「1-Click Drug Discovery from Genome to Therapeutics on the Fugaku Computer」

RIKEN Center for Computational Science
Computational Molecular Science Research Team
Bun CHAN

## 1　研究の背景と目的

The emergence of new variants of the COVID virus has demonstrated the importance of rapid development of new therapeutics.　Despite the worldwide scientific focus on this virus for the past few years, the development of effective treatments remains untimely relative to the rate of viral mutations.　In this regard, a major challenge in this endeavor, and in drug discovery more generally, is to accurately identify drug candidates for a biological target, which is typically an enzyme.

Experimental screening of drug candidates, while being relatively efficient after years of development, remains laborious, resource intensive, and, in some cases, for example, where the target is highly active, difficult and unreliable.　As a complement or an alternative, virtual screening with computational chemistry holds the potential to greatly improve the efficiency with the aid of modern supercomputers.　However, among the many existing computational chemistry methods, the highly accurate ones are excessively demanding on computer resources such that they cannot be applied to enzymes even with the most advanced supercomputers.　On the other hand, those applicable ones lack the reliability to improve on experiment.

There have been tremendous efforts in improving the accuracy of low-cost computational chemistry methods.　Some of the latest methods have reached an adequate level of accuracy for many small molecular species, and it is perceivable that they may be applicable to large enzyme systems for reliable virtual screening.　This project aims to examine the performance of these methods for enzymes and related systems.　The strengths of a collection of methods will be exploited and optimally combined to formulate an accurate yet efficient protocol to enable reliable high throughput computation.　The success of such a development would ultimately enable a “1-click” process to facilitate timely development of drugs for emerging diseases.

## 2　研究方法・研究内容

As the technical background for our execution plan, we first describe relevant aspects of computational chemistry.　The most widely used class of methods is “density functional theory” (DFT).　Another component in essentially all calculations is a “basis set”, and in general the larger the basis set the more accurate are the computed results.　Typically, DFT methods are used with medium-sized basis sets called TZ basis sets, with which the results are close to the maximum achievable accuracy of a given DFT method.　Among the many types of DFT methods, “hybrid DFT” is generally considered to be the best balance between accuracy and computational efficiency.

Some of the best hybrid-DFT/TZ protocols can achieve the "chemical accuracy" of 1 kcal $mol^{-1}$, which is a level of accuracy that is considered acceptable in essentially all chemical applications. A hybrid-DFT/TZ protocol can be viably applied to a molecular system with a maximum of a few hundred atoms using a typical supercomputer. Nonetheless, such computations are time consuming, often requiring weeks of continuous operation. In comparison, a typical enzyme contains thousands of atoms, and their computation with hybrid-DFT/TZ would be intractable.

An established approach to apply a reasonable computation chemistry method to enzymes is the "QM/MM" approximation, which uses, for example, DFT for the most critical part of the enzyme (the active site) with up to a few hundred atoms, and a much lower-cost method for the rest of the enzyme. While QM/MM has been useful, there are pitfalls that have limited its accuracy and, in some cases, its computational viability. Two key issues are (1) the accuracy of the low-cost method, which is usually a "molecular mechanics" (MM) method, is often inadequate even for the treatment of the non-critical part of the enzyme, and (2) the application of hybrid-DFT/TZ to the active site remains costly in many cases.

This project is designed to tackle these two challenges. Regarding the low-cost method, we realize that the "tight-binding" (TB) class of methods, in particular the state-of-the-art "XTB" methods, approach the accuracy of DFT, while being just marginally more costly than MM within the context of QM/MM. Thus, XTB holds the potential to resolve the first issue. Regarding the cost of treating the active site, we note that recent developments in non-hybrid-DFT methods, which are more economical than hybrid-DFT but are traditionally less accurate, have brought their accuracy to a level similar to that of hybrid-DFT. At the same time, nil-cost correction schemes to low-cost "DZ" basis sets have improve their accuracy to a level similar to that of TZ. As a result, using non-hybrid-DFT/DZ may significantly improve the efficient in the computation of the active site and by extension the entire QM/MM calculation.

