desolvation free energy upon complex formation ($\Delta\Delta G_{\text{solv}}$), the change in the conformation "free" energies of the protein and ligand [$\Delta G'^{\text{w}}{}_{\text{conf}}(P,L)$], and the interaction entropy change upon binding ($T\Delta S_{\text{int}}$).[45,46] The PL complexes are optimized using the solution-phase SQM method before scoring. The ΔE_{int} term is favorable for complex formation and usually is the largest in magnitude. It can reach a few hundreds of kcal/mol for charged or polar ligands. The $\Delta\Delta G_{\text{solv}}$ term opposes binding and can be nearly as large as the first term. These two dominant terms thus partially compensate for each other, and the final score is an order of magnitude smaller. Using this SF (eq 1), we have rationalized the binding of series of ligands to a dozen of protein targets,[21,22,46−50] including covalent ligand binding.[27] It should be noted that this SF can also be extended to evaluate explicit solvent effects.[47,48,51]

Recently, we have accelerated our SQM-based SF by considering only the first two dominant terms and replacing the time-demanding SQM optimization with a quick MM relaxation of hydrogens.[23] We have shown in four difficult PL complexes that this SQM/COSMO SF at the PM6-D3H4X level outperforms eight widely used SFs in native ligand pose identification in cognate docking. The number of false-positive (FP) solutions (i.e., those poses that scored better than the native one) was up to 1 order of magnitude lower than that for the classical SFs.[23] In three PL cases, it was even 0. In the challenging system of the tumor necrosis factor-α converting enzyme (TACE) metalloprotein featuring Zn^{2+} in the active site, which is bound by the thiolate group of the ligand, 39 FPs were found.[23] A major improvement (FP = 0)[52] was observed when the ΔE_{int} term was calculated with a more robust SQM method, the self-consistent-charge density-functional tight-binding method augmented with empirical dispersion (previously shown to be useful for the description of biomolecules)[53] and hydrogen-bonding corrections (SCC-DFTB3-D3H4, abbreviated DFTB3-D3H4).[54] The high-quality description of the other three PL systems was retained.[52] The price for the improvement was a higher but not unsurmountable computational cost. It should be noted that two recent studies used the uncorrected SCC−DFTB method in a QM/MM setup and reported success on the correct ligand binding geometries toward metalloproteins.[31,32] Their approach toward the computationally expensive task was to use a rather small QM part consisting of Zn^{2+}, its coordinating protein side chains, and the ligand on a large number of PL systems.[31,32]

In this study, we aim to validate our SQM/COSMO SFs[23,52] for native pose identification in cognate docking on a data set consisting of 17 PL complexes from five diverse classes, selected using strict criteria for physics-based scoring. We apply two variants of the SQM/COSMO SF (ΔE_{int} term at the PM6-D3H4X or DFTB3-D3H4X level)[23,52] and compare them with four standard SFs (Glide XP,[55] AutoDock4,[56] AutoDock Vina,[57] and UCSF Dock[58]). The performance criterion is the number of FPs[23,52] with an extended definition presented here. We show here that the unique behavior of the SQM/COSMO SFs observed in our recent studies[23,52] hold across 17 diverse PL complexes and gives promise of generality after comprehensive large-scale testing in the near future.

RESULTS AND DISCUSSION

Data Set. In this work, we extend our previous pilot studies on four difficult PL systems[23,52] to 17 pharmaceutically relevant and diverse PL complexes from five classes, including three enzyme classes (transferase, hydrolase, and lyase), one chaperone, and two nuclear receptor classes from the PBDbind "core set"[59] (for details, see Methods section). We apply the strict criteria needed for physics-based scoring. Specifically, the crystal structures of the complexes have resolutions better than 2.5 Å, well-resolved electron densities for the ligands, and protein active sites. The ligands have variable chemistries, sizes (molecular weight of 305−666 Da), charges, and flexibilities (for details, see Methods section). Their binding constants toward their targets range from micro- to picomolar.

The crystal poses of the ligands were scored as reference. The ligand poses generated previously by docking[4] with seven docking programs (for details, see Methods section) totaled 4566 poses. RMSD-based clustering (see Methods section) of the poses was carried out to avoid pose redundancy. After this, the number of poses decreased to 3328, corresponding to approximately 250 poses per target. A comprehensive evaluation of the recognition of near-native poses requires a balanced distribution of RMSDs of the docked ligand poses with respect to the crystal (native) geometry from very similar to very dissimilar (up to 10 Å). In most of the PL systems, 20−60% of poses had RMSD <2 Å (Figure S1A). Furthermore, poses were evenly distributed in RMSD ranges of 2−5 and 5−10 Å (roughly 20−40% for each category). Figure S1B shows minimal RMSD (RMSD^{min}) for the poses studied. Near-native poses within the

DOI: 10.1021/acsomega.7b00503
ACS Omega 2017, 2, 4022−4029

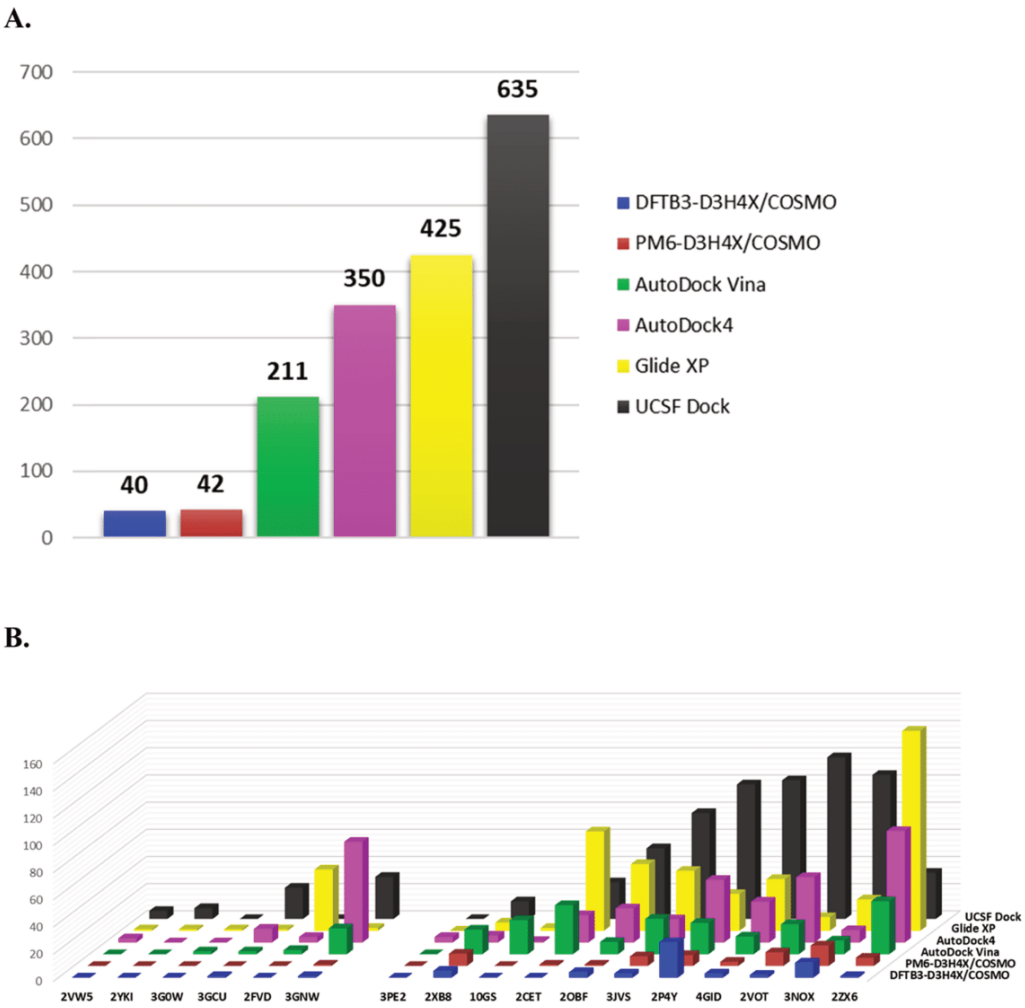


Figure 1. Number of HFP solutions for the six SFs used here across all the 17 PL systems studied. (A) Number of HFPs and (B) HFPs for individual PL complexes sorted by ligand charge: neutral (left) and charged (right).

experimental accuracy of X-ray crystallography of 0.5 Å[60,61] were found in all but two cases (10GS: $RMSD^{min}$ = 0.85 Å and 2VOT: $RMSD^{min}$ = 0.70 Å). However, these cases only slightly exceeded the threshold.

Scoring. For each of the six SFs (two variants of SQM/COSMO SF and four standard SFs; for details, see Methods section), the scores of the docked ligand poses in their respective target proteins were calculated, transformed to relative scores with respect to the score of the crystal pose, and normalized (see Methods section).

The overall sampling power of all the SFs is shown as the enrichment plot (Figure S2), that is, the percent of PL cases (y axis) in which the best-scoring ligand of a given SF has defined RMSD (x axis) to the crystal pose. In the standard range of RMSD up to 2 Å, the SQM/COSMO SFs at the DFTB3-D3H4X level perform the best (88% of PL systems), followed by SQM/COSMO at the PM6-D3H4X level together with UCSF Dock (82% of PL systems). Slightly worse is the performance of Glide XP (76%), followed by AutoDock4 (71%), and AutoDock Vina (65%) (Figure S2). In recognition of near-native poses (RMSD < 0.5 Å), the two SQM/COSMO SFs together with AutoDock Vina perform the best (47%), followed by AutoDock4 and UCSF Dock (41%) and Glide XP, which recognize the poses only in 29% of cases.

The SQM/COSMO SFs also had the lowest number of PL complexes (two for DFTB3-D3H4X/COSMO, Table S1) for which the best-scoring pose exceeded the threshold for success of 2 Å. This was closely followed by SQM/COSMO at the PM6-D3H4X/COSMO level, UCSF Dock, and Glide XP (three cases). Five and six failures were found for Autodock4 and AutoDock Vina, respectively (Table S1). Averaging the RMSDs of the best-scoring poses across all 17 PL complexes (and counting all the failures >2 Å as 2.1 Å), DFTB3-D3H4X/COSMO was the winner (0.71 Å), closely followed by PM6-D3H4X/COSMO and UCSF Dock (0.77 and 0.79 Å, respectively; Table S1). Worse results (around 1 Å) were obtained for AutoDock Vina, AutoDock4, and Glide XP.

For detailed performance evaluations, we use the number of FP solutions criterion[23,52] with an extended definition presented here. Previously, FPs were defined as those poses that scored better than the native pose (defined by a 0.5 Å RMSD cutoff from the crystal pose due to inaccuracies of crystal structures).[23,52] Here, we allow room for larger uncertainties of native pose recognition by defining "hard FPs" (HFP) in which the cutoffs were increased to RMSD >2 Å and score better than −1 kcal/

DOI: 10.1021/acsomega.7b00503
ACS Omega 2017, 2, 4022−4029

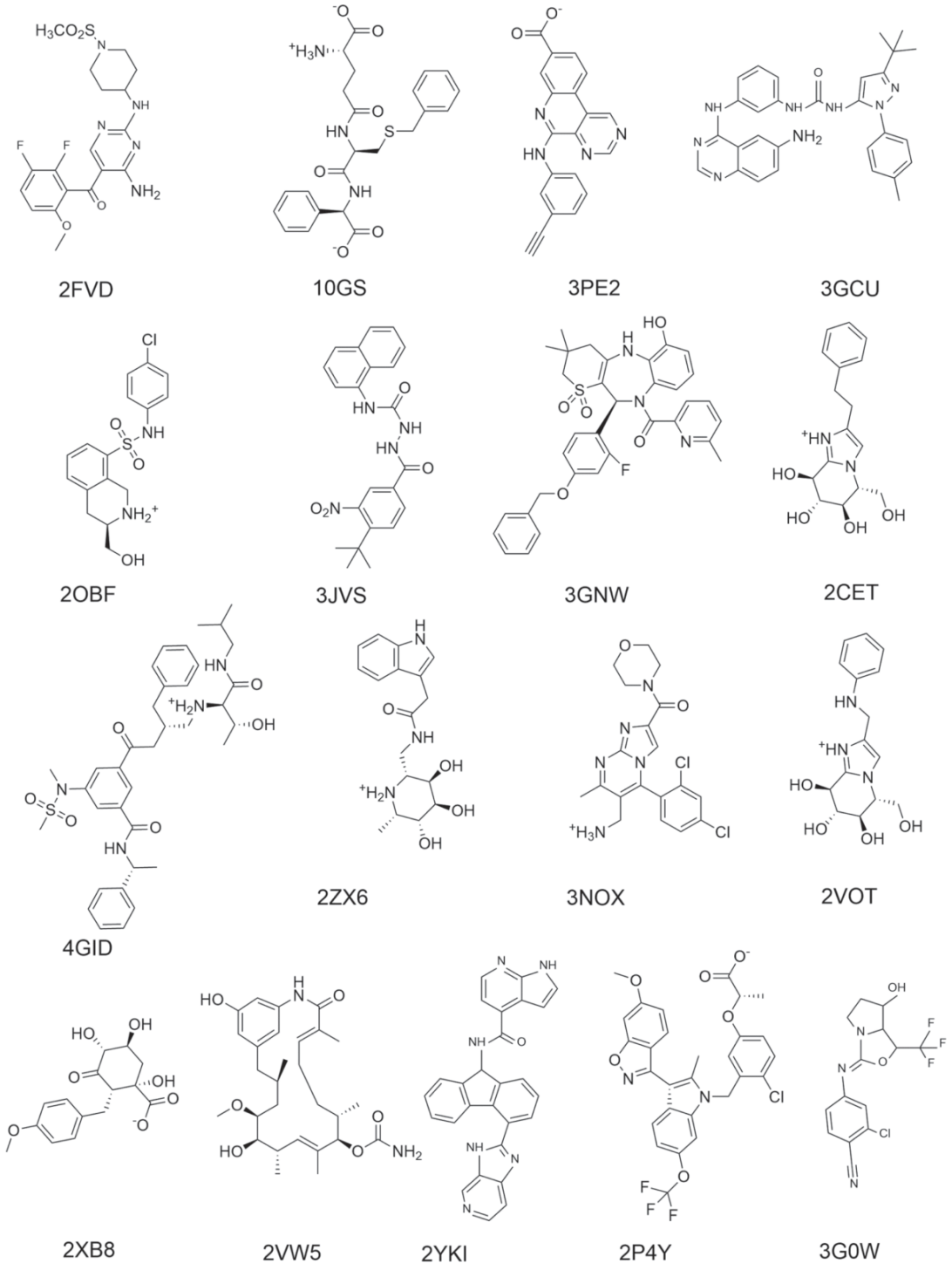


Figure 2. Two-dimensional structures of the ligands studied.

mol. The RMSD cutoff now also includes the effects of flexible parts of the ligands sticking out to the solvent, and the score cutoff corresponds roughly to 2−3 kcal/mol of unscaled energies, which are rough error bounds of the physics-based method. The HFPs for AutoDock Vina, AutoDock4, Glide XP, and UCSF Dock were high—211, 350, 425, and 635, respectively (Figure 1A). The SQM/COSMO SFs performed much better with the numbers of HFPs being up to 1 order of magnitude smaller—40 and 42 for the DFTB3-D3H4X and PM6-D3H4X levels, respectively (Figure 1A).

The number of HFPs for individual PL complexes (Figure 1B and Table S2) differed markedly with respect to the ligand charge: in the case of the neutral ligands (Figure 1B, left), they were by 1 order of magnitude smaller than that for the charged

DOI: 10.1021/acsomega.7b00503
ACS Omega 2017, 2, 4022−4029

Table 1. Summary of the 17 PL Complexes Studied

PDB code	resolution (Å)	protein name	class	ligand charge	rotatable bonds in ligand
2FVD	1.8	CDK2	transferase (E.C.2)	0	6
10GS	2.2	glutathione *S*-transferase		−1	13
3PE2	1.9	casein kinase IIα		−1	4
3GCU	2.1	mitogen-activated protein kinase 14		0	6
2OBF	2.3	phenylethanolamine *N*-methyltransferase		+1	4
3JVS	1.9	checkpoint kinase 1		−1	5
3GNW	2.4	hepatitis C virus NS5B RNA-dependent RNA polymerase		0	5
2CET	1.9	β-glucosidase A	hydrolase (E.C.3)	+1	4
4GID	2.0	β-secretase I		+1	16
2ZX6	2.4	α-L-fucosidase		+1	4
3NOX	2.3	dipeptidyl peptidase 4		+1	3
2VOT	1.9	β-mannosidase		+1	4
2XB8	2.4	3-dehydroquinate dehydratase	lyase (E.C.4)	−1	4
2VW5	1.9	heat shock protein Hsp82	chaperone	0	3
2YKI	1.6	heat shock protein Hsp90-α		0	3
2P4Y	2.2	peroxisome proliferator-activated receptor γ	nuclear receptor	−1	9
3G0W	1.9	androgen receptor		0	2

ones (Figure 1B, right). For SQM/COSMO at the PM6-D3H4X and DFTB3-D3H4X levels, the numbers of HFPs for neutral ligands were single-digit values (1 and 2, respectively). The classical SFs performed worse, with the number of HFPs ranging from 18 to 85 for neutral ligands (Figure 1B, left and Table S2). The complex with the largest number of HFPs was the RNA-dependent RNA polymerase/ligand complex (3GNW) with 71, 28, and 14 HFPs calculated with AutoDock4, UCSF Dock, and AutoDock Vina, respectively. A large number of HFPs (40) was also observed for the cyclin-dependent kinase 2 (CDK2)/ligand complex (2FVD) for Glide XP (Table S2).

