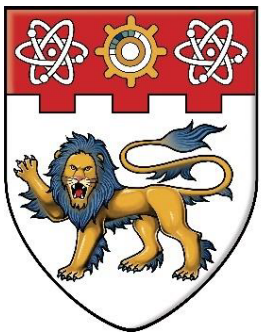


NANYANG TECHNOLOGICAL UNIVERSITY

PROGESTERONE RECEPTOR LIGAND BINDING DOMAIN: FROM CONFORMATIONAL DYNAMICS TO DRUG DISCOVERY

ZHENG LIANGZHEN

SCHOOL OF BIOLOGICAL SCIENCES

2019

PROGESTERONE RECEPTOR LIGAND BINDING DOMAIN: FROM CONFORMATIONAL DYNAMICS TO DRUG DISCOVERY

ZHENG LIANGZHEN

SCHOOL OF BIOLOGICAL SCIENCES

A thesis submitted to the Nanyang Technological University in partial fulfillment of the requirement for the degree of Doctor of Philosophy

2019

Statement of Originality

I hereby certify that the work embodied in this thesis is the result of original research done by me except where otherwise stated in this thesis. The thesis work has not been submitted for a degree or professional qualification to any other university or institution. I declare that this thesis is written by myself and is free of plagiarism and of sufficient grammatical clarity to be examined. I confirm that the investigations were conducted in accord with the ethics policies and integrity standards of Nanyang Technological University and that the research data are presented honestly and without prejudice.

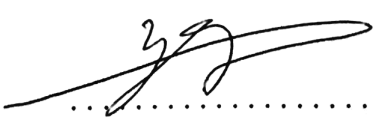
12/03/2019
Date

Liangzhen
Zheng Liangzhen

Supervisor Declaration Statement

I have reviewed the content and presentation style of this thesis and declare it of sufficient grammatical clarity to be examined. To the best of my knowledge, the thesis is free of plagiarism and the research and writing are those of the candidate's except as acknowledged in the Author Attribution Statement. I confirm that the investigations were conducted in accord with the ethics policies and integrity standards of Nanyang Technological University and that the research data are presented honestly and without prejudice.

12 Mar. 2019
.....
Date


.....
Mu Yuguang

Authorship Attribution Statement

This thesis contains material from 2 paper(s) published in the following peer-reviewed journal(s) where I was the first author.

Chapter 1 and Chapter 4 are published as Zheng L, Lin V C, Mu Y. Exploring flexibility of progesterone receptor ligand binding domain using molecular dynamics[J]. *PloS one*, 2016, 11(11): e0165824.

The contributions of the co-authors are as follows:

- A/Prof Mu Yuguang provided the initial project direction and edited the manuscript drafts.
- I prepared the manuscript drafts. The manuscript was revised by Dr Valerie Lin.
- I co-designed the study with A/Prof Mu Yuguang and performed all the laboratory work at the School of Biological Sciences. I also analyzed the data.

Chapter 2 is published as Zheng L., Alhossary A., Kwoh C.K., Mu Y. Molecular Dynamics and Simulation. 2016.

The contributions of the co-authors are as follows:

- I and Dr. Amr Alhossary prepared manuscript drafts.
- The manuscript was revised by A/Prof Mu Yuguang and A/Prof Chee-Keong Kwoh.

10/03/2019 ..... Liangzhen .....

Date Zheng Liangzhen

Acknowledgments

My supervisor, Dr. Mu Yuguang, has never hesitated to guide me in both academic and life. He is not only a mentor but most importantly a friend. Without his kind support, I wouldn't be able to complete the Ph.D. research. I'd like to express my truthful thanks to him.

I can't imagine how my life would be if my wife, HH, wasn't there for me. Words are not enough to express my love and gratitude. And my parents, sister, brother in law, and two young nephews thank you all for your patience and support, love you.

During my four years' Ph.D. study, many professors are very helpful and kind. I'd like to thank Dr. Lu Lanyuan, Dr. Valerie Lin, Dr. Miao Yansong, Dr. Chee-Keong Kwoh, Dr. Surajit Bhattacharjya, and Dr. Shi Jiahai (The City University, Hong Kong), for their guidance, suggestions, and collaborations.

My colleagues, both current and former, are also very nice people to work with. Here, I'd like to thank the following ladies and gentlemen: Dr. Xu Weixin, Dr. Wang Yaofeng, Dr. Zhang Tong, Dr. Luo Di, Dr. Amr A. Alhossary, Dr. Yaw Awuni, Dr. Ning Lulu, Dr. Guo Jingjing, Dr. Fan Jingrong, Dr. Meng Zhenyu, Dr. Chua, Khi Pin, Mr. Liu Yang, Mr. Justin Ng, Mr. Saxena Shikhar, Ms. Nafisa Ali Alhossary.

Finally, I'd like to thank MOE, Singapore, for supporting my Ph.D. studies financially. For NYA HPC (NTU), and NSCC (Singapore), great appreciations are owed to them, for the shared computational resources.

Wish all the best to the people I mentioned! It is very lucky to meet you, work with you, and have you. Without you, I wouldn't be me and wouldn't have gone this far. Thank you!