In this project, we will survey a series of non-hybrid-DFT methods in combination with different DZ-type basis sets to determine a suitable protocol for the calculation of enzyme active sites. The optimal protocol will then be used in combination of TB and related methods in a QM/MM setting. Technical options used in the QM/MM combination will be thoroughly explored to determine an optimal QM/MM methodology for efficient calculation of "binding energies" between drug molecules and enzymes, which is often taken as an indicator for drug efficacy.

We will primarily use the chemical accuracy and computational efficiency as the key metrics to construct the QM/MM methodology. Results will be compared to experimental outcomes of drug screening to validate the protocol. Experimental benchmark data will be sourced from public databases, and we will primarily target enzymes that are of widespread relevance, e.g., to disease such as COVID and cancer, such that abundant high-quality experimental data are available.

3　研究成果

In our previous preliminary study, we have examined a wide range of DFT methods for the computation of binding energies in systems that are relevant to biological settings. We have identified a non-hybrid-DFT method called "B97M-V" to be of sufficiently accurate and reliable for this purpose when it is used with a TZ basis set, with an accuracy that is below 1 kcal $mol^{-1}$. In this project, we have assessed the prospect of using DZ basis sets together with supplements called "geometric counter-poise" (gCP) corrections as a substitute for TZ. We have additionally examined an alternative, widely used, non-hybrid-DFT method (PBE-D3BJ), to explore the possibility of improving on B97M-V. The average errors for our data set of biological-related systems are shown in the table below.

Table 1. Error (kcal $mol^{-1}$) for biologically relevant binding energies

| DFT | basis | gCP | error |
|---|---|---|---|
| B97M-V | TZ | | 0.9 |
| B97M-V | DZ | | 8.4 |
| B97M-V | DZ | pbeh3c | 3.3 |
| B97M-V | DZ | dft/svp | 3.7 |
| B97M-V | DZ | gga/svp | 5.1 |
| PBE-D3BJ | DZ | pbeh3c | 5.5 |

The results show that the use of a DZ basis set leads to substantially larger error (8.4 kcal $mol^{-1}$) than the TZ binding energies (0.9 kcal $mol^{-1}$). However, the addition of gCP to DZ yields notably better accuracies. There are several commonly used options for the calculation of gCP corrections, and we find that the "pbeh3c" option gives the lowest error of 3.3 kcal $mol^{-1}$. Substituting the DFT method of B97M-V with PBE-D3BJ leads to a larger error of 5.5 kcal $mol^{-1}$. While an error of 3.3 kcal $mol^{-1}$ for B97M-V/DZ + gCP(pbeh3c) exceeds the threshold of chemical accuracy, it is unclear whether it may already be adequate for the calculation of drug–enzyme binding energies. Thus, we proceed to apply this protocol in our QM/MM studies of enzyme systems.

In our investigation into QM/MM protocols using B97M-V and XTB, we have explored a wide range of technical options that determine the accuracy and computational efficiency of the results. A key option is the proportion of the enzyme that is calculated using the more accurate but expensive B97M-V method. We have determined that an active site with a radius of 0.4 nm would be sufficient. Other options include the means for combining B97M-V and XTB, and various algorithms in the actual computational chemical calculations. Without going into the details, we find that key technical options for optimal computations are: (1) using "amide caps" to combine B97M-V and XTB, (2) using "GFN-FF" and the "FIRE" algorithm in the intermediate preparation step in QM/MM, and (3) using the "KDIIS" and "SlowConv" algorithms for the B97M-V component of the QM/MM calculations. With these optimal options, a QM/MM calculation for a real-life drug–enzyme system can be completed with a mid-range desktop computer in just several hours.