The results show that the classical SFs had larger troubles in identifying the native binding poses for charged ligands (for the classical SFs, more than 90% of HFPs were found for charged ligands). The largest number of HFPs (140) was found with Glide XP for the α-L-fucosidase (2ZX6) PL complex, which had a positively charged ligand. For UCSF Dock, four systems, 2P4Y, 4GID, 2VOT, and 3NOX, yielded in total 403 HFPs, which is 70% of HFPs for the charged ligands in that method (577; Table S2). In contrast, the number of HFPs for the charged ligands for the SQM/COSMO was in total 38 and 41 for DFTB3-D3H4X and PM6-D3H4X, respectively. This is considerably lower than the classical SFs (193−577 HFPs) (Table S2). For SQM/COSMO at the DFTB3-D3H4X level, the largest number of HFPs was 20 and 8 for 2P4Y and 3NOX, respectively. Also, PM6-D3H4X/COSMO had some troubles with these systems (5 and 10 HFPs, respectively). In both 2P4Y and 3NOX complexes, the HFP poses have the ligand cores placed at very similar positions as the crystal pose, whereas moieties sticking out to the solvent (the benzisoxazol and morpholino groups, respectively) had fewer noncovalent interactions with the protein. This can be one reason why poses with higher RMSD could score well. Other reasons can be some of the approximations embedded in our protocol for speed, such as the neglected terms in the SQM/COSMO SF (change of conformational energy, entropy) or explicit water molecules, which may need to be included in some PL systems for reliable description of the energetics.[47,51]

■ CONCLUSIONS

The sampling (docking) power, that is, the ability to recognize a ligand native pose in cognate PL docking, of two variants of quantum-mechanics-based SQM/COSMO SFs is tested here on 17 PL systems from five diverse protein families carefully selected for physics-based SFs. For comparison, four standard SFs—Glide XP, AutoDock4, AutoDock Vina, and UCSF Dock, are used. The SQM/COSMO SFs at the PM6-D3H4X and DFTB3-D3H4X levels markedly outperform the standard SFs as judged by the number of HFP poses. The time requirements for the SQM/COSMO SF (Table S3) are higher than those for classical SFs, but given the supercomputer power, thousands of docking poses can be evaluated in a reasonable time. The results of the freely available SQM/COSMO SFs give promise of generality, and after comprehensive large-scale testing in the near future, this method could serve as a useful tool in structure-based drug design and reference for SF development.

■ METHODS

Data Set. QM-based interaction energy calculations require sensible geometries and, therefore, we needed good-quality structures of PL complexes. The crystallographic structures should have fair resolution (<2.5 Å) with fully resolved electron density for the entire ligand and surrounding binding site residues. These criteria are fulfilled by the docking/scoring benchmark set PDBbind core set.[3,59,62] In our study, 17 PL complexes (Figure 2 and Table 1) were used with targets from diverse protein families: three enzyme classes (transferase, hydrolase, and lyase), chaperone, and nuclear receptor (Table 1). The ligand structures are shown in Figure 2.

Docking Poses. Ligand poses obtained by seven commonly used docking programs were collected from previously published work.[4] These programs were AutoDock (version 4.2.6),[56] AutoDock Vina (version 1.1.2),[57] LeDock (version 1.0),[63] UCSF Dock (version 6.7),[58] Glide SP (version 67011),[55] Glide XP (version 67011),[55] and Surflex Dock (version 2.706.13302).[64] For each target, the ligand poses were pooled, which amounted to approximately 350 poses per target. To reduce the redundancy, all poses per target were clustered using the "cluster_conformer" script in the Schrödinger suite[65] with an RMSD cutoff of 0.5 Å. The number of poses was thus reduced to approximately 250 poses per PL system. Each ligand pose, as well as X-ray reference geometry, was scored "in-place" using four classical SFs (AutoDock,[56] AutoDock Vina,[57] UCSF Dock,[58] and Glide XP[55]) and compared to that of two variants of SQM/COSMO SFs, see below.[23,52]

DOI: 10.1021/acsomega.7b00503
ACS Omega 2017, 2, 4022−4029

Protein and Ligand Preparation. *Protein Structure Preparation.* Following the standard virtual screening protocol,[4] all the crystal waters were removed from the PL complexes. As noted previously,[23] physics-based SFs require special care in preparing the PL structures. For all proteins, which were not deposited as monomers, chain A was used for protein preparation except 10GS and 2XB8 where the dimer interface makes important contributions to the binding. We used the LEaP program, which is part of the AMBER14 suite,[66] to protonate the proteins. The protonation state of histidine residues was assigned manually on the basis of hydrogen-bonding patterns. Cysteine disulfide bonds were assigned manually on the basis of the sulfur−sulfur distance. In the case of the 3PE2 complex, the B conformation of M163 was used because it forms interactions with the ligand. Hydrogen atom positions in PL complexes were relaxed by the simulated annealing protocol using short molecular dynamics (MD) (for details see Supporting Information).

Ligand Preparation. The protonation states of the ligands were carefully checked by pK_a calculations at pH 7 using Schrödinger "Propka".[65] The collected docking poses from seven different programs had different output file formats. Each ligand was made into one common MOL2 file format without any changes in *X*, *Y*, and *Z* coordinates. Partial charges were derived at the AM1-BCC level using RESP.[67−69]

RMSD Measurements. RMSD values of all the ligand poses were calculated with respect to the corresponding X-ray geometry of the ligand (without any further optimization) with the "heavy atom" option using the "rmsd.py" script by Schrödinger.

Scoring. *SQM/COSMO SFs.* All of the docked PL complexes with close contacts (cutoff of 1.5 Å) between the protein and the ligand were relaxed by short AMBER/GB optimization as in previous studies.[52] Next, optimal hydrogen positions were localized in each complex using a short MDs run using AMBER/GB as in our previous studies.[52] The SQM/COSMO score is a sum of ΔE_{int} and $\Delta\Delta G_{solv}$ terms. For speed-up and without compromising the reliability,[23] the former term was calculated on large parts of the protein (typically the ligand plus 10 Å protein surroundings) using two approaches: (i) the corrected PM6-D3H4X[44] and (ii) DFTB3-D3H4X method, a third-order DFTB[70,71] with the 3OB parameter set[72,73] and the latest version of the D3H4X corrections for noncovalent interactions.[74] The solvation free energy was calculated on the same truncated system as above using a COSMO implicit solvent model at the PM6 level.[40]

Glide XP Score. All scoring calculations were performed with Glide XP[55] and run in the extra precision (XP) workflow framework. Docking grids were generated by Glide using the cocrystallized ligand at the center of the grid box. The compounds were scored with the option "score in place only".

AutoDock4 and AutoDock Vina. For both AutoDock4[56] and AutoDock Vina,[57] the centers of grid boxes were arranged according to the centers of the crystal ligand poses. The grid box sizes were adjusted to make scoring possible for all combinations of ligands and conformations. AM1-BCC RESP partial charges were used.

UCSF DOCK. The grid spacing was 0.3 Å. The cutoff for nonbonded interactions was not used. We used AMBER parameters. For ligands, we used AM1-BCC RESP partial atomic charges.

Score Scaling. The scores of all the poses of the 17 PL complexes obtained by the 6 SFs were transformed into relative numbers with respect to the score of the crystal pose and normalized as done previously.[23,52]

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acsomega.7b00503.

RMSD analyses of docking poses; enrichment plot; RMSD of the best-scoring poses; numbers of total and hard FPs; detailed computational protocols; timing of SQM/COSMO (PDF)

■ AUTHOR INFORMATION

Corresponding Authors

*E-mail: pavel.hobza@uochb.cas.cz (P.H.).
*E-mail: lepsik@uochb.cas.cz (M.L.).

ORCID

Jan Řezáč: 0000-0001-6849-7314
Pavel Hobza: 0000-0001-5292-6719
Martin Lepšík: 0000-0003-2607-8132

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

We thank Kristian Kříž for helpful ideas on the structures of PL complexes. This work was part of Research Project RVO: 61388963 of the Institute of Organic Chemistry and Biochemistry, Academy of Sciences of the Czech Republic. This work was also supported by the Czech Science Foundation (H.A., S.H., S.M.E., J.F., A.P., P.H., and M.L. from grant No. P208/12/G016 and J.Ř. from grant No. P208/16-11321Y). This work was supported by the Ministry of Education, Youth and Sports from the Large Infrastructures for Research, Experimental Development, and Innovations project "IT4 Innovations National Supercomputing Center—LM2015070," as well as from project LO1305 (P.H.).

■ REFERENCES

(1) Schneider, G. Virtual screening: an endless staircase? *Nat. Rev. Drug Discovery* **2010**, *9*, 273−276.
(2) Irwin, J. J.; Shoichet, B. K. Docking Screens for Novel Ligands Conferring New Biology. *J. Med. Chem.* **2016**, *59*, 4103−4120.
(3) Cheng, T.; Li, X.; Li, Y.; Liu, Z.; Wang, R. Comparative Assessment of Scoring Functions on a Diverse Test Set. *J. Chem. Inf. Model.* **2009**, *49*, 1079−1093.
(4) Wang, Z.; Sun, H.; Yao, X.; Li, D.; Xu, L.; Li, Y.; Tian, S.; Hou, T. Comprehensive evaluation of ten docking programs on a diverse set of protein-ligand complexes: the prediction accuracy of sampling power and scoring power. *Phys. Chem. Chem. Phys.* **2016**, *18*, 12964−12975.
(5) Yuriev, E.; Agostino, M.; Ramsland, P. A. Challenges and advances in computational docking: 2009 in review. *J. Mol. Recognit.* **2011**, *24*, 149−164.
(6) Jain, A. N.; Nicholls, A. Recommendations for evaluation of computational methods. *J. Comput.-Aided Mol. Des.* **2008**, *22*, 133−139.
(7) Kontoyianni, M.; McClellan, L. M.; Sokol, G. S. Evaluation of docking performance: comparative data on docking algorithms. *J. Med. Chem.* **2004**, *47*, 558−565.
(8) Gohlke, H.; Hendlich, M.; Klebe, G. Knowledge-based scoring function to predict protein-ligand interactions. *J. Mol. Biol.* **2000**, *295*, 337−356.
(9) 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.;

DOI: 10.1021/acsomega.7b00503
ACS Omega 2017, 2, 4022−4029

Head, M. S. A critical assessment of docking programs and scoring functions. *J. Med. Chem.* **2006**, *49*, 5912−5931.
(10) Böhm, H.-J. The development of a simple empirical scoring function to estimate the binding constant for a protein ligand complex of known 3-dimensional structure. *J. Comput.-Aided Mol. Des.* **1994**, *8*, 243−256.
(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) Gohlke, H.; Klebe, G. Statistical potentials and scoring functions applied to protein-ligand binding. *Curr. Opin. Struct. Biol.* **2001**, *11*, 231−235.
(13) Spitzmüller, A.; Velec, H. F. G.; Klebe, G. MiniMuDS: A New Optimizer using Knowledge-Based Potentials Improves Scoring of Docking Solutions. *J. Chem. Inf. Model.* **2011**, *51*, 1423−1430.
(14) Ishchenko, A. V.; Shakhnovich, E. I. SMall molecule growth 2001 (SMoG2001): An improved knowledge-based scoring function for protein-ligand interactions. *J. Med. Chem.* **2002**, *45*, 2770−2780.
(15) Ain, Q. U.; Aleksandrova, A.; Roessler, F. D.; Ballester, P. J. Machine-learning scoring functions to improve structure-based binding affinity prediction and virtual screening. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2015**, *5*, 405−424.
(16) Khamis, M. A.; Gomaa, W. Comparative assessment of machine-learning scoring functions on PDBbind 2013. *Eng. Appl. Artif. Intell.* **2015**, *45*, 136−151.
(17) Ewing, T. J. A.; Makino, S.; Skillman, A. G.; Kuntz, I. D. DOCK 4.0: Search strategies for automated molecular docking of flexible molecule databases. *J. Comput.-Aided Mol. Des.* **2001**, *15*, 411−428.
(18) 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.
(19) Li, L.; Wang, B.; Meroueh, S. O. Support vector regression scoring of receptor-ligand complexes for rank-ordering and virtual screening of chemical libraries. *J. Chem. Inf. Model.* **2011**, *51*, 2132−2138.
(20) Řezáč, J.; Hobza, P. Benchmark Calculations of Interaction Energies in Noncovalent Complexes and Their Applications. *Chem. Rev.* **2016**, *116*, 5038−5071.
(21) Fanfrlík, J.; Kolář, M.; Kamlar, M.; Hurný, D.; Ruiz, F. X.; Cousido-Siah, A.; Mitschler, A.; Řezáč, J.; Munusamy, E.; Lepšík, M.; Matějíček, P.; Veselý, J.; Podjarny, A.; Hobza, P. Modulation of Aldose Reductase Inhibition by Halogen Bond Tuning. *ACS Chem. Biol.* **2013**, *8*, 2484−2492.
(22) Fanfrlík, J.; Ruiz, F. X.; Kadlčíková, A.; Řezáč, J.; Cousido-Siah, A.; Mitschler, A.; Haldar, S.; Lepšík, M.; Kolář, M. H.; Majer, P.; Podjarny, A. D.; Hobza, P. The Effect of Halogen-to-Hydrogen Bond Substitution on Human Aldose Reductase Inhibition. *ACS Chem. Biol.* **2015**, *10*, 1637−1642.
(23) Pecina, A.; Meier, R.; Fanfrlík, J.; Lepšík, M.; Řezáč, J.; Hobza, P.; Baldauf, C. The SQM/COSMO filter: reliable native pose identification based on the quantum-mechanical description of protein-ligand interactions and implicit COSMO solvation. *Chem. Commun.* **2016**, *52*, 3312−3315.
(24) Ciancetta, A.; Genheden, S.; Ryde, U. A QM/MM study of the binding of RAPTA ligands to cathepsin B. *J. Comput.-Aided Mol. Des.* **2011**, *25*, 729−742.
(25) Pecina, A.; Lepšík, M.; Řezáč, J.; Brynda, J.; Mader, P.; Řezáčová, P.; Hobza, P.; Fanfrlík, J. QM/MM calculations reveal the different nature of the interaction of two carborane-based sulfamide inhibitors of human carbonic anhydrase II. *J. Phys. Chem. B* **2013**, *117*, 16096−16104.
(26) Mader, P.; Pecina, A.; Cigler, P.; Lepšík, M.; Šícha, V.; Hobza, P.; Grüner, B.; Fanfrlík, J.; Brynda, J.; Řezáčová, P. Carborane-Based Carbonic Anhydrase Inhibitors: Insight into CAII/CAIX Specificity from a High-Resolution Crystal Structure, Modeling, and Quantum Chemical Calculations. *BioMed Res. Int.* **2014**, No. 389869.
(27) Fanfrlík, J.; Brahmkshatriya, P. S.; Řezáč, J.; Jílková, A.; Horn, M.; Mareš, M.; Hobza, P.; Lepšík, M. Quantum Mechanics-Based Scoring Rationalizes the Irreversible Inactivation of Parasitic *Schistosoma mansoni* Cysteine Peptidase by Vinyl Sulfone Inhibitors. *J. Phys. Chem. B* **2013**, *117*, 14973−14982.
(28) Söderhjelm, P.; Ryde, U. How Accurate Can a Force Field Become? A Polarizable Multipole Model Combined with Fragment-wise Quantum-Mechanical Calculations. *J. Phys. Chem. A* **2009**, *113*, 617−627.
(29) Berg, L.; Mishra, B. K.; Andersson, C. D.; Ekström, F.; Linusson, A. The Nature of Activated Non-classical Hydrogen Bonds: A Case Study on Acetylcholinesterase-Ligand Complexes. *Chem. − Eur. J.* **2016**, *22*, 2672−2681.
(30) Wichapong, K.; Rohe, A.; Platzer, C.; Slynko, I.; Erdmann, F.; Schmidt, M.; Sippl, W. Application of Docking and QM/MM-GBSA Rescoring to Screen for Novel Myt1 Kinase Inhibitors. *J. Chem. Inf. Model.* **2014**, *54*, 881−893.
(31) Chaskar, P.; Zoete, V.; Röhrig, U. F. Toward On-The-Fly Quantum Mechanical/Molecular Mechanical (QM/MM) Docking: Development and Benchmark of a Scoring Function. *J. Chem. Inf. Model.* **2014**, *54*, 3137−3152.
(32) Chaskar, P.; Zoete, V.; Röhrig, U. F. On-the-Fly QM/MM Docking with Attracting Cavities. *J. Chem. Inf. Model.* **2017**, *57*, 73−84.
(33) Burger, S. K.; Thompson, D. C.; Ayers, P. W. Quantum Mechanics/Molecular Mechanics Strategies for Docking Pose Refinement: Distinguishing between Binders and Decoys in Cytochrome c Peroxidase. *J. Chem. Inf. Model.* **2011**, *51*, 93−101.
(34) Antony, J.; Grimme, S.; Liakos, D. G.; Neese, F. Protein-Ligand Interaction Energies with Dispersion Corrected Density Functional Theory and High-Level Wave Function Based Methods. *J. Phys. Chem. A* **2011**, *115*, 11210−11220.
(35) Raha, K.; Merz, K. M. A quantum mechanics-based scoring function: Study of zinc ion-mediated ligand binding. *J. Am. Chem. Soc.* **2004**, *126*, 1020−1021.
(36) Raha, K.; Merz, K. M. Large-scale validation of a quantum mechanics based scoring function: Predicting the binding affinity and the binding mode of a diverse set of protein-ligand complexes. *J. Med. Chem.* **2005**, *48*, 4558−4575.
(37) Řezáč, J.; Hobza, P. Advanced Corrections of Hydrogen Bonding and Dispersion for Semiempirical Quantum Mechanical Methods. *J. Chem. Theory Comput.* **2012**, *8*, 141−151.
(38) Kolář, M.; Fanfrlík, J.; Lepšík, M.; Forti, F.; Luque, F. J.; Hobza, P. Assessing the Accuracy and Performance of Implicit Solvent Models for Drug Molecules: Conformational Ensemble Approaches. *J. Phys. Chem. B* **2013**, *117*, 5950−5962.
(39) Stewart, J. J. P. Optimization of parameters for semiempirical methods VI: more modifications to the NDDO approximations and re-optimization of parameters. *J. Mol. Model.* **2013**, *19*, 1−32.
(40) Klamt, A.; Schüürmann, G. COSMO - A new approach to dielectric screening in solvents with explicit expressions for the screening energy and its gradient. *J. Chem. Soc., Perkin Trans. 2* **1993**, 799−805.
(41) Sulimov, A. V.; Kutov, D. C.; Katkova, E. V.; Sulimov, V. B. Combined Docking with Classical Force Field and Quantum Chemical Semiempirical Method PM7. *Adv. Bioinf.* **2017**, No. 7167691.
(42) Oferkin, I. V.; Katkova, E. V.; Sulimov, A. V.; Kutov, D. C.; Sobolev, S. I.; Voevodin, V. V.; Sulimov, V. B. Evaluation of Docking Target Functions by the Comprehensive Investigation of Protein-Ligand Energy Minima. *Adv. Bioinf.* **2015**, No. 126858.
(43) Hostaš, J.; Řezáč, J.; Hobza, P. On the performance of the semiempirical quantum mechanical PM6 and PM7 methods for noncovalent interactions. *Chem. Phys. Lett.* **2013**, *568*−*569*, 161−166.
(44) Řezáč, J.; Hobza, P. A halogen-bonding correction for the semiempirical PM6 method. *Chem. Phys. Lett.* **2011**, *506*, 286−289.
(45) Fanfrlík, J.; Bronowska, A. K.; Řezáč, J.; Přenosil, O.; Konvalinka, J.; Hobza, P. A Reliable Docking/Scoring Scheme Based on the Semiempirical Quantum Mechanical PM6-DH2 Method Accurately Covering Dispersion and H-Bonding: HIV-1 Protease with 22 Ligands. *J. Phys. Chem. B* **2010**, *114*, 12666−12678.
(46) Lepšík, M.; Řezáč, J.; Kolář, M.; Pecina, A.; Hobza, P.; Fanfrlík, J. The Semiempirical Quantum Mechanical Scoring Function for In Silico Drug Design. *ChemPlusChem* **2013**, *78*, 921−931.