Table of Contents

Table of Figures

Table of Tables

Abbreviations

cMD	Conventional molecular dynamics
COM	Center of mass
CTD	C-terminal domain
CV	Collective variable
DT	Decision tree
EDA	Essential dynamics analysis
FES	Free energy surface
HREMD	Hamiltonian replica exchange molecular dynamics
HTVS	High throughput virtual screening
LBD	Ligand binding domain
LBVS	Ligand based virtual screening
MD	Molecular dynamics
MLP	Multiple-layer perceptron
MSE	Mean square error
NMA	Normal mode analysis
NR	Nuclear receptor
NTD	N-terminal domain
PCA	Principle component analysis
PMF	Potential of mean force
PR	Progesterone receptor
REMD	Replica exchange molecular dynamics
RMSD	Root mean square deviation
RMSF	Root mean square fluctuation
RF	Random forest
SBVS	Structure based virtual screening
SVM	Support vector machine

Abstract

Hormones are vital molecules for human differentiation, development, and health. Among them, the female reproduction, development, and cancer-related hormone progesterone has been cast interests for its involvement in nearly all aspects of female health. Recently researches have demonstrated that progesterone, in alliance with estrogen, would suppress breast cancer cell proliferation in cell lines and mouse models. Its receptor, the progesterone receptor (PR), which is a member of the nuclear receptor superfamily, has gradually been recognized as a potential drug target. The dysfunction of this PR would contribute to different types of cancers, including breast cancer. Although there have been several drug molecules designed or discovered against this PR, the side-effects of these drugs are not to be ignored. The necessity of synthesizing new types of drugs against PR, especially its ligand binding domain (LBD), is never decreased. Current studies about PR mainly focus on the biological functions, signaling pathway, its involvement in cancer models. The structural and dynamics of PR, which lay the foundation of the biological understandings, however, are limited, partially due to the difficulty of performing X-ray crystallography or nuclear magnetic resonance spectroscopy (NMR) of the PR, or its domains.

We hope to incorporate the structure and dynamics information of PR, or more specifically, the LBD of PR, to identify high potential PR lead molecules for pharmaceutical purposes. Therefore, starting from the available experimental solved structures, we applied biophysics methods to 1) explore the dynamics of PR LBD, 2) and understand the interactions between PR LBD and its agonists and antagonists, 3) and screen large compound libraries to identify potential lead-like molecules. Molecular dynamics (MD) simulations and advanced sampling methods generate good models to estimate the energy landscape of the *apo*-form PR LBD. The results suggest that the agonistic PR LBD conformation is a meta-stable state. Several other meta-stable states have also been identified and they could be adopted for further virtual screening studies. And the ligand binding (either an agonist or an antagonist) would induce the LBD conformational adaptations more towards the “closed” agonistic state, though different ligands may induce different PR LBD dynamics. We also found that the co-repressor peptide is a necessary component for maintaining the antagonistic conformation. In meanwhile, the ligand-induced conformational adaptations result from both hydrophobic and electrostatic forces. Subsequently, large-scale virtual screening (VS), as well as machine learning-based rescoring calculations, were performed. Towards PR LBD VS, we constructed highly accurate and low false positive rate models and identified 21 potential lead-like molecules, which would be further tested in future.

The computational methods or models we used here to fill in the void in structural and dynamics knowledge of PR. The identified molecules could be potential lead molecules for further analysis. The drug discovery strategy we applied would also be useful and could be applied to other targets.

Chapter 1. PR LBD, from dynamics to virtual screening

PR is a member of the nuclear receptor (NR) superfamily, which regulates a complex gene network by their transcription effects [1]. The receptors of NR superfamily could be further divided into three classes. The first class includes many steroid hormone receptors, such as PR, estrogen receptor (ER) and androgen receptor (AR); the second class may be called the thyroid/retinoid family, including receptors such as vitamin D receptor (VDR) and retinoic acid receptor (RAR); while the third class is the orphan receptor family, which is relatively less studied [2]. In the following section, we would provide a brief introduction to the steroid receptor PR.

Nucleic PR is one of the cellular natural receptors for steroid hormone progesterone and it plays important roles in female reproductive development, differentiation, and maintenance. Besides this PR (within cellular or nucleus), there is another type of receptors for progesterone locating in the cell membrane, mPR.

The full-length PR is composed by an unstructured N terminal domain (NTD), and two other structural conserved domains: a DNA binding domain (DBD) and an LBD (Figure 1). Human PR is the expression product of the *PRG* gene in chromosome 11. *PRG* could use different promotors to synthesize two receptor isoforms, PR-B, and PR-A, whose amino-acid sequences are almost identical, except a 165-amino-acids NTD missing in PR-A [2-4]. For their sharing sequences, there exist two activation function (AF1 and AF2) regions. The NTD contains a third activation function (AF3) region. This AF3 region enables the PR-B a relatively stronger transcription factor both in cellular experiments [5] and mouse models [6]. Besides, PR-B differs from PR-A in more physical effects. For example, the RU486 (Mifepristone) acts as a pure antagonist for PR-A, however, it is a selective modulator for PR-B [2, 7]. Moreover, the two isoforms may activate different subsets of downstream genes [8]. Recent studies have shown that PR-A and PR-B also contribute to the progress of breast cancer differently [9], and the ratio of these two isoforms could regulate the population of mammary gland stem cell in mouse [10]. However, the detailed mechanisms of these differences remain unclear.