We have applied the optimal QM/MM protocol to several sets of drug–enzyme systems to validate its utility for drug discovery. These systems include the enzymes, (1) the COVID protease, (2) cancer cell Bcl-xL protein, (3) the ricin-A toxin, and (4) the HIV protease. They cover a wide range of biological functions and structural diversity for rigorous testing. In all cases, we find that the calculated binding energies correlate well with the known drug activities. These correlations are shown in the figure below. Our results have been submitted for publication. As part of this project, we have also carried out additional validations of DFT methods, and the findings have been published.

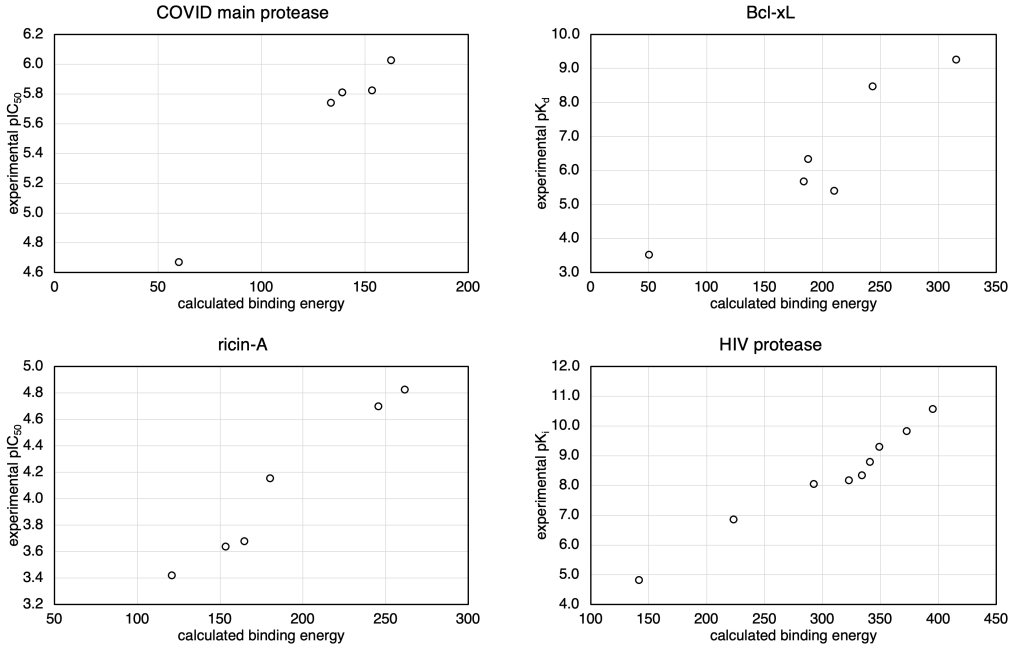


Figure 1. Calculated binding energies for durg–enzyme complexes for four enzyme targets.

## 4　生活や産業への貢献および波及効果

The new QM/MM computational chemistry protocol of this project will enable rapid and reliable drug discovery for a wide range of pathogen enzyme targets. More generally, a major challenge in drug development is testing for side effects. Much of this is due to non-specific inhibition of proteins. The efficient QM/MM method would further enable large-scale screening of molecule–protein interactions and thus the discovery of potential side effects. In turn, this would facilitate targeted experiments for more precise discovery of therapeutics. In future developments, the protocol can be integrated with freely available computational tools, notably the recently developed AlphaFold, to construct a fully automated high-throughput, and, importantly, reliable, drug discovery progress. The COVID situation of the past few years, which has caused tremendous societal and economic disruptions, has shown that the only viable solution is likely the rapid development of therapeutics. This project has contributed to paving a way towards such an improved response to future public health challenges.

Publications: (1) Electron. Struct. 2022, 4, 044001, (2) J. Phys. Chem. A 2022, 126, 4981, (3) J. Comput. Chem. 2022, 43, 1394, and (4) J. Phys. Chem. A 2022, 126, 2397.