DOI: 10.1021/acsomega.7b00503
ACS Omega 2017, 2, 4022−4029

(47) Vorlová, B.; Nachtigallová, D.; Jirásková-Vaníčková, J.; Ajani, H.; Jansa, P.; Řezáč, J.; Fanfrlík, J.; Otyepka, M.; Hobza, P.; Konvalinka, J.; Lepšík, M. Malonate-based inhibitors of mammalian serine racemase: kinetic characterization and structure-based computational study. *Eur. J. Med. Chem.* **2015**, *89*, 189−197.
(48) Cousido-Siah, A.; Ruiz, F. X.; Fanfrlík, J.; Giménez-Dejoz, J.; Mitschler, A.; Kamlar, M.; Veselý, J.; Ajani, H.; Parés, X.; Farrés, J.; Hobza, P.; Podjarny, A. D. IDD388 Polyhalogenated Derivatives as Probes for an Improved Structure-Based Selectivity of AKR1B10 Inhibitors. *ACS Chem. Biol.* **2016**, *11*, 2693−2705.
(49) Dostál, J.; Pecina, A.; Hrušková-Heidingsfeldová, O.; Marečková, L.; Pichová, I.; Řezáčová, P.; Lepšík, M.; Brynda, J. Atomic resolution crystal structure of Sapp2p, a secreted aspartic protease from *Candida parapsilosis*. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **2015**, *71*, 2494−2504.
(50) Nekardová, M.; Vymětalová, L.; Khirsariya, P.; Kováčová, S.; Hylsová, M.; Jorda, R.; Kryštof, V.; Fanfrlík, J.; Hobza, P.; Paruch, K. Structural Basis of the Interaction of Cyclin-Dependent Kinase 2 with Roscovitine and Its Analogues Having Bioisosteric Central Heterocycles. *ChemPhysChem* **2017**, 785.
(51) Hylsová, M.; Carbain, B.; Fanfrlík, J.; Musilová, L.; Haldar, S.; Köprülüoğlu, C.; Ajani, H.; Brahmkshatriya, P. S.; Jorda, R.; Kryštof, V.; Hobza, P.; Echalier, A.; Paruch, K.; Lepšík, M. Explicit treatment of active-site waters enhances quantum mechanical/implicit solvent scoring: Inhibition of CDK2 by new pyrazolo[1,5-a]pyrimidines. *Eur. J. Med. Chem.* **2017**, *126*, 1118−1128.
(52) Pecina, A.; Haldar, S.; Fanfrlík, J.; Meier, R.; Řezáč, J.; Lepšík, M.; Hobza, P. SQM/COSMO Scoring Function at the DFTB3-D3H4 Level: Unique Identification of Native Protein-Ligand Poses. *J. Chem. Inf. Model.* **2017**, *57*, 127−132.
(53) Elstner, M.; Hobza, P.; Frauenheim, T.; Suhai, S.; Kaxiras, E. Hydrogen bonding and stacking interactions of nucleic acid base pairs: A density-functional-theory based treatment. *J. Chem. Phys.* **2001**, *114*, 5149−5155.
(54) Miriyala, V. M.; Řezáč, J. Description of non-covalent interactions in SCC-DFTB methods. *J. Comput. Chem.* **2017**, *38*, 688−697.
(55) Friesner, R. A.; Banks, J. L.; Murphy, R. B.; Halgren, T. A.; Klicic, J. J.; Mainz, D. T.; Repasky, M. P.; Knoll, E. H.; Shelley, M.; Perry, J. K.; Shaw, D. E.; Francis, P.; Shenkin, P. S. Glide: A new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. *J. Med. Chem.* **2004**, *47*, 1739−1749.
(56) 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. Comput. Chem.* **2009**, *30*, 2785−2791.
(57) Trott, O.; Olson, A. J. Software News and Update AutoDock Vina: Improving the Speed and Accuracy of Docking with a New Scoring Function, Efficient Optimization, and Multithreading. *J. Comput. Chem.* **2010**, *31*, 455−461.
(58) Allen, W. J.; Balius, T. E.; Mukherjee, S.; Brozell, S. R.; Moustakas, D. T.; Lang, P. T.; Case, D. A.; Kuntz, I. D.; Rizzo, R. C. DOCK 6: Impact of New Features and Current Docking Performance. *J. Comput. Chem.* **2015**, *36*, 1132−1156.
(59) Li, Y.; Liu, Z.; Li, J.; Han, L.; Liu, J.; Zhao, Z.; Wang, R. Comparative Assessment of Scoring Functions on an Updated Benchmark: 1. Compilation of the Test Set. *J. Chem. Inf. Model.* **2014**, *54*, 1700−1716.
(60) Kirchmair, J.; Wolber, G.; Laggner, C.; Langer, T. Comparative performance assessment of the conformational model generators omega and catalyst: A large-scale survey on the retrieval of protein-bound ligand conformations. *J. Chem. Inf. Model.* **2006**, *46*, 1848−1861.
(61) Warren, G. L.; Do, T. D.; Kelley, B. P.; Nicholls, A.; Warren, S. D. Essential considerations for using protein-ligand structures in drug discovery. *Drug Discovery Today* **2012**, *17*, 1270−1281.
(62) Zilian, D.; Sotriffer, C. A. SFCscore(RF): A Random Forest-Based Scoring Function for Improved Affinity Prediction of Protein-Ligand Complexes. *J. Chem. Inf. Model.* **2013**, *53*, 1923−1933.
(63) Zhao, H.; Caflisch, A. Discovery of ZAP70 inhibitors by high-throughput docking into a conformation of its kinase domain generated by molecular dynamics. *Bioorg. Med. Chem. Lett.* **2013**, *23*, 5721−5726.
(64) Jain, A. N. Surflex: Fully automatic flexible molecular docking using a molecular similarity-based search engine. *J. Med. Chem.* **2003**, *46*, 499−511.
(65) *Small-Molecule Drug Discovery Suite 2016-1*; Schrödinger, LLC: New York, NY, 2016.
(66) Case, D. A.; Babin, V.; Berryman, J. T.; Betz, R. M.; Cai, Q.; Cerutti, D. S.; Cheatham, T. E.; Darden, T. A.; Duke, R. E.; Gohlke, H.; Goetz, A. W.; Gusarov, S.; Homeyer, N.; Janowski, P.; Kaus, J.; Kolossváry, I.; Kovalenko, A.; Lee, T. S.; LeGrand, S.; Luchko, T.; Luo, R.; Madej, B.; Merz, K. M.; Paesani, F.; Roe, D. R.; Roitberg, A.; Sagui, C.; Salomon-Ferrer, R.; Seabra, G.; Simmerling, C. L.; Smith, W.; Swails, J.; Walker; Wang, J.; Wolf, R. M.; Wu, X.; Kollman, P. A. *AMBER 14*; University of California: San Francisco, 2014.
(67) 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.
(68) 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.
(69) 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.
(70) Gaus, M.; Cui, Q.; Elstner, M. DFTB3: Extension of the Self-Consistent-Charge Density-Functional Tight-Binding Method (SCC-DFTB). *J. Chem. Theory Comput.* **2011**, *7*, 931−948.
(71) Gaus, M.; Goez, A.; Elstner, M. Parametrization and Benchmark of DFTB3 for Organic Molecules. *J. Chem. Theory Comput.* **2013**, *9*, 338−354.
(72) Gaus, M.; Lu, X.; Elstner, M.; Cui, Q. Parameterization of DFTB3/3OB for Sulfur and Phosphorus for Chemical and Biological Applications. *J. Chem. Theory Comput.* **2014**, *10*, 1518−1537.
(73) Lu, X.; Gaus, M.; Elstner, M.; Cui, Q. Parametrization of DFTB3/3OB for magnesium and zinc for chemical and biological applications. *J. Phys. Chem. B* **2015**, *119*, 1062−1082.
(74) Kubillus, M.; Kubař, T.; Gaus, M.; Řezáč, J.; Elstner, M. Parameterization of the DFTB3 method for Br, Ca, Cl, F, I, K, and Na in organic and biological systems. *J. Chem. Theory Comput.* **2015**, *11*, 332−342.

DOI: 10.1021/acsomega.7b00503
ACS Omega 2017, 2, 4022−4029

Publication C

ChemPubSoc Europe

DOI: 10.1002/cphc.201701104

CHEMPHYSCHEM Articles

Ranking Power of the SQM/COSMO Scoring Function on Carbonic Anhydrase II–Inhibitor Complexes

Adam Pecina[+],[a] Jiří Brynda[+],[a, b] Lukáš Vrzal,[a] Ramachandran Gnanasekaran,[a, e] Magdalena Hořejší,[b] Saltuk M. Eyrilmez,[a, c] Jan Řezáč,[a] Martin Lepšík,[a] Pavlína Řezáčová,[a, b] Pavel Hobza,[a, d] Pavel Majer,[a] Václav Veverka,*[a] and Jindřich Fanfrlík*[a]

Accurate prediction of protein–ligand binding affinities is essential for hit-to-lead optimization and virtual screening. The reliability of scoring functions can be improved by including quantum effects. Here, we demonstrate the ranking power of the semiempirical quantum mechanics (SQM)/implicit solvent (COSMO) scoring function by using a challenging set of 10 inhibitors binding to carbonic anhydrase II through Zn^{2+} in the active site. This new dataset consists of the high-resolution (1.1–1.4 Å) crystal structures and experimentally determined inhibitory constant (K_i) values. It allows for evaluation of the common approximations, such as representing the solvent implicitly or by using a single target conformation combined with a set of ligand docking poses. SQM/COSMO attained a good correlation of R^2 of 0.56–0.77 with the experimental inhibitory activities, benefiting from careful handling of both noncovalent interactions (e.g. charge transfer) and solvation. This proof-of-concept study of SQM/COSMO ranking for metalloprotein–ligand systems demonstrates its potential for hit-to-lead applications.

1. Introduction

The ultimate goal of computational drug design is to accurately predict the binding affinities of ligands to their targets. Structure-based approaches traditionally employ docking/scoring methodology but molecular-dynamics approaches have been increasingly successful. Scoring functions (SFs) have certain requirements for reliability and speed.[1] The approximations used in physics-based scoring limit accuracy. Reliable prediction of protein-ligand (P–L) binding affinities depends on accurate description of noncovalent interactions, as well as other factors.[2–3]

Quantum mechanics (QM) approaches can accurately describe noncovalent interactions. QM qualitatively and quantitatively describes quantum phenomena occurring in P-L binding, such as charge transfer in metalloprotein binding,[4–8] halogen (X)-bonding,[9] and covalent bond formation.[10] The high computational requirements of QM descriptions can be reduced by using linear-scaling semiempirical QM (SQM) methods to handle systems up to 10,000 atoms.[11] The quality of original SQM methods (such as PM6) was low, and thus we developed transferable corrections for dispersion (D), hydrogen (H)-bonding, and X-bonding.[12] Combining these corrections with PM6 or the more advanced density-functional tight-binding (DFTB3) method yields the PM6-D3H4X and DFTB3-D3H4X methods, which give highly accurate descriptions of noncovalent interactions comparable to very demanding "gold standard" QM methods.[12] PM6-D3H4X formed the core of SQM-based SF,[13,14] designed as a sum of several terms of thought decomposition of drug binding, including the gas-phase interaction energy (ΔE_{int}), changes in the solvation free energy upon complex formation ($\Delta\Delta G_{solv}$), and the conformational "free" energy change upon complex formation ($\Delta G'_{confw}$).[13,14] P-L complexes are optimized in an aqueous environment prior to scoring. The full SQM-based SF has already been successfully applied for analysis of small molecule series inhibiting various enzyme classes (oxidoreductases, proteases, and kinases).[9,13–15]

We have recently simplified and accelerated the SQM-based SF by employing only the two dominant terms, ΔE_{int} and $\Delta\Delta G_{solv}$, and neglecting SQM optimization.[6] This novel

[a] Dr. A. Pecina,[+] Dr. J. Brynda,[+] L. Vrzal, Dr. R. Gnanasekaran, S. M. Eyrilmez, Dr. J. Řezáč, Dr. M. Lepšík, Dr. P. Řezáčová, Prof. P. Hobza, Dr. P. Majer, Dr. V. Veverka, Dr. J. Fanfrlík
Institute of Organic Chemistry and Biochemistry of the Czech Academy of Sciences
Flemingovo nam. 2, 16610 Prague 6 (Czech Republic)
E-mail: veverka@uochb.cas.cz
fanfrlik@uochb.cas.cz

[b] Dr. J. Brynda,[+] M. Hořejší, Dr. P. Řezáčová
Institute of Molecular Genetics of Czech Academy of Sciences
Videnska 1083, 14220 Prague 4 (Czech Republic)

[c] S. M. Eyrilmez
Palacký University, 77146 Olomouc (Czech Republic)

[d] Prof. P. Hobza
Regional Centre of Advanced Technologies and Materials
Palacký University, 77146 Olomouc (Czech Republic)

[e] Dr. R. Gnanasekaran
Current address: Department of Chemistry
Pondicherry University,Puducherry, 605014 (India)

[+] These authors contributed equally to this work

Supporting Information and the ORCID identification number(s) for the author(s) of this article can be found under: https://doi.org/10.1002/cphc.201701104.

scheme, called the SQM/COSMO SF, is by two order of magnitude faster than the full version of the SQM-based SF. The time requirements were thus reduced from days to tens of minutes, which makes the SQM/COSMO SF applicable to drug design even in an industrial context. It was demonstrated that the SQM/COSMO SF outperformed standard SFs at the PM6-D3H4X and DFTB3-D3H4X (for Zinc metalloproteins) levels in identifying the native binding pose,[6,7,16] which is a critical prerequisite for affinity estimation in physics-based scoring.[17,18] However, it was not clear whether the simplified SQM/COSMO SF could reliably estimate P-L binding affinities and provide the valuable ranking for an inhibitor series.