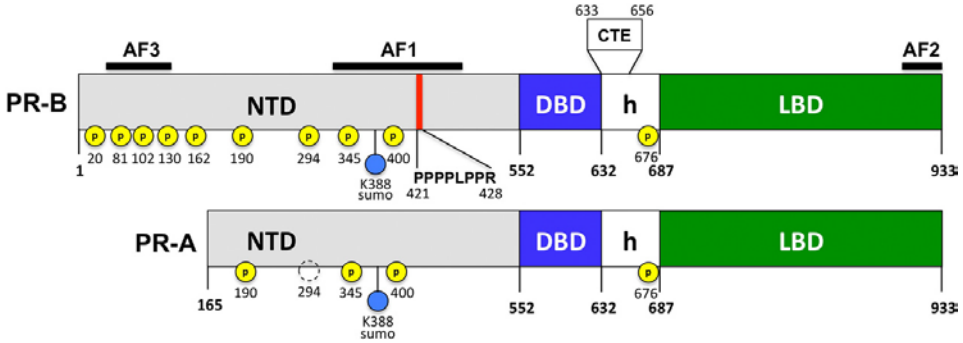


Figure 1. The domains of two isoforms of PR. PR-B is the full-length PR with 3 AFs.

PR-A lacks the first 164 residues, which contain the AF3. PR-B and PR-A share an identical sequence containing DBD and LBD. Figure adopted from Hill et al [3].

1.1 Agonists, antagonists, and their biological functions

Upon its agonist progesterone binding (*holo*-form), conformational changes of helix 12 are trigged. And two PRs form a dimer and is translocated into the nucleus and further binds to specific DNA sequences through its DBDs and recruits coactivators' binding near AF2 region, thus mediates the subsequent transcription activation [2, 11]. It is suggested that the dimerization (either homo-dimer or hetero-dimer) are necessary for functional DNA binding unit. However, monomeric form of PR could also localize near the PRE and induce endogenous gene expressions [12].

What's more, induced by progesterone, PR could initiate protein phosphorylation signaling cascades [13] in the cytosol. Many experiments have proved that progestins, the synthetic agonists of PR, play a key role in promoting and maintaining mammary gland neoplastic phenotype due to its agonistic effects. Synthetic progestins, and agonists of PR LBD have been widely used in gynecological diseases and menopausal therapy [14, 15]. Meanwhile, progesterone (P4) and synthetic progestins would also orchestra highly complicated interaction networks with other hormones, as well as nuclear receptors. P4 binding would induce the activation state of PR, with AF2 (helix 12) patched on the ligand binding pocket and exposed the co-activators binding surface.

Early findings indicate that P4, or other progestins, independent of estrogen, is a risk factor for human breast cancer [16]. However, recent studies suggest that the synthetic progestin's androgenic properties, rather than native PR agonistic effect, may increase the risk of breast cancer development in hormone replacement therapy [16-18]. Other evidence, however, indicate that in some situations, P4 or synthetic progestins could inhibit breast cancer proliferation [19-21]. A recent work provides the evidence of the interplay between ERα and PR. It was proved that the co-existence of estrogen and P4 [22] enhancing anti-proliferative genes' expressions, partially due to the chromatin re-localization induced by estrogen-bound ERα and P4-bound PR heterodimer, therefore suppressing the tumor growth in mouse models, as well as human breast cancer cell lines [23]. The collaboration of ERα and PR illustrates the importance of the PR agonist for $ER\alpha^+$ breast cancer therapy [24], therefore it highlights the value of designing novel pure agonistic molecules for PR.

In contrast to agonists, PR LBD antagonists may inhibit the PR-related activation of downstream gene expressions. Antagonists would also bind to PR LBD in the classic binding pocket near helix 12. However, antagonist-bound LBD then forms the antagonistic conformation with helix 12 far away from the binding pocket and parallel aligned with helix 11 [25], therefore recruits co-repressor binding around the cleft formed by helix 3 and helix 4, then inhibits the activation of specific genes' expressions.

Animal model studies suggest that antagonists could be employed to inhibit progestin-dependent mammary carcinogenesis [26-28]. The first clinic trial antagonist for PR LBD is RU486, which also invokes antagonistic effects with AR and glucocorticoid receptor (GR) [29]. Initially, Mifepristone was used for pregnancy termination, however, it is effective for fibroid treatment as well [30]. After that, large varieties of PR agonists, antagonists and selective PR modulators (SPRMs), have been found or designed.

Some PR LBD antagonists could be employed for gynecological disorder treatment [31] and sometimes for carcinogenesis inhibition [32], long-term use would cause severe side-effects [20, 31, 32]. The safety issues always arise since these antagonists, however, are not pure antagonists for PR LBD, displaying mix-profile effects, and are possible selective modulators for other nuclear receptors. The non-selectivity or low selectivity against PR could trigger complicate biological effects inside a cell, thus hinder the successful usage of these molecules in clinical trials.

The discovery and designing of highly selective molecules (agonists or antagonists) for PR LBD only, would be extremely valuable to female health. Therefore, the discovery and design of novel PR LBD binding molecules are urgent and important.

1.2 PR LBD structural models

Full-length PR LBD contains an NTD, a DBD, and an LBD. For NTD, no defined secondary structures could be modeled or solved. It is speculated that this domain is an intrinsically disordered region (IDR) which response to co-factor binding and ligand binding in LBD [1, 2].

The globular DBD is composed of two α-helices and two zinc-binding motifs according to the crystal model published in 2006 [3, 33]. The interactions between DBD and progesterone response element (PRE) are quite conserved among the nuclear receptors [3]. In that model, two DBDs form a head-to-head homo-dimer, which make contacts with a repeat DNA region, the PRE [3, 34]. Helix 1 in DBD binds to the major groove of the PRE, and the second which forms a ~90° crossing angle with helix 1, also interacts with the PRE. The helix 2 extension residues could make contacts with the same region of the other DBD monomer and mediates the DBD-DNA interactions.

Near all the NR LBDs share a common folding topology, a three-layer globular motif formed by 11 or 12 α-helices and 1 or 2 β-turn [2]. There are diverse states of LBD structures, the *apo*-form extended conformations, *holo*-form agonistic conformations, and antagonistic conformations. The *apo*-form extended conformation was firstly described in *apo*-form retinoid X receptor alpha (RXRα) [35] and later in ERα LBD, where helix 12 is totally dispatched from core LBD surface [36, 37].

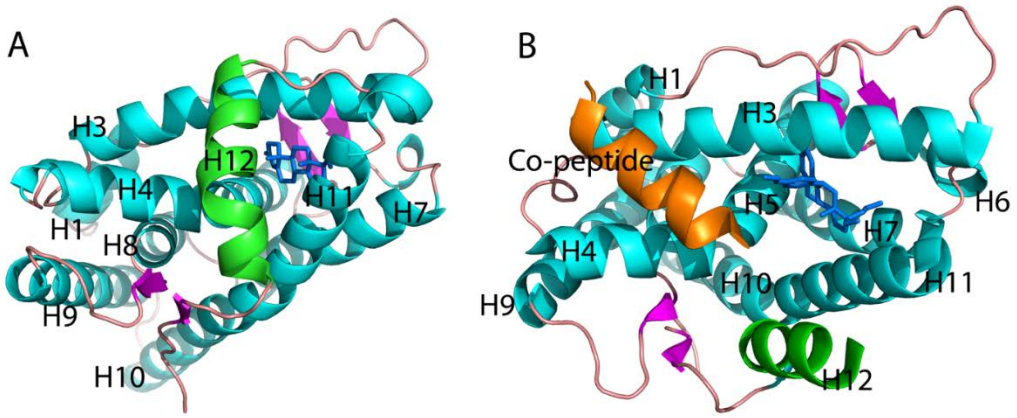


Figure 2. Crystal structure models of *holo*-form PR LBD.

A, in agonistic conformation (PDB ID: 1A28), helix 12 (green, labeled by H12) is closely patched on the core LBD; the ligand is progesterone (marine). B, in antagonistic conformation (PDB ID: 2OVH), helix 12 is displaced to nearly parallel aligned with helix 11, while a co-peptide binds to LBD in the cleft formed by helix 3 and helix 4 and the ligand asoprisnil (marine) is positioned in a similar manner as progesterone in agonist conformation; the loop between helix 12 and helix 11 is missing in crystal structure. Helices are labeled as 'H', such as H1 represents helix 1, and are shown by cyan color (except helix 12); ligands, loops, and β-sheet are marine blue, pink and magnet, respectively. The co-peptide in antagonist conformation is colored as orange.

The abundant NR LBD crystal or NMR structures deposited in RCSB Protein Data Bank suggest that helix 12 exhibit large flexibility while the other 11 or 10 α-helices remain the three-layer sandwich folding. For example, in *apo*-form RXRα LBD and *apo*-form ERα LBD, helix 12 is shifted away from the core LBD in a manner where helix 12 no longer forms hydrophobic interactions with other helices. However, in steroid PRs, upon the agonist binding, helix 12 folds back blocking the solvent access from the binding pocket and forms interactions with helix 3 and helix 4 to create a hydrophobic binding groove [2, 35] facilitating coactivators' binding [38] (Figure 2A). The structure model of antagonist-bound NR LBD, however, is a different story, and the antagonistic conformations also differ for diverse NR LBDs. Taking PR LBD as an example, AF2 helix 12 orientates far apart from helix 3 but almost aligns in parallel with helix 11 (Figure 2B). The conformation is further stabilized by a corepressor peptide containing the LXXLL helical motif. The existence of such a corepressor peptide would block helix 12 forming the hydrophobic cleft thus may prevent the forming of active conformation and the recruitment of the other coactivators [2, 39]. The helix 12 orientations may also affect the pocket size to accommodate various ligands [39]. The structural plasticity, therefore, affects the coactivator binding and thus subsequent gene expression activation, although the exact co-activator binding site in PR is not clearly determined yet.

The first ligand bound PR LBD structure model [40] (PDB ID: 1A28), which adopts a similar helices packing pattern as other NR LBDs, was reported in 1998. In this model, the LBD adopts the agonistic conformation with progesterone co-crystalized with helix 12 tightly patched on the pocket. Several groups reported that the Met909 and Met908 in the N-terminal region of helix 12 play vital roles for

the stabilization of the agonistic conformation. Firstly, the Met909 and Met908 side chains are deeply tangled with the progesterone by hydrophobic interactions [41]. Besides, the Glu723 (in helix 3) side chain anchors the backbone atoms of Met908 and Met909 by a so-called dipole system [40, 42]. These interactions may restrict the flexibility of the N-terminal region of helix 12 closing to helix 3. The crystal structure model of the antagonistic conformation of PR LBD (PDB ID: 2OVH and 2OVM), together with the SPRM ligand asoprisnil and the helical corepressor peptide, was published in 2007 by Madauss *et al* [25]. In this model, helix 12 is displaced from the core LBD surface and the co-peptide occupies the H3-H4 hydrophobic groove. It thus postulated that the accessibility of the H3-H4 hydrophobic groove in PR LBD is nessary for the inhibitory function of PR upon co-repressor binding.

However, not all SPRM ligands and PR antagonists could induce the antagonistic conformations. Interestingly, RU486 and asoprisnil were witnessed co-crystalized with agonistic PR LBD [42, 43]. However, helix 12, as well as H11-H12 loop and partial helix 11, in the antagonist-bound agonistic conformation of PR LBD shows higher b-factors, which indicates the more flexible instinct of helix 12 upon antagonist binding.