In the present study, we tested the ranking power of the SQM/COSMO SF at the DFTB3-D3H4X level-specifically, its ability to reliably rank a series of structurally similar carbonic anhydrase II (CAII) inhibitors, an important task during the hit-to-lead development phase-and compared it to those of widely used classical SFs (GOLD, DOCK 6, AutoDock Vina, Autodock4)[19–26] and AMBER molecular-mechanics (MM) force field.[27] Inhibition of CAII, a zinc metalloenzyme essential for maintaining general acid-base equilibrium, has been studied extensively.[28–30] Here, we compared virtual scoring results with experimentally determined inhibition constants (K_i) and high-resolution crystal structures of CAII–inhibitor complexes for 10 compounds containing a benzenesulfonamide moiety.

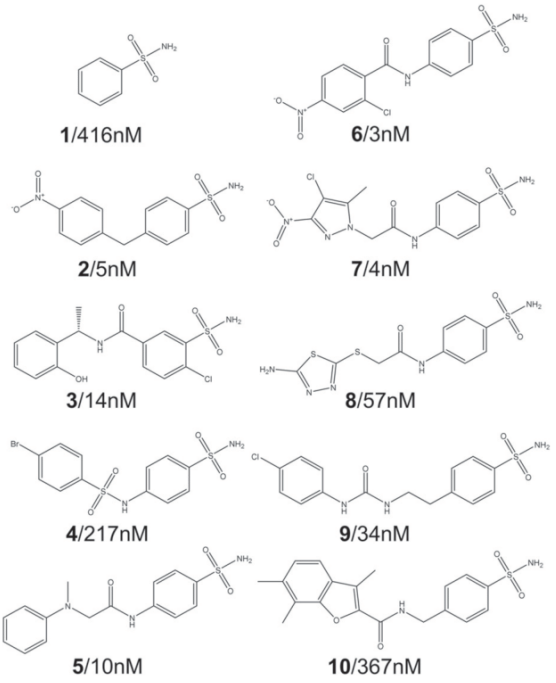


Figure 1. Benzenesulfonamide-containing CAII inhibitors and experimentally determined inhibition constants (K_i).

2. Results and Discussion

2.1. Compounds

From the ZINC database,[31] we selected 10 small molecules with a benzenesulfonamide moiety, known to inhibit CAII,[29,30] and determined their ability to inhibit recombinant CAII using a standard assay (Figure 1).[32] The compounds were ordered and numbered according to the length of the extension of the benzenesulfonamide core. The molecular mass and octanol-water partition coefficient ($\log P_{ow}$) of the ligands ranged from 157 to 391 Da and 0.62 to 3.29, respectively (Table S1 in the Supporting Information). The K_i values of the ligands spanned three orders of magnitude, ranging from 416 nM for the basic scaffold (**1**) to 3 nM (**6**), and were not correlated to either the size or the lipophilicity of the molecules (R^2 of 0.19 and 0.01, respectively; data in Table S1).

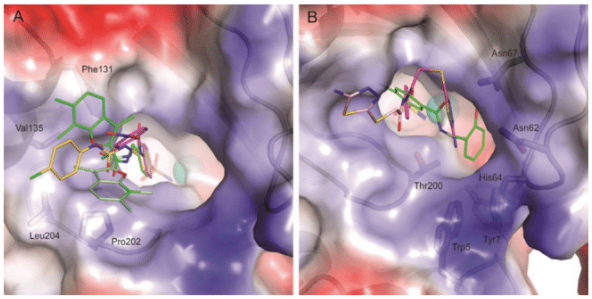


Figure 2. Binding mode of inhibitors in the CAII active site. The solvent accessible surface of the protein is colored by electrostatic potential; interacting residues are highlighted in sticks and labeled. (A) Inhibitors **6**, **9**, and **10** interacting with the hydrophobic pocket are shown as sticks with carbon atoms colored magenta for **6**, yellow for **9**, and green for **10** (two alternative conformations are shown). (B) Inhibitors **3** and **8** interacting with the hydrophilic pocket are shown as sticks with carbon atoms colored green for **3** (one conformation is shown for clarity) and magenta for **8** (two alternative conformations are shown).

2.2. Crystal Structures

All compounds were co-crystallized with CAII. Structures were determined at 1.1–1.4 Å resolution (X-ray statistics are given in Table S2). The compounds were modeled into well-defined electron density within the active site. The high quality of the structures allowed us to interpret electron density maps revealing two alternative inhibitor conformations (for **4**, **5**, **7**, **8**, and **10**). These inhibitors interact with CAII through the deeply buried sulfonamide group, which establishes charged/polar interactions with Zn^{2+} and residues distributed at the bottom of the active site (Figure 2A). The ionized amine ($-NH^-$) of the sulfonamide coordinates Zn^{2+} at a distance of 2.0 Å and forms an H-bond with the hydroxyl group of the Thr199 side chain. One O atom of the sulfonamide moiety forms an H-bond with the backbone amine of Thr199. Additionally, the inhibitors interact with residues in the pockets at the active site entrance. Most interact with the hydrophobic pocket formed by Phe131, Val135, Leu198, Pro200, and Leu204 (Figure 2A). The exceptions are **3** and **8**, which interact with the hydrophilic pocket formed by Trp5, Tyr7, Asn62, Asn67, His64, and Thr200 (Figure 2B). The position of the benzenesulfonamide moiety is conserved for all inhibitors except **3**, for which this ring is ro-

tated by roughly 60° due to a substituent in the *ortho* position (Figure 2B).

2.3. Scoring

For computational modeling, we scored the crystal structures of the CAII-inhibitor complexes with the SQM/COSMO SF at the DFTB3-D3H4X level, classical SFs (GOLD, DOCK 6, AutoDock Vina, Autodock4) and AMBER MM force field coupled with different implicit solvent models ($AMBER^{AM1}$/GB, $AMBER^{HF}$/GB, $AMBER^{HF}$/PB, $AMBER^{HF}$/COSMO, see Methods). Using information from the structures (binding mode, position of active-site waters, protein conformation), we obtained a reasonable correlation between the SQM/COSMO scores and experimental binding data (R^2 of 0.69, predictive index PI[33] of 0.81; see Figures 3 and Graphs S1 and S2). The other SFs had much worse or no correlation with the experimental data, with the best results found for AutoDock Vina (R^2 of 0.39, PI of 0.67), DOCK 6 (R^2 of 0.38, PI of 0.50) and $AMBER^{HF}$/COSMO (R^2 of 0.31, PI of 0.45). Poor correlation was obtained with the $AMBER^{AM1}$/GB, $AMBER^{HF}$/GB and $AMBER^{HF}$/PB. This may be due to the missing description of the charge transfer between the ligand and the metal ion in MM in this difficult case. MM based SFs might thus be expected to give better results for other targets.

Scores obtained on this series of crystal structures show the performance limit of SFs in the "single-structure approach,"[2] that is, without sampling P–L conformations. It should be noted however that it is not common scenario for a drug design project to have crystal structures of all the compounds in complex with the target. In order to follow a more realistic scenario, such as the standard "implicit solvent" approximation, we omitted crystal waters in scoring. The correlation with the experimental binding data decreased with the exception of $AMBER^{HF}$/PB (R^2 of 0.28, PI of 0.60) and $AMBER^{HF}$/COSMO (R^2 of 0.51, PI of 0.67) SFs. The correlation of $AMBER^{HF}$/COSMO was even comparable to SQM/COSMO, which again, yielded the best results (R^2 of 0.58 and PI of 0.76). It distinguished strong (inhibitors with K_i ranging from 2.8 to 13.5 nM had scores ranging from −46.3 to −39.0 kcal mol^{-1}) from weaker binders (K_i ranging from 33.9 to 415.6 nM and scores from −34.9 to −29.6 kcal mol^{-1}), while the remaining SFs did not show correlation (max. R^2 of 0.19 and PI of 0.34).

Subsequently, we used a single protein conformation from the CAII-1 structure to score the other inhibitors and test the approximations in the "rigid protein conformation" approach. The binding poses were aligned to the CAII-**1** complex, and explicit waters were not considered. The correlation with the experimental data was not considerably worse for the SQM/COSMO SF (R^2 of 0.56 and PI of 0.64), and its ability to distinguish between strong and weak binders was preserved. The low correlation of $AMBER^{HF}$/COSMO (R^2 of 0.18, PI of 0.48) was caused by the low score of **6**, that is, omitting this data point resulted in R^2/PI of 0.44/0.62. These results may be due in part to the relatively well-conserved geometry of the CAII binding site. Obtaining correlation without considering explicit waters does not contradict previous findings that waters are important in P–L binding.[9,15,34] Rather, it demonstrates that the implicit COSMO model can capture some of these effects in some P–L systems. The correlation of the classical SFs remained low. Interestingly, the "single protein conformation" approximation improved the DOCK 6 results. However, the R^2/PI of 0.33/0.46 for DOCK 6 was mainly due to the very low score of **10** (omitting this data point resulted in R^2/PI of 0.18/

A

Method	10 crystals; crystal waters	10 crystals; no waters	single crystal; no waters
	R^2/PI		
SQM/COSMO	**0.69/0.81**	**0.58/0.76**	**0.56/0.64**
Vina	0.39/0.67	0.07/0.20	0.01/-0.11
DOCK 6	0.38/0.50	0.10/-0.22	0.33/0.46
$AMBER^{HF}$/COSMO	0.31/0.45	0.51/0.67	0.18/0.48
Autodock 4	0.19/0.53	0.09/0.06	0.04/0.13
Gold ASP	0.15/0.47	0.12/0.35	0.11/0.33
GoldPLP	0.12/0.39	0.13/0.39	0.08/0.22
$AMBER^{HF}$/GB	0.09/0.45	0.13/0.30	0.01/0.16
$AMBER^{AM1}$/GB	0.05/0.23	0.14/0.37	0.12/0.32
$AMBER^{HF}$/PB	0.02/-0.05	0.28/0.60	0.00/-0.18
GoldScore	0.01/0.15	0.01/-0.10	0.17/0.37
Chemscore	0.01/0.08	0.01/-0.10	0.02/0.08

B

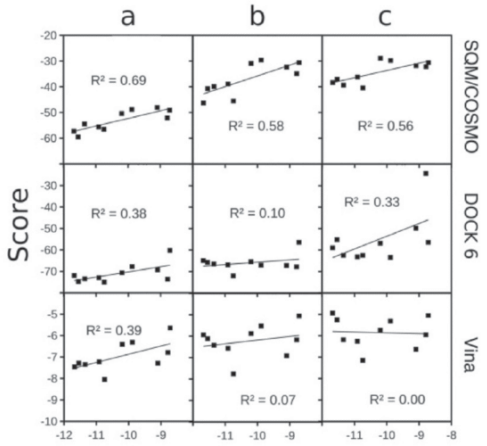


Figure 3. A) Correlation between experimental and calculated binding data expressed by coefficient of determination (R^2) and predictive index (PI). B) SQM/COSMO, DOCK 6, and Vina scores plotted against experimental binding free energy values. a) Ten crystal structures used, b) crystal water molecules not considered, and c) single crystal conformation used. Energies in kcal mol^{-1}.

0.32), which was caused by a clash between the H atoms of **10** and Phe131. Notably, the clash introduced due to the soft repulsive potential of the docking function was eliminated during the H optimization step, which is present in SQM/COSMO and AMBER scoring.[6]

Finally, to simulate a real-world scenario, the performance of the SQM/COSMO was tested on docked poses (using DOCK 6) of all the inhibitors into the single CAII conformation (taken from the CAII-**1** crystal). The generated poses were rescored by SQM/COSMO SF and the most negative scores were further used. The comparison of obtained scores for docked poses and 10 crystal structures confirmed the ability of the SQM/COSMO SF to identify the X-ray pose,[6,7,16] that is, crystal structures had the most negative scores in all the cases (Figure 4). Importantly, the docked poses did not worsen the correlation with the experimental binding data. The obtained correlation (R^2/PI of 0.77/0.92) was even slightly better than that for the crystal structures. This might be due to error cancelation on consistently prepared docked poses. However, this effect should not be overestimated, for example, DOCK 6 scores of the docked poses did not have a better correlation with the experimental binding data (R^2/PI of 0.10/0.18).

A

	Docked poses; no waters		10 crystals; no waters
Inhibitor	RMSD	SQM/COSMO	SQM/COSMO
1	0.19	-28.3	-31.0
2	0.47	-34.8	-39.9
3	0.49	-33.4	-45.4
4	0.44	-27.0	-33.5
5	0.19	-30.7	-39.0
6	0.86	-39.6	-46.3
7	0.35	-35.3	-40.8
8	0.33	-29.6	-29.6
9	0.56	-27.5	-31.2
10	0.86	-26.6	-34.9

B

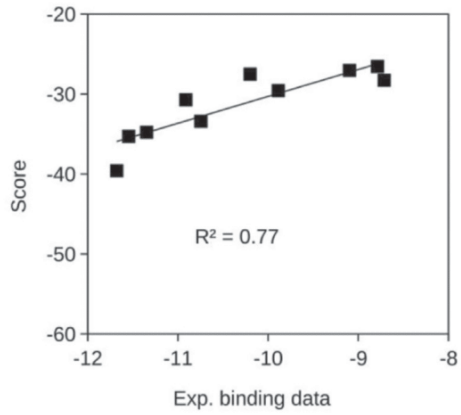


Figure 4. A) RMSD and SQM/COSMO scores for docked poses, and SQM/COSMO scores for crystal structures. B) SQM/COSMO obtained on DOCK 6 poses plotted against experimental binding free energy values. Distances in Å and energies in kcal mol^{-1}.

3. Conclusions

We present the ranking power of SQM/COSMO scoring function that captures quantum effects in protein–ligand binding. We designed a challenging set of 10 inhibitors of carbonic anhydrase II binding through Zn^{2+}, determined their inhibitory constants (K_i) and high-resolution (1.1–1.4 Å) crystal structures which allowed us to assess the effects of using implicit solvent models, a single protein conformation with a set of docked ligand poses. The standard scoring functions (GOLD, AMBER/GB, AMBER/PB, DOCK 6, Vina, Autodock4) did not correlate with the experimental binding data ($R^2 < 0.39$). In contrast, SQM/COSMO provided a fair correlation (R^2 of 0.56–0.77). This proof-of-concept study demonstrates the advantages of SQM/COSMO ranking for zinc metalloprotein inhibitors. A large-scale testing on diverse protein–ligands complexes will show whether the increased computational cost of SQM/COSMO scoring can be outweighed by its benefits.

Experimental Section

Protein Preparation

Recombinant CAII was expressed in *E.coli* BL21(DE3) and purified as described by Pinard et al.[35] Protein was stored in 25 mM NaH_2PO_4, pH 6.5, 100 mM NaCl and used in inhibition assays and crystallization experiments.