So far, no *apo*-form PR LBD model has been proposed, and no biophysics efforts have been paid to reveal the *apo*-form PR LBD conformations and free energy landscape. Molecular simulations, as well as other enhanced sampling methods, are powerful tools and were utilized to explore the *apo*-form LBD conformations, as well as the ligand-induced conformational changes, which would widen our understanding of PR LBD dynamics, therefore opening the gate way of large-scale virtual screening towards this target.

1.3 Objective of the thesis

Computer aided drug discovery is the first step of drug development. The continuous improvement of the techniques and protocols of the computational methods is required for efficient and high accurate drug development. However, current methods in computational drug discovery bring too high false positive rate. Especially for PR, new drugs would be applied for breast cancer therapy, abortion, and female reproductive maintenance. But current PR targeting drugs are not capable to treat breast cancer and other female specific cancers and may cause serve side-effects. Meanwhile, the dynamics of PR have been rarely explored. The structure and dynamics are essential information for understanding the structure-function relationship of PR but could hardly be obtained through experimental techniques. What's more, the dynamics of PR would further be applied for VS targeting the LBD.

To address the issues, therefore, the major objectives of this thesis are: 1) exploring the underline structure-function relationship of PR LBD; 2) developing new computational strategies for PR drug

discovery and screening for possible high potential lead molecules for subsequent biochemistry experiment validations.

To fill in the gaps in PR structure-dynamics relationship, in this thesis, we for the first time conducted thorough investigations of PR LBD using extensive theoretical methods including MD simulations, metadynamics, and umbrella sampling. We also introduced a new workflow of drug discovery towards PR including ensemble-initial-structures-based VS and machine learning based rescoring of the docking results and identified. The methods and protocols used in this thesis could also be useful and applied to other drug targets.

1.4 Structure of the contents

The PR dynamics studies and drug discoveries in this thesis have been organized in the following structure: In Chapter 1, an introduction of PR and the research objectives has been provided. The computational and theoretical methods, algorithms, and principles used in our studies have been explained thoroughly in Chapter 2. And in Chapter 3, the dynamics and structures of *apo*-form PR LBD have been discussed, following by the ligand-induced helix 12 dynamics studies in Chapter 4. In Chapter 5, the VS targeting PR LBD has been explained and discussed. While the rescoring strategy of the VS results has been developed and compared with traditional scoring functions in Chapter 6. Lastly, in Chapter 7, an overall summary of the studies in this thesis and future directions have been provided. In addition, the references are also listed in the Reference section.

In summary, the thesis is consisted of 3 parts: 1) the background information of both the biology of the target (PR) (Chapter 1) and theoretical details of the computational techniques used in the studies (Chapter 2); 2) and the structure and dynamics studies of PR LBD (Chapter 3 and Chapter 4), as well as the VS, rescoring, and prediction of the lead-like molecules (Chapter 5 and Chapter 6); 3) a summary and future perspective for PR LBD dynamics and drug discovery (Chapter 7).

Chapter 2. Computational Methodologies

2.1 Conventional MD (cMD) simulations

With the advancement of the new technologies and tools of computational biology, our abilities have greatly increased to explore the unknown parts of the "unknown world" of macromolecule organizations and dynamics mechanisms, which would ultimately voyage human beings more close to the ambitious goal – the biological world in silica [44, 45].

Modern sciences witness the rise of computational biology, which has been deeply interplayed with every aspect of biological fields, thus we should ask ourselves what is not computational biology, rather than what is computational biology [46]? The topics of computational biology today have basically involved the majority of the biological fields, such as genomics, protein secondary structure predictions and dynamics, drug discovery and library screening, macromolecules interaction pathway and network, neurosciences, and the other "omics" and any other related fields where computation is required. Among the diverse fields, the valuable structural information, which would be relevant to biological functions and mechanisms, from any aspect, is of general interest to scientists. The understanding of molecular structures, as well as dynamics, offer us the hints of underlying atomic or sub-molecular levels mechanisms of certain biological effects. The structure and dynamics data thus form the foundations of modern biological sciences [47].

Multiple tools, either experimental or computational methods, have been developed to capture a clear picture of macromolecule structures and dynamics at the atomic scale. Among computational tools, the molecular simulations [47, 48] emerge as very powerful tools to explore the structures and energy states of molecules in the biological relevant environment. Here, we provide a general introduction to the popular method: MD simulation.

Mc Cammon *et al* [49] in 1977 reported the first biological system MD simulation study of the folding of a protein, the bovine pancreatic trypsin inhibitor. Since then, MD simulation has been developed as a powerful tool in studying not only biological macromolecules but also some molecules in chemical sciences and material sciences. Ogata, K., and others reported an MD simulation study of the photosynthesis system II [50], which is a signal that we are embracing an era where complicated system simulation becomes possible.

With the accumulating of chemical and physical knowledge, much more complicated details and principles of micro-world behaviors would be gradually elucidated. Meanwhile, as traditional silicon-based computers and quantum computers are evolving quickly, we are going to embrace a theoretic biology future, when every wet-lab experiment could be replaced by in silica experiments. And this

goal, using computers to simulate every level of life science, though difficult to reach, would be achieved finally some day in the future.

At moment, MD simulation mainly focuses on the following five parts: 1) protein (and other biochemical molecules, such as lipids and sugar) structure and dynamics, 2) nucleic acid structure, dynamics, and folding simulation, 3) membrane and transport protein, 4) computer-aided drug design and molecular docking, and 5) large proteins-complex system simulation. We shall offer a relatively detailed explanation about the basic principles of MD simulation [48, 49, 51] in the following part.