CAII Inhibition Assay

An Applied Photophysics stopped-flow instrument was used to assess CAII-catalyzed CO_2 hydration activity in the presence of inhibitors. Phenol red at a concentration of 0.1 mM was used as a pH indicator, with an absorbance maximum of 557 nm. Reactions were performed at 298 K temperature in 20 mM HEPES, pH 7.5, containing 20 mM Na_2SO_4. Rates of the CAII-catalyzed CO_2 hydration reaction were followed for a period of 30 s; the CO_2 concentration was 8.5 mM. For each inhibitor, at least three traces of the initial 5–10% of the reaction were used to determine the initial velocity. The uncatalyzed rates were determined in the same manner and subtracted from the total observed rates. Stock solutions of inhibitor (30 mM) were prepared in dimethylsulfoxide (DMSO), and dilutions up to 5 nM were prepared in 20 mM HEPES, pH 7.5, 20 mM Na_2SO_4. Inhibitor and enzyme solutions were preincubated for 5 min at 298 K temperature prior to substrate addition to allow for formation of the E-I complex. The inhibition constants (Table S1) were obtained by nonlinear least-squares methods using EXCEL spreadsheets.

Protein Crystallization and X-ray Data Collection

Complexes of recombinant CAII and ligands were prepared by addition of a 2-fold molar excess of inhibitor (dissolved in pure DMSO) to 24 mg mL^{-1} protein solution in 50 mM Tris, pH 7.8. The best crystals were prepared by the vapor-diffusion hanging drop method at 291 K using a precipitation solution containing 1.6 M sodium citrate, 50 mM Tris-HCl, pH 7.8. Drops containing 2 μl complex solution and 1 μl precipitant solution were equilibrated over reservoirs containing 1 mL precipitant solution. The final DMSO concentration in the drop did not exceed 5% (v/v). Crystals suitable for X-ray measurement typically grew within 1–2 weeks.

Before data collection, the crystals were soaked for 5–10 s in a reservoir solution supplemented with 20% (v/v) sucrose and stored in liquid N_2. Diffraction data at 100 K were collected on BL14.1 operated by the Helmholtz-Zentrum Berlin (HZB) at the BESSY II electron storage ring (Berlin-Adlershof, Germany).[36] Diffraction data were processed using the XDS suite of programs.[37,38] Crystal parameters and data collection statistics are summarized in Table S1.

Structure Determination, Refinement, and Analyses

Crystal structures were determined by the difference Fourier technique using the coordinates of the CAII structure (PDB entry 3PO6)[39] as a model. Atomic coordinates of inhibitor molecules were generated by quantum mechanical (QM) optimizations in the Turbomole package[40] using the density functional theory (DFT) method with the B-LYP functional and the SVP basis set, augmented with empirical dispersion correction.[41] The geometric libraries for the inhibitors were generated using the Libcheck program.[42] The Coot program[43] was used for inhibitor fitting, model rebuilding, and addition of water molecules. Refinement was carried out with Refmac5,[44] with roughly 1,000 reflections reserved for cross-validation.

The structures were first refined with isotropic atomic displacement parameters (ADPs). After adding solvent atoms and zinc ions, building inhibitor molecules in the active site, and determining several alternate conformations for a number of residues, anisotropic ADPs were refined for nearly all atoms (with the exception of spatially overlapping atoms in segments with alternate conformations; additionally, oxygen atoms of water molecules with an unrealistic ratio of ellipsoid axes were refined with isotropic ADPs) including atoms in the inhibitor molecules. The structure of CAII in complex with 9 was refined with isotropic ADPs using diffraction data to a resolution of 1.4 Å. The quality of crystallographic models was assessed with MolProbity.[45] The final refinement statistics are summarized in Table S2. All figures representing structures were created using PyMOL.[46] Atomic coordinates and structure factors for the crystal structures were deposited in the PDB with accession codes specified in Table S2.

Computational Section

Preparation of Proteins

Ten crystal structures of CAII in complex with sulfonamide-based ligands were determined. Computational models were prepared according to the following procedure. Only A conformations of side chains were considered. Hydrogens for the proteins were added using the Reduce[47] and LEaP[48] modules in the AMBER10 package.[49] The protonation states of individual histidines were assigned based on visual inspection of their surroundings. The protein N-terminus and all lysines and arginines were considered positively charged and the C-terminus and all glutamates and aspartates were considered negatively charged to reflect the predominant state at pH 7. Using information from the crystal structures, the important water molecules in the binding cavities (bridging water molecules or water molecules within the water network between the ligand and the protein) were identified and retained. The total number of such waters ranged from 15 to 21 molecules. One water molecule bridging the inhibitors and CAII residues Tyr7, Glu106, and Thr198 was retained to maintain the integrity of the active site. Other waters were discarded for all further calculations. The protein parameters were obtained from the ff03 force field,[28] and the positions of the added hydrogen atoms were relaxed in vacuo using the FIRE algorithm followed by annealing (5 ps) from 1700 K to 0 K using the Berendsen thermostat in the SANDER module of the AMBER 10 package.[49]

Preparation of Ligands

The inhibitors were protonated with the UCSF Chimera program.[50] The sulfamide moiety that binds to the Zn^{2+} of CAII was modeled in a deprotonated NH^- form.[51,52] The force field parameters were taken from GAFF,[53] and partial atomic charges were determined by the RESP procedure at the AM1-BCC level.[54]

Computational Methods

Physics-based scoring functions are sensitive to molecular details, such as hydrogen bond orientation or geometric clashes. Therefore, we automated the three-step geometry optimization of hydrogen atoms that are in contact with the ligand. For each protein–ligand complex, hydrogens were first optimized in GB with the ff03 force field using the SD algorithm, then by simulated annealing (3 ps) from 1700 K to 0 K using SANDER, and finally by optimization with the FIRE algorithm in AMBER10.[49] Consistently prepared complexes were then rescored using the following set-ups:

The SQM/COSMO scoring function consists of two terms. The first term is an interaction energy in the gas phase calculated with the self-consistent charge density functional tight-binding scheme, including the third-order terms and the 3OB Slater-Koster parameters, augmented with the dispersion, hydrogen-bonding and halogen bonding corrections (SCC-DFTB3-D3H4X).[7,55] The second term is the desolvation free energy change upon binding, evaluated at the COSMO level[56] from the PM6/COSMO calculation using MOPAC with default parameters.[57] To make the calculations faster, we truncated the systems by defining a sphere of 12 Å (roughly 3,000 atoms) around the aligned ligand poses as a region representing the binding site. This region was treated by SQM and was the same for all complexes. We had previously demonstrated at the PM6-D3H4 level on four P–L complexes that results obtained on similarly truncated or full sized systems show nearly identical behavior.[6]

The AMBER^{AM1}/ GB, AMBER^{HF}/ GB, AMBER^{HF}/ PB, AMBER^{HF}/ COSMO scoring functions combine the ff03/GAFF[27,53] force fields with the GB (IGB = 1 option),[58] PB[59] and COSMO[56] implicit solvent models. GB and PB were evaluated by AMBER, while COSMO by PM6/COSMO calculation in MOPAC. The AMBER^{AM1} and AMBER^{HF} stand for the RESP procedure at AM1-BCC and HF/6-31G* levels, respectively.

AutoDock4 (v.4.2.6)[24] and AutoDock Vina 1.1.2:[25] Grid box centers were placed on the center of mass of the crystal ligand heavy

atoms for each protein structure. The box size was chosen as 20 Å. Ligands and proteins were prepared using Ligand4.py and receptor4.py preparation scripts in AutoDockTools-1.5.6[24] with default settings. In AutoDock4, we used 0.375 Å grid point spacing and AutoDock4$_{Zn}$ improved force field for zinc metalloproteins.[26]

GOLD: We used ChemPLP[19] GoldScore,[20] Chemscore,[21] and ASP[22] scoring functions implemented in GOLD Suite v5.4.1 software for rescoring. Binding site origins were defined as the same coordinates for AutoDock Vina and AutoDock4. The radius was set to 20 Å. Zinc coordination geometries were set to tetrahedral for Zn-containing proteins. All other settings were kept as software defaults.

DOCK 6:[23] The size of the box was obtained by adding an extra margin of 15 Å in all 6 directions. A grid spacing of 0.3 Å was used. The cutoff for nonbonded interactions was not used. We used AMBER parameters. For the ligands, we used AM1-BCC RESP partial atomic charges. DOCK 6 was also used to generate docked poses. For the SQM/COSMO rescoring, we considered five top-ranked poses with RMSD of heavy atoms up to 2.5 Å.

Acknowledgements

This work was supported by research project RVO 61388963 of the Czech Academy of Sciences. We acknowledge the financial support of the Czech Science Foundation from grant no. P208/16-11321Y (J.Ř.), and GA15-05677S (J.B., M.L., J.F.). A.P., J.F., M.L. P.H. and S.M.E. acknowledge the support from European Regional Development Fund; OP RDE; Project: "Chemical biology for drugging undruggable targets (ChemBioDrug)" (No. CZ.02.1.01/ 0.0/ 0.0/ 16_019/ 0000729). This work was also supported by the Ministry of Education, Youth and Sports from the Large Infrastructures for Research, Experimental Development and Innovations project "IT4Innovations National Supercomputing Center—LM2015070" as well as from projects LO1304 (JB, LV, PR, VV), LO1305 (PH), and LK11205 (LV, VV).

Keywords: binding affinities · crystallography · lead optimization · quantum mechanics · scoring function

[1] J. J. Irwin, B. K. Shoichet, *J. Med. Chem.* **2016**, *59*, 4103 – 4120.
[2] U. Ryde, P. Söderhjelm, *Chem. Rev.* **2016**, *116*, 5520 – 5566.
[3] L. Berg, B. K. Mishra, C. D. Andersson, F. Ekström, A. Linusson, *Chem. Eur. J.* **2016**, *22*, 2672 – 2681.
[4] K. Raha, K. M. Merz, Jr., *J. Med. Chem.* **2005**, *48*, 4558 – 4575.
[5] P. Chaskar, V. Zoete, U. F. Rohrig, *J. Chem. Inf. Model.* **2014**, *54*, 3137 – 3152.
[6] A. Pecina, R. Meier, J. Fanfrlik, M. Lepsik, J. Rezac, P. Hobza, C. Baldauf, *Chem. Commun.* **2016**, *52*, 3312 – 3315.
[7] A. Pecina, S. Haldar, J. Fanfrlik, R. Meier, J. Řezáč, M. Lepsik, P. Hobza, *J. Chem. Inf. Model.* **2017**, *57*, 127 – 132.
[8] P. Chaskar, V. Zoete, U. F. Röhrig, *J. Chem. Inf. Model.* **2017**, *57*, 73 – 84.
[9] A. Cousido-Siah, F. X. Ruiz, J. Fanfrlík, J. Geménez-Dejoz, A. Mitschler, M. Kamlar, J. Veselý, H. Ajani, X. Perés, J. Farrés, P. Hobza, A. D. Podjarny, *ACS Chem. Biol.* **2016**, *11*, 2693 – 2705.
[10] J. Fanfrlik, P. S. Brahmkshatriya, J. Rezac, A. Jilkova, M. Horn, M. Mares, P. Hobza, M. Lepsik, *J. Phys. Chem. B* **2013**, *117*, 14973 – 14982.
[11] J. J. P. Stewart, *J.Mol. Model.* **2007**, *13*, 1173 – 1213.
[12] J. Rezac, P. Hobza, *Chem. Rev.* **2016**, *116*, 5038 – 5071.
[13] J. Fanfrlik, A. K. Bronowska, J. Rezac, O. Prenosil, J. Konvalinka, P. Hobza, *J. Phys. Chem. B* **2010**, *114*, 12666 – 12678.
[14] M. Lepsik, J. Rezac, M. Kolar, A. Pecina, P. Hobza, J. Fanfrlik, *ChemPlusChem* **2013**, *78*, 921 – 931.
[15] M. Hylsová, B. Carbain, J. Fanfrlík, L. Musilová, S. Haldar, C. Köprülüoğlu, H. Ajani, P. S. Brahmkshatriya, R. Jorda, V. Kryštof, P. Hobza, A. Echalier, K. Paruch, M. Lepšík, *Eur. J. Med. Chem.* **2017**, *126*, 1118 – 1128.
[16] H. Ajani, A. Pecina, S. M. Eyrilmez, J. Fanfrlik, S. Haldar, J. Rezac, P. Hobza, M. Lepsik, *ACS Omega* **2017**, *2*, 4022 – 4029.
[17] E. Yuriev, M. Agostino, P. A. Ramsland, *J. Mol. Recognit.* **2011**, *24*, 149 – 164.
[18] A. V. Sulimov, D. C. Kutov, E. V. Katkova, V. B. Sulimov, *Adv. in Bioinformatics* **2017**, 7167691.
[19] O. Korb, T. Stutzle, T. E. Exner, *J. Chem. Inf. Model.* **2009**, *49*, 84 – 96.
[20] G. Jones, P. R. Willett, C. Glen, A. R. Leach, R. Taylor, *J. Mol. Biol.* **1997**, *267*, 727 – 748.
[21] M. D. Eldridge, C. W. Murray, T. R. Auton, G. V. Paolini, R. P. Mee, *J. Comput. Aided Mol. Des.* **1997**, *11*, 425 – 445.
[22] W. T. M. Mooij, M. L. Verdonk, *Proteins* **2005**, *61*, 272 – 287.
[23] P. T. Lang, S. R. Brozell, S. Mukherjee, E. F. Pettersen, E. C. Meng, V. Thomas, R. C. Rizzo, D. A. Case, T. L. James, I. D. Kuntz, *RNA* **2009**, *15*, 1219 – 1230.
[24] G. M. Morris, *J. Comput. Chem.* **2009**, *30*, 2785 – 2791.
[25] O. Trott, A. J. Olson, *J. Comput. Chem.* **2010**, *31*, 455 – 461.
[26] D. Santos-Martins, S. Forli, M. J. Ramos, A. J. Olson, *J. Chem. Inf. Model.* **2014**, *54*, 2371 – 2379.
[27] Y. Duan, C. Wu, S. Chowdhury, M. C. Lee, G. Xiong, W. Zhang, R. Yang, P. Cieplak, R. Luo, T. Lee, *J. Comput. Chem.* **2003**, *24*, 1999 – 2012.
[28] J. Brynda, P. Mader, V. Šícha, M. Fábry, K. Poncová, M. Bakardiev, B. Grüner, P. Cígler, P. Řezáčová, *Angew. Chem. Int. Ed.* **2013**, *52*, 13760 – 13763; *Angew. Chem.* **2013**, *125*, 14005 – 14008.
[29] V. M. Krishnamurthy, G. K. Kaufman, A. R. Urbach, I. Gitlin, K. L. Gudiksen, D. B. Weibel, G. M. Whitesides, *Chem. Rev.* **2008**, *108*, 946 – 1051.
[30] C. T. Supuran, *Nat. Rev. Drug. Discovery* **2008**, *7*, 168 – 181.
[31] J. J. Irwin, T. Sterling, M. M. Mysinger, E. S. Bolstad, R. G. Coleman, *J. Chem. Inf. Model.* **2012**, *52*, 1757 – 1768.
[32] R. G. Khalifah, *J. Biol. Chem.* **1971**, *246*, 2561 – 2573.
[33] D. A. Pearlman, P. S. Charifson, *J. Med. Chem.* **2001**, *44*, 3417 – 3423.
[34] S. G. Krimmer, J. Cramer, M. Betz, V. Fridh, R. Karlsson, A. Heine, G. Klebe, *J. Med. Chem.* **2016**, *59*, 10530 – 10548.
[35] M. A. Pinard, C. D. Boone, B. D. Rife, C. T. Supuran, R. McKenna, *Bioorg. Med. Chem.* **2013**, *21*, 7210 – 7215.
[36] U. Mueller, N. Darowski, M. R. Fuchs, R. Förster, M. Hellmig, K. S. Paithankar, S. Pühringer, M. Steffien, G. Zocher, M. S. Weiss, *J. Synchrotron Radiat.* **2012**, *19*, 442 – 449.
[37] W. Kabsch, *Acta Crystallogr. Sect. D* **2010**, *66*, 125 – 132.
[38] W. Kabsch, *Acta Crystallogr. Sect. D* **2010**, *66*, 133 – 144.
[39] P. Mader, J. Brynda, R. Gitto, S. Agnello, P. Pachl, C. T. Supuran, A. Chimirri, P. Rezacova, *J. Med. Chem.* **2011**, *54*, 2522 – 2526.
[40] R. Ahlrichs, M. Bar, M. Haser, H. Horn, C. Kolmel, *Chem. Phys. Lett.* **1989**, *162*, 165 – 169.
[41] P. Jurečka, J. Černý, P. Hobza, D. Salahub, *J. Comput. Chem.* **2007**, *28*, 555 – 569.
[42] M. Winn, C. C. Ballard, K. D. Cowtan, E. J. Dodson, P. Emsley, P. R. Evans, R. M. Keegan, E. B. Krissinel, A. G. W. Leslie, A. McCoy, S. J. McNicholas, G. N. Murshudov, N. S. Pannu, E. A. Potterton, H. R. Powell, R. J. Read, A. Vagin, K. S. Wilson, *Acta. Cryst. D* **2011**, *67*, 235 – 242.
[43] P. Emsley, B. Lohkamp, W. Scott, K. Cowtan, *Acta Crystallogr. Sect. D* **2010**, *66*, 486 – 501.
[44] A. A. Vagin, R. S. Steiner, A. A. Lebedev, L. Potterton, S. McNicholas, F. Long, G. N. Murshudov, *Acta Crystallogr. Sect. D* **2004**, *60*, 2184 – 2195.
[45] V. B. Chen, W. B. Arendall, J. J. Headd, D. A. Keedy, R. M. Immormino, G. J. Kapral, L. W. Murray, J. S. Richardson, D. C. Richardson, *Acta Crys. D* **2010**, *66*, 12 – 21.
[46] Schrödinger, LLC. The PyMOL Molecular Graphics System **2010**, version 1.3r1.
[47] J. M. Word, S. C. Lovell, J. S. Richardson, D. C. Richardson, *J. Mol. Biol.* **1999**, *285*, 1735 – 1747.
[48] C. E. A. F. Schafmeister, W. S. Ross, V. Romanovski, V., LeaP, University of California, San Francisco, **1995**.
[49] D. A. Case, T. A. Darden, T. E. Cheatham, C. L. Simmerling III, J. Wang, R. E. Duke, R. Luo, M. Crowley, R. C. Walker, W. Zhang, K. M. Merz, B. Wang, S. Hayik, A. Roitberg, G. Seabra, I. Kolossváry, K. F. Wong, F. Paesani, J. Vanicek, X. Wu, S. R. Brozell, T. Steinbrecher, H. Gohlke, L. Yang, C. Tan, J. Mongan, V. Hornak, G. Cui, D. H. Mathews, M. G. Seetin, C. Sagui,