2.1.1 Integration methods

In a real situation, besides the influence of potential energy, external factors such as temperature and pressure, also affect the systems. The combination of internal factor (potential energy) and external factors makes molecular dynamics method more resemble the real systems. Molecular dynamics simulation repeatedly calculates the velocity and force on every atom to simulation system motion and other behavior. This method searches to compute total energy composing by potential energy and kinetic energy. (See equation 1)

$$E_{total} = E_{motion} + U_{potential} \qquad (1)$$

$$U_{potential} = U_{int} + U_{vdw} \qquad (2)$$

$$U_{int} = U_r + U_b + U_\theta + U_\alpha + U_{columb} \qquad (3)$$

Potential energy *U* is composed by internal potential energy and van der Waals term potential energy. While U_{int} could be divided to several terms we have describes above. (See equation 2)

The van der Waals part of the total potential energy U_{vdw} is calculated between all non-bonded atom pairs and is the function of distances.

$U_{vdw} = u_{14} + u_{15} + \cdots + u_{1n} + u_{23} + u_{24} + \cdots = \sum_{i=1}^{n-3} \sum_{j=i+3}^{n} u_{ij}(r_{ij})$ (4);while *i* and *j* are the atom identifier numbers, and *n* is the total number of atoms in the system. Thus, total potential energy could be calculated from atom positions, bonding patterns, several potential terms and atom velocities, which are also needed for kinetic energy calculation. Then the potential gradient of each atom is also calculated. If knowing the forces on atom *i* (See equation 5), we can learn the motion behavior of these atoms according to calculation following Newton's Second Law.

$$\vec{F}_i = -\frac{\partial U}{\partial r_i} \qquad (5)$$

$$\overrightarrow{a_i} = \frac{\vec{F}_i}{m_i} \qquad (6)$$

According to equation 6, we calculate the integral with respect to time, then we have equation 7, 8 and 9.

$$\frac{d^2}{dt^2}\vec{r_i} = \frac{d}{dt}\vec{v_i} = \vec{a_i} \qquad (7)$$

$$\vec{v_i} = \vec{v_i}^0 + \vec{a_i}t \qquad (8)$$

$$\vec{r_i} = \vec{r_i}^0 + \vec{v_i}^0 + \frac{1}{2}\vec{a_i}t^2 \qquad (9)$$

If we could know $\vec{r_i}$, together with *U*, then we would be able to know *v*, which is important for us to know molecular motion.

People come up with several methods to calculate v according to the atom position $\vec{r_i}$. Velet method, leap frog method, Beeman method and predicator-corrector method are most frequently used. For example, in the Verlet method, the atom *i* position could be worked out by Taylor's expansion [48]. (See equation 10)

$$r(t+\delta t) = r(t) + \frac{d}{dt}r(t) + \frac{1}{2}\frac{d^2}{dt^2}r(t)(\delta t)^2 + \qquad (10)$$

$$r(t-\delta t) = r(t) - \frac{d}{dt}r(t) + \frac{1}{2}\frac{d^2}{dt^2}r(t)(\delta t)^2 + \qquad (11)$$

When we combine equation 10 and 11, we have the equation 12.

$$r(t+\delta t) = -r(t-\delta t) + 2r(t) + \frac{d^2}{dt^2}r(t)(\delta t)^2 + \qquad (12)$$

Or, the subtractive of equation 10 and 11 is equation 12.

$$v(t) = \frac{dr}{dt} = \frac{1}{2\delta t}[r(t+\delta t) - r(t-\delta t)] \qquad (12)$$

Then according to equation 13, we could know *v*(*t*) is the function of $t+\delta t$ and $t-\delta t$. And the δt is called time-step and is often set as 1 or 2 femtosecond (fs).

In molecular dynamics simulation, it is a continuous iteration process: we repeatedly calculate the motions every 1 or 2 fs time step. After enough time steps, we can have the integral process of all the motion functions to obtain the result of a multiple components system and the trajectory (including positions, momentum and coordinates, and even velocities and energy terms) of the system in space. By multi-step calculations, the system is towards a more stable state theoretically. We can view the multiple steps as sampling statistics. Using the sampling statistics, we could get access to the macroscopic properties, which are based on a continuous process. In the same time, if we use all the states of a system at one defined time point to replace all states a single system going through in a trajectory, time average is achieved and replaced by systems average. And that is how molecular dynamics used to fulfill molecular simulation in a dynamic manner.

2.1.2 Force fields

Nearly all the simulation methods rely on force field for calculation. Force field, a potential energy (equation 14) field in atomic and molecular level, describes the topology and motion behavior of atoms in molecules. When we use the force field to describe the properties of molecules, spectrum constant force field and empirical potential function force field are often introduced. And there are also Dreiding force field and universal force field. (Setubal and Meidanis, 1997)

$$E_{total} = E_{Kinetic} + U_{potential} \qquad (14)$$

In spectrum constant field, the energy is simply linked with the distance between two atoms. (See equation 15, K_i is a constant measured by spectrum data, ΔR_i is the distance between atom *i* and atom *j*.) This only applies for simple molecules such as water molecules.

$$E_{vibrant} = \frac{1}{2}\sum_i K_i \left(\Delta R_i\right)^2 \qquad (15)$$

For multi-atom molecules, internal coordinates are introduced. Four internal coordinates are commonly used as depicted in Figure 3. These four coordinates (bond stretching, bond angle, torsion angle, and out-of-plane angle) do not change when the whole molecule moves in the space, but they are enough to describe the internal motion of the molecules.

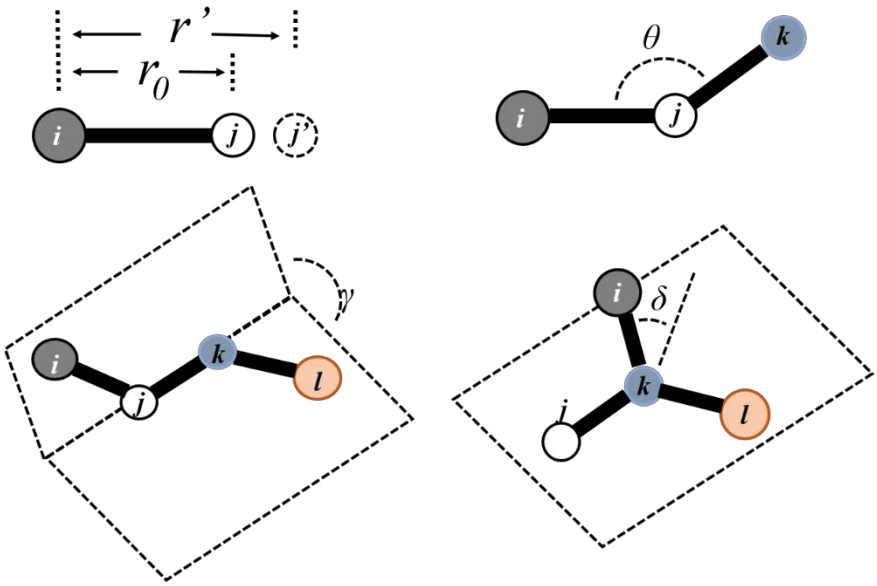


Figure 3. Four types of internal coordinates used to potential energies in MD force fields.

The 4 types of internal coordinates are bond, angle, dihedral angle, and out-of-plane angle.

In other simulation methods, potential energy is composed by several terms: non-binding potential, bonding stretching term potential, angle bending term potential, torsion (dihedral) angle term

potential, out-of-plane bending term potential, columbic interaction term potential (Daan Smit, 2002). For many force fields, the calculation is based on the terms mentioned above. The most frequently applied non-bonding potential is contributed by Lennard-Jones (LJ) potential force. (See equation 16, r is the distance between two atoms, ε and σ are atomic specific constants)

$$U_r = 4\varepsilon\left[\left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^{6}\right] \qquad (16)$$

In equation 16, for the U_r of two atoms A and B, the parameter σ could be approximated by the following equation.

$$\sigma_{AB} = \frac{1}{2}\left(\sigma_A + \sigma_B\right)$$
$$\varepsilon_{AB} = \sqrt{\varepsilon_A \varepsilon_B} \qquad (17)$$

$$U_b = \frac{1}{2}\sum_i k_b\left(r_i - r_i^0\right)^2 \qquad (18)$$

Equation 18 is a common expression of bonding stretching term potential function (a simple harmonic vibration model). r_i is the bond length of the no. ith bond, while r_i^0 is the average bond length of the bond i:

$$U_\theta = \tfrac{1}{2}\sum_i k_\theta(\theta_i - \theta_i^0) \qquad (19)$$

$$U_\tau = \frac{1}{2}\sum_i\left[V_1\left(1+\cos\tau\right) + V_2\left(1-\cos 2\tau\right) + V_3\left(1+\cos 3\tau\right)\right] \qquad (20)$$

Common bending angle term potential is like equation 19. θ_i is the angle between the two bonds on the atom i. The equation 20 concerns the torsion angle term potential energy. While τ is torsion angle using the no. i atom as the center. V_1, V_2 and V_3 are constants.

$$U_\alpha = \frac{1}{2}\sum_i k_\alpha . \alpha^2 \qquad (21)$$

Equation 21 is out-of-plane term potential energy function. k_α is a constant and α is the out-of-plane angle.

$$U_{columbic} = \sum_{ij}\frac{q_i . q_j}{D . r_{ij}} \qquad (22)$$

This columbic term potential energy is the electrostatic energy between the no. i and no. j atom. D is dielectric constant, and r_{ij} is the distance between two atoms. q is the electric charge of atoms.

Another group of the force fields called second generation force fields, integrates more complicated quantum mechanics and experimental results to pursuit more precise simulation. Some examples are CFF91, PCFF, CFF95, MMFF94 [52], etc. There are also some special force fields (such as COMPASS and Gay-Berne) "costumed" for specific molecules, and thus are not explained in detail here.

For water molecules, there are also force fields. The explicit and implicit water force fields are developed. For example, TIP (Transferable intermolecular potential serials, such as TIP3P and TIP4P) force field is an explicit water force field, which is based on comparison with quantum calculation and has very simple formation thus makes it quite popular. SPC (Simple point charge) force field is another example of explicit water force field. The SPC force field, proposed by Berendsen *et al*, treats water molecules as rigid three-atom molecules with defined length and angle of bonds.

2.1.3 Ensemble

The output of the simulation comes in the form of an ensemble of frames. All frames share the same macroscopic/thermodynamic state but may differ in the microscopic states. Each frame represents the system at a specific point of time (a specific microscopic state). If the ensemble is sequence (time) dependent, it is called a trajectory. In this case, the trajectory represents the time-dependent evolution of the system. For Canonical Ensemble (NVT), it contains all possible states in thermal equilibrium with a heat path. The system remains in the absolute temperature T but may exchange energy with the heat path. Three Parameters of the system are fixed throughout the simulation: The absolute temperature (T), the number of atoms (N), and the volume (V). Temperature is the most influential parameter in the system states among them. For Micro canonical Ensemble (NVE), it represents an isolated system. No change in the mass/ number of atoms (N), Volume (V), nor exchange of Energy (E) is allowed. As for Isothermal-isobaric ensemble (NPT), the system has a fixed temperature (T), hence it is isothermal, and fixed pressure (P), hence it is isobaric.