V. Babin, P. A. Kollman, AMBER 10, University of California, San Francisco, **2008**.
[50] E. F. Pettersen, T. D. Goddard, C. C. Huang, G. S. Couch, D. M. Greenblatt, E. C. Meng, T. E. Ferrin, *J. Comput. Chem.* **2004**, *25*, 1605 – 1612.
[51] K. Kanamori, J. D. Roberts, *Biochemistry* **1983**, *22*, 2658 – 2664.
[52] A. Pecina, M. Lepšík, J. Řezáč, J. Brynda, P. Mader, P. Řezáčová, P. Hobza, J. Fanfrlík, *J. Phys. Chem. B* **2013**, *117*, 16096 – 16104.
[53] J. Wang, R. M. Wolf, J. W. Caldwell, P. A. Kollman, D. A. Case, *J. Comput. Chem.* **2004**, *25*, 1157 – 1174.
[54] A. Jakalian, B. L. Bush, D. B. Jack, C. I. Bayly, *J. Comput. Chem.* **2000**, *21*, 132 – 146.
[55] V. M. Miriyala, J. Rezac, *J. Comput. Chem.* **2017**, *38*, 688 – 697.
[56] A. Klamt, G. Schuurmann, *Perkin Trans. 2* **1993**, *0*, 799 – 805.
[57] J. J. P. Stewart, MOPAC 2016, Stewart Computational Chemistry, Colorado Springs, CO: **2016**.
[58] V. Tsui, D. A. Case, *Biopolymers* **2001**, *56*, 275 – 291.
[59] C. Tan, Y. H. Tan, R. Luo, *J. Phys. Chem. B* **2007**, *111*, 12263 – 12274.

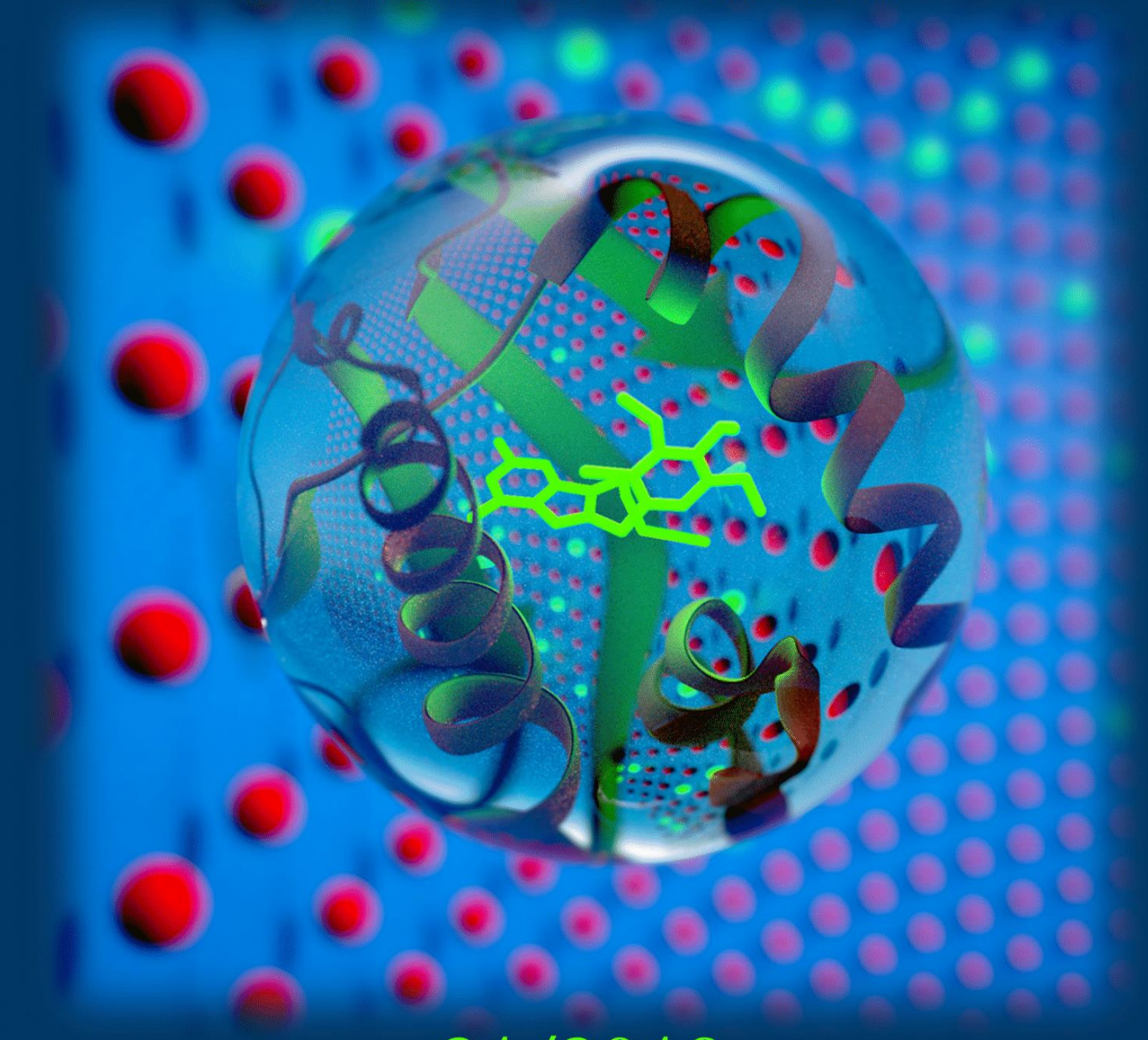
Manuscript received: October 11, 2017
Accepted manuscript online: January 8, 2018
Version of record online: February 19, 2018

Publication D

A EUROPEAN JOURNAL

CHEMPHYSCHEM

OF CHEMICAL PHYSICS AND PHYSICAL CHEMISTRY



21/2019

Front Cover:

S. M. Eyrilmez, C. Köprülüoğlu et al.

Impressive Enrichment of Semiempirical Quantum Mechanics-Based Scoring Function: HSP90 Protein with 4541 Inhibitors and Decoys

A Journal of

ChemPubSoc Europe

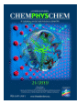
WILEY-VCH

www.chemphyschem.org

ChemPubSoc Europe

DOI: 10.1002/cphc.201900628

CHEMPHYSCHEM Articles



Impressive Enrichment of Semiempirical Quantum Mechanics-Based Scoring Function: HSP90 Protein with 4541 Inhibitors and Decoys

Saltuk M. Eyrilmez[+],[a, b] Cemal Köprülüoğlu[+],[a, b] Jan Řezáč,[a] and Pavel Hobza*[a, b]

This paper describes the excellent performance of a newly developed scoring function (SF), based on the semiempirical QM (SQM) PM6-D3H4X method combined with the conductor-like screening implicit solvent model (COSMO). The SQM/COSMO, Amber/GB and nine widely used SFs have been evaluated in terms of ranking power on the HSP90 protein with 72 biologically active compounds and 4469 structurally similar decoys. Among conventional SFs, the highest early and overall enrichment measured by EF_1 and AUC% obtained using single-scoring-function ranking has been found for Glide SP and GoldASP SFs, respectively (7, 75% and 3, 76%). The performance of other standard SFs has not been satisfactory, mostly even decreasing below random values. The SQM/COSMO SF, where P–L structures were optimised at the advanced Amber level, has resulted in a dramatic enrichment increase (47, 98%), almost reaching the best possible receiver operator characteristic (ROC) curve. The best SQM frame thus inserts about seven times more active compounds into the selected dataset than the best standard SF.

1. Introduction

The determination of the structure and properties of protein-ligand (P–L) complexes is a key task in structure-based drug design.[1] To this end, numerous docking and scoring functions (DF, SF) have been developed and tested. Besides classical approach to docking/scoring based on empirical, knowledge- or physics-based methods,[2] a quantum mechanics (QM) approach was pioneered by Merz et al.[3] The former approaches, dependent on the existence of a sufficiently broad training set, are based on molecular mechanics (MM) methods (or their simplification), whereas the latter approach utilises the 'objective' QM method. The main advantage of QM methods is the fact that they cover quantum effects (e.g. charge transfer or σ-hole binding), which might play an important role in P–L interactions. Due to the size of P–L complexes, mostly semiempirical QM (SQM) methods as AM1,[4] PM3,[5] PM6,[6] PM7[7] and DFTB3[8] are applied. None of these methods is, however, directly suitable for the investigation of noncovalent complexes.[9] For this purpose, correction terms ensuring the proper description of dispersion, electrostatic and σ-hole interactions should be included. Throughout the present paper, the advanced D3H4X[10] correction term has been used in combination with the PM6 method (PM6-D3H4X). The method can be applied to extended P–L complexes with several thousand atoms because the PM6 method can be combined with the MOZYME linear scaling algorithm implemented in MOPAC.[11] P–L complexes exist in a solvent environment, which affects their structure and properties. To model the solvent, we used the conductor-like screening implicit solvent model (COSMO).[12] The binding free energy between the protein and ligand in a solvent is approximated by the score [Eq. (1)]:

$$SCORE = \Delta E_{int} + \Delta\Delta G_{solv} + \Delta G^{'w}_{conf}(P) + \Delta G^{'w}_{conf}(L) + T\Delta S_{int} \quad (1)$$

expressed as the sum of the gas-phase P–L interaction energy (the first term), the change of solvation/desolvation free energy upon complex formation (the second term), the change of the conformation 'free' energies of the protein and ligand (the third and fourth terms), and the entropy change upon binding (the fifth term).[13,14] The first two terms, having the opposite sign (the first term is always stabilising while the second one is destabilising), are clearly dominant and SQM/COSMO-based SFs are mostly based only on them.[15]

The evaluation of the performance of DFs and SFs is mainly based on the estimation of their sampling and ranking power, where the first refers to the ability of SF to predict the position of a native ligand correctly. We have recently investigated the sampling power of several widely used classical SFs as well as quantum mechanics-based SFs developed in our laboratory. PM6 and DFTB3 SQM methods were combined with PM6/COSMO, systematically based on the PM6 characteristics. Four[16] and seventeen[17] different P–L complexes were studied, and the SQM/COSMO SFs clearly outperformed all classical SFs. Slight improvement was achieved when the less empirical but considerably more expensive DFTB3 method was applied.[18–20] Ranking power, describing the ability of SF to rank different ligands of the same target protein (based on binding affinities), is a more difficult task. In the first study,[21] we investigated a

[a] S. M. Eyrilmez,[+] C. Köprülüoğlu,[+] Dr. J. Řezáč, Prof. P. Hobza
Institute of Organic Chemistry and Biochemistry of the Czech Academy of Sciences
Flemingovo nam. 2, 16610 Prague 6 (Czech Republic)
E-mail: pavel.hobza@uochb.cas.cz
[b] S. M. Eyrilmez,[+] C. Köprülüoğlu,[+] Prof. P. Hobza
Regional Centre of Advanced Technologies and Materials
Palacký University, 77146 Olomouc (CzechRepublic)
[+] These authors contributed equally to this work

relatively easy case, where the structures of the complexes of carbonic anhydrase II with ten different ligands were known, what reduces the problem to the determination of binding free energy. In addition, in this case the SQM/COSMO SF provided much better results than ten different classical SFs.

Structure-based virtual screening can only be successful if the method can reliably predict the geometry of the P–L complex (the binding mode) and the SF used provides reliable ranking. It is now widely accepted that it is beyond the ability of the currently used SFs to meet both requirements. As shown above, the SQM/COSMO-based SF was more successful in both these scenarios. The present study has applied this methodology (with further extensions) to a larger and more diverse set of P–L complexes, the HSP90 protein and its about 5000 compounds from the DUD–E database. The HSP90 (heat shock protein 90) is a protein that stabilises various growth receptors[22] and some signalling molecules[23] required for the survival of cancer cells.

The structures of P–L complexes were not known and were thus determined by docking. The PM6/COSMO and Amber/GB SFs and nine widely used classical SFs were evaluated in terms of docking and scoring. The SQM/COSMO and Amber/GB SFs were only applied for ranking, and the respective poses were generated by different classical SFs. Since the biological activities of all ligands and decoys are not known in detail and are characterised only in terms 'active' and 'inactive', direct correlation between activities and theoretical scores is impractical. It should be added here that even if ligand affinities are known, correlation between them and calculated scores is difficult, sometimes[24] even denoted as being 'beyond the current methods'. To describe the ability of the method to distinguish between active and inactive ligands, we use the enrichment factor, a quantity that distinguishes known ligands from decoys.

1.1. Strategy

The knowledge of the native structure of the P–L complex is crucial for the estimation of the biological activity of a ligand. If the experimental structure of the P–L complex is missing, it is possible to use the theoretical structure determined by the gradient optimisation of binding free energy (Eq. 1). Such an approach is not only CPU-time demanding but, and this is more serious, it mostly leads to a local minimum at the free energy landscape. There is an enormous number of the local minima for P–L complexes, and it is clearly impractical to search the whole landscape at the SQM level. We have chosen an alternative route; DFs are used to generate a large number of poses, which will be scored in the next step by means of SQM/COSMO SF. We are aware that a reliable identification of a native P–L pose with a single DF is a difficult task. Therefore, we have intentionally used nine different DFs to increase the possibility of finding a native binding pose. The disadvantage of this procedure is clear – the SQM/COSMO calculations should be performed for thousands of P–L structures. Therefore, the SQM/COSMO SF should be computationally as efficient as possible. For that reason, we have introduced a semiempirical quantum mechanics-based virtual screening frame, which eliminates redundant poses and produces high-quality structures to increase the efficiency and applicability of demanding PM6/COSMO calculations. The structures of P–L complexes determined by DFs should be optimised at the molecular mechanics level. This work has used the AMBER[25] biomolecular simulation package for geometry optimisations (MM_N). To increase the reliability of subsequent SQM scoring, restrained AMBER optimisations have been applied as well. The bond length and bond angle values have been taken from PM6 optimised compound structures (MM_A). In the present study, we have used test compounds taken from 'A Database of Useful Decoys: Enhanced' (DUD–E)[24] for the HSP90 target. This set contains 4850 decoys, 25 experimental inactive compounds with similar physical properties (e.g. molecular weight, calculated logP) but dissimilar 2-D topology, and 88 actives with known experimental binding affinities. The ligands included 19 macrocycle-containing molecules. Since considerable effort might be needed to rationalise the protocol, these compounds were excluded from the actives in the first place.[26] 406 of the decoys were also not considered due to the computational reasons.

1.2. Scoring

Within the present scoring framework, the score was approximated without the entropy change [Eq. (2)]:

$$SCORE = \Delta E_{int} + \Delta\Delta G_{solv} + \Delta G^{'w}_{conf}(P^{H\ atoms}) + \Delta G^{'w}_{conf}(L) \qquad (2)$$

where the first, second and fourth terms were identical to these in Equation (1), while in the third term only hydrogens were considered in optimisation.

Three types of scoring were applied: MM scoring using Amber/GB SF and two types of SQM scoring based on SQM/COSMO SFs, denoted as SQM_1 and SQM_2, where MM_N and MM_A optimised structures were utilised.

In the case of multiple protonation states, each state was scored individually and the one with the minimum score was used for enrichment analysis.

MM scoring: The scoring scheme shown in Equation (2) was applied for MM scoring using the MM_N optimised structures of the complex and ligands. Since the application of MM_A optimisations deteriorated the Amber energies, the MM scoring over MM_A-optimised structures was not performed.

SQM scoring: The key point for any SQM scoring in this virtual screening study was to decrease the redundant poses before processing them at the PM6/COSMO level. To achieve this, we first applied RSMD clustering with a 1 Å cut-off to eliminate similar poses produced by MM_N optimisations. Representative poses were selected as the MM_N minimum complex structures. The complexes for the subsequent SQM scoring were selected on the basis of MM_N optimisations (the complexes within the 10 kcal/mol energy interval were taken into consideration).

Two types of SQM scoring (SQM_1 and SQM_2) using the same SQM/COSMO SF but different optimisation schemes, denoted as MM_N and MM_A, were considered. We have used Cuby4[27] software to automate our fragmentation, optimisation and energy calculation protocols.

1.3. Analysis

For each of these SFs, the score of all ligand poses binding to the respective target protein was calculated, ranked and plotted. The performance of the scoring functions was evaluated based on the analysis of the enrichment factor (EF) and receiver operator characteristic (ROC) plots.[28] The accuracy of virtual screening was evaluated using EF. Calculated score values were ranked, and EF was defined as [Eq. (3)]:

$$EF_{subset} = (ligand_{selected}/N_{subset})/(ligand_{total}/N_{total}) \quad (3)$$

where $ligand_{total}$ is the number of known ligands with activity against the target, N_{total} is the number of all compounds in the dataset, $ligand_{selected}$ is the number of found ligands in a given subset, and N_{subset} is the total number of the compounds in the subset. EF_{subset} provides information about the number of the true positives among the decoys in the given subset in comparison with a random selection.[29] Generally, the top part of the library of the ranked compounds was used for further evaluation and was strongly dependent on the initial library size. The size might range from 0.1% to 10% and it was considered as 1% in the present study.[30]

ROC curves were obtained by plotting sensitivity (Se) and specificity (Sp), where:

$$Se_{subset} = (ligand_{selected}/Ligands_{total}) \times 100$$

$$Sp_{subset} = [(Decoys_{total} - Decoys_{selected})/Decoys_{total}] \times 100$$

The ROC curves were plotted as (100%–Sp%) (i.e. % of selected decoys) versus Se% (i.e. % of selected active compounds).[31]

The AUC was defined as the area under a ROC curve. It is simply the probability that a randomly chosen active has a higher score than a randomly chosen inactive. In other words, the AUC is the average of this property over all inactive fractions.[28]

The results were also supported by the pROC AUC values, which focus on early enrichment.[32][33] pROC AUC values for random enrichment were determined as follows [Eq. (4)]:

$$\lim_{a\to 0}\int_a^1(-\log_{10}X)dX = \frac{-1}{\log 10}\lim_{a\to 0}\int_a^1(\log X)dX = 0.434\lim_{a\to 0}\{X - X\log X\}|0pt1a = 0.434 \quad (4)$$

Table 1. The ROC enrichment factors (EF1), AUC (in %) and pROC AUC obtained for single-docking-function ranking (DF).

DF	EF1	AUC [%]	pROC AUC
AD4	1	49	0.383
VINA	0	30	0.192
SMINA	0	34	0.224
GlideSP	7	75	0.880
GlideXP	4	71	0.730
ASP	3	76	0.787
Gscr	0	60	0.488
Cscr	0	34	0.270
PLP	1	51	0.383

2. Results and Discussion

Table 1 shows the performance of nine different SFs where a single SF was used for both scoring and ranking. The analysis is based on EF_1, AUC and pROC AUC characteristics determined for the average property over all inactive fractions.

Evidently, the results are not satisfactory, especially concerning the early-stage enrichment (EF_1 values). The EF_1 values of six out of nine SFs were equal to or below random values ($EF_1 = 1$) and only SP, XP and ASP SFs provided EF_1 above this limit. The overall performance measured by AUC values was slightly better – the AUC values of five SFs were below random values (50%), whereas SP, XP, ASP as well as Gscr were above them, while PLP equalled the random performance. The best among the single-scoring-function ranking SFs were GlideSP, GlideXP and ASP, having the highest EF_1, AUC% and pROC AUC characteristics (7, 75%, 0.880; 4, 71%, 0.730 and 3, 76%, 0.787). Note that GlideSP exhibits higher early enrichment, which is more important for drug-design purposes, while ASP has better overall performance. The combined performance of all SFs was, however, poor. In all the cases, the single-scoring-function ranking SFs had been applied. It was thus necessary to decide whether the problem originated in incorrect structures determined by docking or in the incorrect score determined by ranking. The question was to be answered in the next step by a combined study using different SFs for scoring and ranking. It is known that rescoring with the different docking functions can improve the enrichment significantly.[34] Each scoring function has been therefore sampled extensively to fill the active pocket as complete as possible. We collected 100 poses from each docking software. Comparing results presented in the Tables 1 and 2 we found that none of empirical functions (cf. Table 1) could reach the respective results presented in the Table 2.

Table 2 summarises the enrichment where docking was made by a single standard DF while ranking was performed by MM, SQM_1 and SQM_2 SFs. A comparison with the corresponding values from Table 1 clearly shows the dramatic improvement when binding free energies (SCOREs) have been evaluated at the MM and both SQM levels. The SQM_1 results will be discussed first. All combinations of SQM_1 ranking with poses generated by different DFs have provided the enrichment values considerably above the random values. The highest early and overall enrichment was obtained for Gscr, SMINA, AD4, GlideSP and PLP structures. Considering the pROC AUC values

Table 2. The ROC enrichment factors (EF_1), AUC (in %) and pROC AUC obtained for SQM_2//DF (a combination of scoring and docking; the P–L structures were optimised with the MM_A method), SQM_1//DF (the P–L structures were optimised with the MM_N method) and MM//SF.

SF//DF	EF1	AUC [%]	pROC AUC
SQM_2//AD4	40	91	2.104
SQM_1//AD4	25	70	1.262
MM//AD4	15	86	1.304
SQM_2//VINA	42	93	2.052
SQM_1//VINA	27	67	1.277
MM//VINA	13	83	1.208
SQM_2//SMINA	37	93	1.997
SQM_1//SMINA	31	69	1.371
MM//SMINA	17	84	1.343
SQM_2//SP	34	81	1.670
SQM_1//SP	15	82	1.368
MM//SP	15	76	1.277
SQM_2//XP	32	85	1.710
SQM_1//XP	14	65	0.872
MM//XP	23	83	1.493
SQM_2//ASP	31	93	1.877
SQM_1//ASP	24	66	1.093
MM//ASP	18	91	1.418
SQM_2//Gscr	44	97	2.329
SQM_1//Gscr	31	74	1.350
MM//Gscr	14	90	1.352
SQM_2//Cscr	29	89	1.757
SQM_1//Cscr	19	62	0.880
MM//Cscr	14	82	1.229
SQM_2//PLP	31	95	2.096
SQM_1//PLP	27	68	1.193
MM//PLP	3	88	1.333

the VINA DF shows a better performance than PLP even though both DFs have similar EF_1 and AUC results. Glide SP exhibits the best overall performance but low EF_1 values. On the other hand, the pROC AUC result was the second best. The SP DF thus provides early stage success as demonstrated by pROC AUC value and the best overall performance (see AUC% value in the Table 2). For drug discovery, as mentioned above, early evaluation is more important; therefore, preference should be given to SMINA, SP, Gscr and VINA DFs combined with SQM/COSMO SF. Much better enrichment performed by combined SQM_1 ranking and DF docking provides evidence that all standard SFs have problems with the determination of binding free energies while their geometries are reliable. Surprisingly high enrichment, especially an overall one, was obtained when MM SF was applied. The best results in the overall performance were obtained with ASP and Gscr DFs. The very good performance of MM is promising for the future investigation of extended P–L complexes, because MM is much less CPU-time demanding. It should be noted that the SQM_1 results discussed above were obtained with the P–L structures optimised with the standard MM_N method. On the other hand, the SQM_2 results in Table 2 were obtained with the P–L structures optimised with the MM_A method. Evidently, this systematically resulted in significantly higher enrichment. Considering the SQM_1 values, only the Gscr structures provided AUC values higher than 70%. When the SQM_2 values were considered, five of the SF structures exceeded 90% limit and the PLP and Gscr values even reached 95 and 97%, respectively.

A similar dramatic increase was found for the early enrichment, where five out of nine EF_1 values were higher than 31 (this value was not exceeded by any EF_1 for SQM_1 and MM) and the highest EF_1 was detected for SQM_2//Gscr (44), beside this, pROC AUC value is five times better than the random. A comparison of the entries in the Table 2 clearly shows that high enrichment is only obtained if reliable binding modes are used. Evidently, the poses generated by docking are not sufficiently accurate and significant enrichment increase is only obtained after their re-optimisation at the MM_A level. The question arises whether comparable results can be expected for other proteins as well. The necessary condition for it is the generation of reliable structures. There is no reason to expect that a DF that has generated reliable structures for some protein will also succeed for another one. To make the method more robust, it is thus beneficial to use more DFs for the generation of ligand poses. To test this approach, we have collected ligand poses from all the DFs considered in the present paper; the subsequent ranking was performed with SQM_1, SQM_2 and MM SFs. The consideration of the poses from all DFs provided an enrichment increase when SQM_1 and SQM_2 SFs were used (cf Table 3). When SQM_1, SQM_2 and MM methods were applied, the

Table 3. The ROC enrichment factors (EF_1), AUC (in %) and pROC AUC obtained for SQM_2//ALL (a combination of scoring and docking; the P–L geometries from all SFs were optimised with the MM_A method), SQM_1//ALL (a combination of scoring and docking; the P–L geometries from all SFs were optimised with the MM_N method) and MM//ALL.

SF//DF	EF1	AUC [%]	pROC AUC
SQM_2//ALL	47	98	2.477
SQM_1//ALL	32	75	1.426
MM//ALL	10	92	1.395

AUC values reached 75%, 98% and 92%, respectively, and highest enrichment was achieved when the SQM_2 method was used. The highest AUC values obtained with the same methods where only the structures generated by a single DF were used equalled 74%, 97% and 91%, respectively. The EF_1 values for SQM_1, SQM_2 and MM methods (where the structures of all DFs were used) amounted to 32, 47, 10, and, again, the highest EF_1 was obtained for SQM_2. The pROC AUC value is 2.477 which means the performance is six times better than the random case. When only the structures generated by a single DF were used EF_1 equalled 31, 44 and 23, respectively. The consideration of the structures from all DFs improved early and overall enrichment for SQM_1 and SQM_2, the effect was not dramatic. We have seen a decline of EF_1 for MM but improved overall performance. The reason for the decline might be due to the energy ranking of all the structures generated by all DFs. The above-mentioned results are valid for the present protein. For different targets situation might be different and the use of structures from more DFs is thus recommendable.

3. Conclusions

The enrichment obtained with single-scoring-function ranking was low for all nine conventional SFs. Several SFs provided enrichment even below the random-value limit. Only four SFs (SP, XP, ASP and Gscr) provided enrichment above random values. Evidently, no single SF succeeds in both docking/scoring and ranking.

The enrichment increased when SQM_1, SQM_2, and MM ranking was determined for poses generated by standard SFs. This gives evidence that standardly used SFs provide reliable poses but fail for ranking. On the other hand, SQM_1, SQM_2 as well as MM SFs yield reliable ranking.

A significant enrichment increase was achieved when P–L structure optimisation was performed within the SQM_2 frame. The enrichment (AUC) obtained by five out of nine SFs exceeded 90%, and the PLP and Gscr AUC values even reached 95% and 97%. Impressive enrichment in terms of both EF_1 and AUC resulted when the Gscr, AD4 and PLP structures were re-optimised at the MM_A level (44, 97%, 2.329; 40, 91%, 2.104 and 31, 95%, 2.096 respectively). Using the PM6 parameters in the MM treatment improves the geometry of the ligand, what leads to better geometries of the P–L complex and, consequently, to higher enrichment. The consideration of all poses provided an enrichment increase for SQM_1 and SQM_2 methods, EF_1, AUC and pROC AUC values rose to 32, 75%, 1.426 and 47, 98%, 2.477 respectively.

The overall enrichment after MM_A application to P–L structure optimisation was very close to the best ROC AUC limits.

The standard approach to virtual screening is based on the use of single-scoring-function ranking. The highest enrichment in EF_1, AUC and pROC AUC (7, 75%, 0.880 and 3, 76%,0.787 respectively) was obtained using the GlideSP and ASP SFs. Passing from the best single-scoring-function ranking to the advanced SQM treatment led to a dramatic increase. A combination of the SQM SF with P–L optimisation using MM_A provided impressively high EF_1, AUC and pROC AUC values (47, 98%, 2.477). The enrichment factor obtained included 34 (out of 72) experimentally active structures in the subset. This means that nearly 50% of actives are found in 1% of the whole dataset. In other words, the present SQM_2 SF frame inserts about seven times more active compounds into the selected dataset and three times better pROC performance than the best SF. This clearly demonstrates the impressive performance of the SQM_2 frame in both early and overall enrichment. The values of the overall enrichment are close to the best ROC curve. We are certainly aware that all these findings are based on the investigation of a single protein. Intensive work in our laboratory is currently being performed for targets from other protein families.

The above-mentioned findings clearly demonstrate the advantage of using SQM SFs over the standard ones. We believe that despite higher CPU demands, the wider application of SQM SFs could be beneficial not only for structure-based drug design but also for related applications.

Figure 1 shows the visualisation of docking results using a novel Post Dock tool[35] implemented in MOE Software.[36] Six (A, C, D, E, F, G) out of nine DFs provided binding modes close to the crystal pose. Their transparencies were high which means the respective DFs provided correct poses with the worse score. In another word, they are good in sampling but failed in ranking. On the contrary, figures B, H and I demonstrated that the individual DFs failed to generate the correct binding modes. As it seen in the Figure 1J SQM provided less transparent yellow colour which means that the binding mode totally matches with the crystal pose. It implies that SQM SF was able to select the crystal pose with the highest score. These results show that increasing the number of poses for the individual DFs was important for finding the crystal pose. For instance, in the case of SQM SF (Figure 1J) the binding mode with the best score fully agreed with the crystal pose while AD4 DF (Figure 1A) found the best agreement with the crystal pose for pose number 90 having the worse score. Evidently that the use of SQM SF is required for obtaining both successful sampling and ranking.

Computational Section

Compound Preparation

The compounds were downloaded in the SMILES format and prepared using the LigPrep module with an Optimised Potentials for Liquid Simulations (OPLS3e)[37] force field. Their ionisation states were generated at pH 7.0 ± 2.0 using Epik[38] in LigPrep.[39] Specific chiralities were retrained during the ligand preparations. The structures generated within the state penalty value of 0–1 were saved for docking calculations.

Protein Preparation

The process of protein preparation requires special care in physics-based SFs.[16] Protein has been downloaded from the Protein Databank[40] with the 1UYG PDB code.[41] We have decided to keep three conserved water molecules (W2121, W2123 and W2236). To implement the selection, we first aligned the PDB structures with 100% sequence similarity to 1UYG. According to the Ref.[42] we selected the intersection set of the most favourable water molecules (W1, W3,W4). Hydrogens of the protein were added by using reduce program, which is part of the AMBER18 suite. The protonation states of each histidine residue were assigned manually based on hydrogen-bonding patterns. Hydrogen positions were relaxed by the simulated annealing protocol using short molecular dynamics (MD). The protocol includes the optimisation of hydrogens, annealing and optimisation in the solvent igb7[43] model. The MD protocol was the following: the initial temperatures were assigned following Maxwell Boltzmann distribution to the target temperature of 1000 K. They were kept at 1500 K for 1 ps and then cooled down to 0 K over 2 ps. Optimisation was carried out employing the Broyden-Fletcher-Goldfarb-Shanno algorithm using a limited amount of computer memory with the igb7 solvent model.

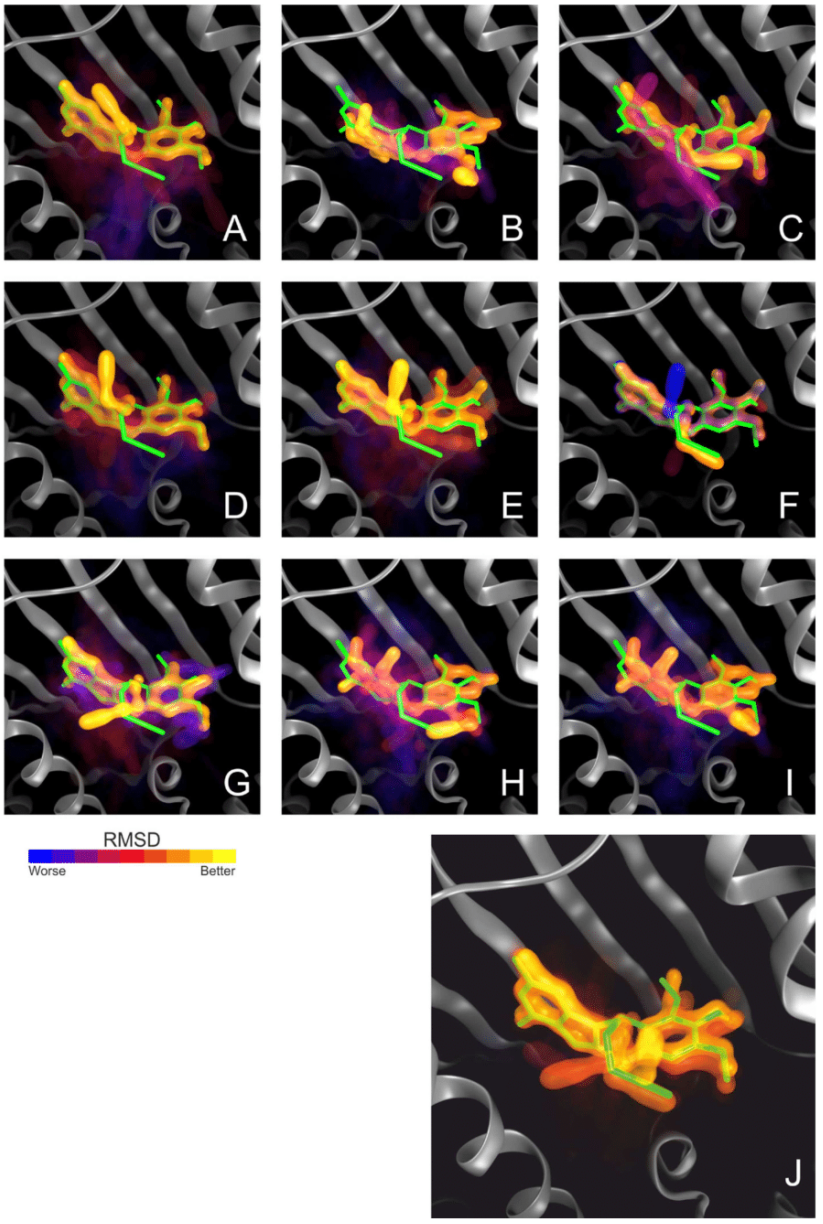


Figure 1. Individual representation of RMSD vs Energy for docking and scoring results of 9 different DFs and the scoring result of SQM-based SF: A) AD4, B) ASP, C) CSCR, D) GSCR, E) GlideSP, F) GlideXP, G) PLP, H) SMINA, I) VINA and J) represent the results of SQM. For these results, 1UYF crystal ligand has been used for the comparison in Post Dock.[35] It displays an interactive pseudo-3D snapshot of multiple docked ligand poses such that both the docking poses and docking scores are encoded visually for rapid assessment. The docking energies are represented by a transparency scale whereas the docking poses are visually encoded by a colour scale. Reference ligand localization in the binding site is shown in green colour with the stick model. The poses from the docking functions are shown in the surface model. The colours from the tinted yellow to the faint blue represent the RMSD values. Yellow colour corresponds to the lowest RMSD and a blue colour corresponds to the highest RMSD. Regarding the opaqueness, the opaquest surface represents the lowest energy pose with a better score and the score is getting worse when the transparency increase.

Dockings

We have examined the poses generated by nine docking functions: Glide(SP,XP),[44] AutoDock4,[45] Vina,[46] Smina,[47] and GOLD software,[48–50] using ASP,[51] GoldScore',[48][49] ChemPLP[52] and ChemScore.[51] All hydrogens of the compounds and the receptor were explicitly preserved during all docking calculations to make it possible to see every possible interaction for different protonation states of the same molecule. We have changed the upper limit of the pose production to 100 for all DFs while keeping other settings as default. The centre coordinates of the grid were assigned as the geometrical centre of crystal inhibitor and used for all docking functions.

The grid centre was adjusted in MGLtools. In our grid box all the possible interactions have been checked and presented. Based on the x,y,z centres 20 $Å^3$ grid box covered all the poses. 20 $Å^3$ grid

box was prepared for AutoDock4, AutoDock Vina as well as SMINA. Grid earch space size was specified in "grid points" (0.375 Angstrom), as in AutoDock 4. Docking receptor grid in Glide was generated using the Glide Receptor Grid Generation module. A cubic box of 20 $Å^3$ was placed at the grid box centre. For Gold dockings, Grid centre coordinates were used for binding site origin. The radius value was defined as 12.4 Å in order to produce the same volume as the grid boxes for previous DFs.

Fragmentation

The fragmentation of the protein step was applied to reduce the computational cost for demanding PM6/COSMO calculations for the complex. For this reason, all docked pose coordinates were gathered to generate a reference volume for the fragmentation. The fragmented protein part (receptor) was defined as a selection of protein residues within a 4 Å distance from the reference volume, truncated, and capped by using Cuby4. Hydrogen sampling and optimisation processes were applied as explained in the protein preparation section. MM_N and PM6/COSMO energies of the receptor were noted for scoring calculations.

Scoring Preparation

We used ffPM3[53] for protein, tip3p[54] for water molecules, gaff2 for compounds and the igb7 model for the solvation of AMBER calculations. We assigned partial atomic charges by means of the AM1-BCC[55] charge model implemented in antechamber.[56] Individual input complex structures were generated from docked poses and the receptor. The MM preparation of the complexes were initiated as a 2 ps MD step and the optimisation of the ligand hydrogen atoms and the surrounding H atoms of the receptor within 4 Å with respect to ligand heavy atoms. The hydrogen sampling of the complex step was followed by another optimisation of all ligand atoms along with 4 Å surrounding hydrogen atoms of the receptor.

All compound conformations were also optimised by MM_N and MM_A protocols and single-point PM6/COSMO energy calculations were applied for further deformation penalty inclusion.

Acknowledgements

We are grateful to Dr. Federico Urban for helpful discussion. This work was part of the Research Project RVO: 61388963 of the Institute of Organic Chemistry and Biochemistry, Czech Academy of Sciences. We acknowledge the support from the European Regional Development Fund; OP RDE; Project: 'Chemical Biology for Drugging Undruggable Targets (ChemBioDrug)' (No. CZ.02.1.01/0.0/0.0/16_019/0000729). This work was also supported by the Ministry of Education, Youth and Sports from the Large Infrastructures for Research, Experimental Development and Innovations project 'IT4Innovations National Supercomputing Center – LM2015070'.

Conflict of Interest

The authors declare no conflict of interest.

Keywords: docking · enrichment · non-covalent interactions · semiempirical quantum mechanics-based scoring function · virtual screening

[1] Z. Wang, H. Sun, X. Yao, D. Li, L. Xu, Y. Li, S. Tian, T. Hou, *Phys. Chem. Chem. Phys.* **2016**, *18*, 12964–12975.
[2] H. Gohlke, M. Hendlich, G. Klebe, *J. Mol. Biol.* **2000**, *295*, 337–356.
[3] K. Raha, K. M. Merz, *J. Am. Chem. Soc.* **2004**, *4*, 1020–1021.
[4] M. J. S. Dewar, E. G. Zoebisch, E. F. Healy, J. J. P. Stewart, *J. Am. Chem. Soc.* **1985**, *13*, 3902–3909.
[5] J. J. P. Stewart, *J. Comput. Chem.* **1991**, *3*, 320–341.
[6] J. J. P. Stewart, *J. Mol. Model.* **2007**, *12*, 1173–1213.
[7] J. J. P. Stewart, *J. Mol. Model.* **2013**, *1*, 1–32.
[8] M. Kubillus, T. Kubař, M. Gaus, J. Řezáč, M. Elstner, *J. Chem. Theory Comput.* **2015**, *11*, 332–342.
[9] J. Řezáč, P. Hobza, *Chem. Rev.* **2016**, *116*, 5038–5071.
[10] J. Řezáč, K. E. Riley, P. Hobza, *J. Chem. Theory Comput.* **2012**, *8*, 4285–4292.
[11] J. J. P. Stewart, MOPAC 2016, Stewart Computational Chemistry, Colorado Springs, CO: **2016**.
[12] A. Klamt, G. Schüürmann, *J. Chem. Soc. Perkin Trans. 2* **1993**, *0*, 799–805.
[13] J. Fanfrlík, A. K. Bronowska, J. Řezáč, O. Přenosil, J. Konvalinka, P. Ho, *J. Phys. Chem. Btry B* **2010**, *114*, 12666–12678.
[14] M. Lepšík, J. Řezáč, M. Kolář, A. Pecina, P. Hobza, J. Fanfrlík, *ChemPlusChem* **2013**, *78*, 921–931.
[15] M. Lepšík, J. Řezáč, M. Kolář, A. Pecina, P. Hobza, J. Fanfrlík, *ChemPlusChem* **2013**, *78*, 921–931.
[16] A. Pecina, R. Meier, J. Ich Fanfrlí, M. Lepš, Í K. J. Řezá, P. Hobza, C. Baldauf, *Chem. Commun. Chem. Commun* **2016**, *3312*, 3312–3315.
[17] H. Ajani, A. Pecina, S. M. Eyrilmez, J. Fanfrlík, S. Haldar, J. Řezáč, P. Hobza, M. Lepšík, *ACS Omega* **2017**, *2*, 4022–4029.
[18] J. Řezáč, *J. Chem. Theory Comput.* **2017**, *13*, 4804–4817.
[19] J. Řezáč, P. Hobza, *J. Chem. Theory Comput.* **2012**, *8*, 141–151.
[20] J. Řezáč, K. E. Riley, P. Hobz, *J. Chem. Theory Comput.Ion* **2011**, *7*, 2427–2438.
[21] A. Pecina, J. Brynda, L. Vrzal, R. Gnanasekaran, M. Hořejší, S. M. Eyrilmez, J. Řezáč, M. Lepšík, P. Řezáčová, P. Hobza, P. Majer, V. Veverka, J. Fanfrlík, *ChemPhysChem* **2018**, *7*, 873–879.
[22] A. Sawai, S. Chandarlapaty, H. Greulich, M. Gonen, Q. Ye, C. L. Arteaga, W. Sellers, N. Rosen, D. B. Solit, *Cancer Res.* **2008**, *68*, 589–596.
[23] C. E. Stebbins, A. A. Russo, C. Schneider, N. Rosen, F. U. Hartl, N. P. Pavletich, *Cell* **1997**, 89, 239–250.
[24] M. M. Mysinger, M. Carchia, J. J. Irwin, B. K. Shoichet, *J. Med. Chem.* **2012**, *55*, 6582–6594.
[25] D. A. Case, T. E. Cheatham, T. Darden, H. Gohlke, R. Luo, K. M. Merz, A. Onufriev, C. Simmerling, B. Wang, R. J. Woods, *J. Comput. Chem.* **2005**, *26*, 1668–1688.
[26] O. Gutten, D. Bím, J. Řezáč, L. Rulíšek, *J. Chem. Inf. Model.* **2018**, *58*, 48–60.
[27] J. Řezáč, *J. Comput. Chem.* **2016**, *37*, 1230–1237.
[28] A. Nicholls, *J. Comput.-Aided Mol. Des.* **2008**, *22*, 239–255.
[29] H. Fan, J. J. Irwin, B. M. Webb, G. Klebe, B. K. Shoichet, A. Sali, *J. Chem. Inf. Model.* **2009**, *11*, 2512–2527.
[30] H. Chen, P. D. Lyne, F. Giordanetto, T. Lovell, J. Li, *J. Chem. Inf. Model.* **2005**, *1*,401-415.
[31] N. Huang, B. K. Shoichet, J. J. Irwin, *J. Med. Chem.* **2006**, *23*, 6789–6801.
[32] R. D. Clark, D. J. Webster-Clark, *J. Comput.-Aided Mol. Des.* **2008**, *3–4*, 141–146.
[33] S. M. Vogel, M. R. Bauer, F. M. Boeckler, *J. Chem. Inf. Model.* **2011**, *10*, 2650–2665.
[34] O. Korb, T. Ten Brink, F. R. D. V. P. Raj, M. Keil, T. E. Exner, *J. Comput.-Aided Mol. Des.* **2012**, *2*,185-197.
[35] E. A. Wiley, G. Deslongchamps, *Computing and Visualization in Science* **2009**, *1*, 1–7.
[36] Chemical Computing Group ULC, *Molecular Operating Environment (MOE)*, Montreal, Quebec, Canada **2018**.
[37] E. Harder, W. Damm, J. Maple, C. Wu, M. Reboul, J. Y. Xiang, L. Wang, D. Lupyan, M. K. Dahlgren, J. L. Knight, *J. Chem. Theory Comput.* **2016**, *12*, 281–296.
[38] J. C. Shelley, A. Cholleti, L. L. Frye, J. R. Greenwood, M. R. Timlin, M. Uchimaya, *J. Comput.-Aided Mol. Des.* **2007**, *21*, 681–691.
[39] Schrödinger, *Schrödinger Release 2018–2* **2018**.

[40] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, P. E. Bourne, *Nucleic Acids Res.* **2000**, *28*, 235–42.
[41] L. Wright, X. Barril, B. Dymock, L. Sheridan, A. Surgenor, M. Beswick, M. Drysdale, A. Collier, A. Massey, N. Davies, *Chem. Biol.* **2004**, *11*, 775.
[42] K. Haider, D. J. Huggins, *J. Chem. Inf. Model.* **2013**, *53*, 2571–2586.
[43] J. Mongan, C. Simmerling, J. A. McCammon, D. A. Case, A. Onufriev, *J. Chem. Theory Comput.* **2006**, *1*, 156–169.
[44] R. A. Friesner, J. L. Banks, R. B. Murphy, T. A. Halgren, J. J. Klicic, D. T. Mainz, M. P. Repasky, E. H. Knoll, M. Shelley, J. K. Perry, P. S. Shenkin, *J. Med. Chem.* **2004**, *7*, 1739–1749
[45] G. M. Morris, D. S. Goodsell, M. E. Pique, R. Huey, S. Forli, W. E. Hart, S. Halliday, R. Belew, A. J. Olson, *J. Comput. Chem.* **2009**, *16*, 2785–2791.
[46] O. Trott, A. J. Olson, *J. Comput. Chem.* **2009**, *31*, 455–461.
[47] D. R. Koes, M. P. Baumgartner, C. J. Camacho, *J. Chem. Inf. Model.* **2013**, *53*, 1893–1904.
[48] J. W. M. Nissink, C. Murray, M. Hartshorn, M. L. Verdonk, J. C. Cole, R. Taylor, *Proteins Struct. Funct. Genet.* **2002**, *49*, 457–471.
[49] G. Jones, P. Willett, R. C. Glen, A. R. Leach, R. Taylor, *J. Mol. Biol.* **1997**, *267*, 727–748.
[50] G. Jones, P. Willett, R. C. Glen, *J. Mol. Biol.* **1995**, *1*, 43–53.
[51] W. T. M. Mooij, M. L. Verdonk, *Proteins Struct. Funct. Bioinf.* **2005**, *61*, 272–287.
[52] O. Korb, T. Stützle, T. E. Exner, *J. Chem. Inf. Model.* **2009**, *1*, 84–96.
[53] A. M. Wollacott, K. M. Merz, *J. Chem. Theory Comput.* **2006**, *4*,1070-1077.
[54] W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, M. L. Klein, *J. Chem. Phys.* **1983**, *79*, 926–935.
[55] A. Jakalian, D. B. Jack, C. I. Bayly, *J. Comput. Chem.* **2002**, *23*, 1623–1641.
[56] J. Wang, W. Wang, P. A. Kollman, D. A. Case, *J. Mol. Graphics Modell.* **2006**, *25*, 247–260.

Manuscript received: June 26, 2019
Revised manuscript received: August 21, 2019
Accepted manuscript online: August 28, 2019
Version of record online: September 11, 2019