2.2 Advanced sampling methods

The metastable stables with large-scale conformation changes between each other, however, are separated by large energy barriers. Based on the transition state theory, it is an exponential relationship between the time scale of state transition and the energy barrier height [53]. Therefore, the rare dynamical events of interest, such as the folding process of a polypeptide, are often happening in quite large time scale [54]. For example, according to Jan Kubelka *et al*, the folding time scale of a generic single-domain protein from the unfolded state to folded state could be expressed as

following: $N/100$ μs, where N is the number of residues in the protein [54]. In order to obtain the statistical meaningful transition free energies, multiple times of trans-passing between the energy local minima are required according to the assumption of the ergodicity [55], then it must need an extremely long trajectory. Therefore, for larger proteins, using cMD simulations to address the large energy barrier crossing events could be an impossible mission.

Several enhanced sampling schemes have been developed to cross large potential energy barriers. The replica exchange MD method (REMD) [56], solute tempering [57], Hamiltonian replica exchange MD (HREMD) [57], Wang-Landau method and simulated tempering (ST) [58, 59] and so on, are all belong to one general class. Besides, the second class of the enhanced sampling methods, including umbrella sampling [60] and metadynamics [61], relying on introducing bias potentials, could visit the rare states by reconstructing the probability distribution along one or a few well-chosen CVs [62]. What's more, the hybrids of these methods are also proved to be efficient for FES reconstruction. For example, the hybrid Hamiltonian method by introducing the Generalized Born polar solvation energies, is a useful tool to quickly explore the FES which is similar to the standard TRE simulation [63]. Also, the combination of parallel tempering with metadynamics by Bussi *et al* is also proved to be a powerful method [64].

However, the cMD simulations could sample canonical ensembles of conformations in a time-series continuous manner. The kinetics of the macro-molecule could be directly estimated through the cMD simulations. Whereas for almost all enhanced sampling methods, bias potentials were applied to the system, therefor the canonical ensemble could not always be maintained. Meanwhile, the kinetics of the trajectories sampled in enhanced sampling simulations make no sense and could not be directly used. To obtain enough sampling, both cMD and enhanced sampling simulations are capable and useful, though the later one would be more effective.

2.2.1 Umbrella sampling

Umbrella sampling [60, 65] could sample the structure dynamics along one or many reaction coordinates, thus could estimate the free energies of various states in the reaction coordinates. A reaction coordinate (ξ) could be a kind of continuous parameter, either high dimensional or low dimensional. If the reaction coordinate is good enough to differentiate distinct states of a system, by biasing the system along the reaction coordinate, thus one would be able to reconstruct the FES.

For a general umbrella sampling scheme, multiple windows were set starting from different initial structures with a sequential list of reaction coordinate values. Harmonic bias potentials or adaptive bias potentials [66] would be applied to the system along the reaction coordinate. The system in each window would be constrained to sample a narrow phase space along the reaction coordinate to ensure potential energies overlapping between adjacent windows. After simulation completion, a post-

processing method, the weighted histogram analysis method (WHAM) [67], could recover the unbiased free energy profile by the umbrella integration. Multiple integration methods and biasing potential schemes are available, among them, WHAM as the integration method and harmonic biasing potential algorithm are relatively widely popular. The following part is a brief introduction of the general process of umbrella sampling applying harmonic bias potential and WHAM post-processing method.

When we deposit the reaction coordinate dependent bias potentials $\omega_i(\xi)$ to the system in a window, the total potential energies of the biased system could be expressed as following:

$$E_{bias} = E_{unbias} + \omega_i(\xi) \qquad (13)$$

where *i* represents the *i*th window of the umbrella sampling. The harmonic bias potentials applying to the system is expressed as following:

$$\omega_i(\xi) = \frac{1}{2}k(\xi - \xi_i)^2 \qquad (24)$$

ξ_i works as a reference coordinate. During the MD simulations, if the system is escaping the reaction coordinate, then a bias potential is added to push the system back.

To calculate the unbiased free energy, we need to obtain the unbiased distribution of the reaction coordinate, according to the following equation:

$$P_i^u(\xi) = \frac{\int exp[-\beta E(r)]\delta[\xi^r(r)-\xi]d^{N}r}{\int exp[-\beta E(r)]d^{N}r} \qquad (14)$$

Furthermore, the unbiased probability $P_i^u(\xi)$ could be determined by:

$$P_i^u(\xi) = P_i^b(\xi)exp[\beta\omega_i(\xi)]\langle exp[-\beta\omega_i(\xi)]\rangle \qquad (15)$$

From the MD simulation in each window, the biased probability $P_i^b(\xi)$ is known, and the free energy of the window thus could be deduced by:

$$E_{unbias} = -\left(\frac{1}{\beta}\right) ln\, P_i^b(\xi) - \omega_i(\xi) + F_i \qquad (16)$$

where F_i is a constant which could be solved by self-iteration until a convergence reached [68]. Besides, the unbiased free energy could be combined by all the windows to produce a whole free energy along the reaction coordinate.

2.2.2 Metadynamics simulation

Different from the cMD simulations, metadynamics [61] is an enhanced sampling method. By depositing bias potential along the predefined CVs, post-processing of metadynamics sampling could reconstruct the free energy surface (FES) and be able to accelerate the sampling process to capture rare dynamics of the system, such as the large-scale conformational changes (Figure 4).

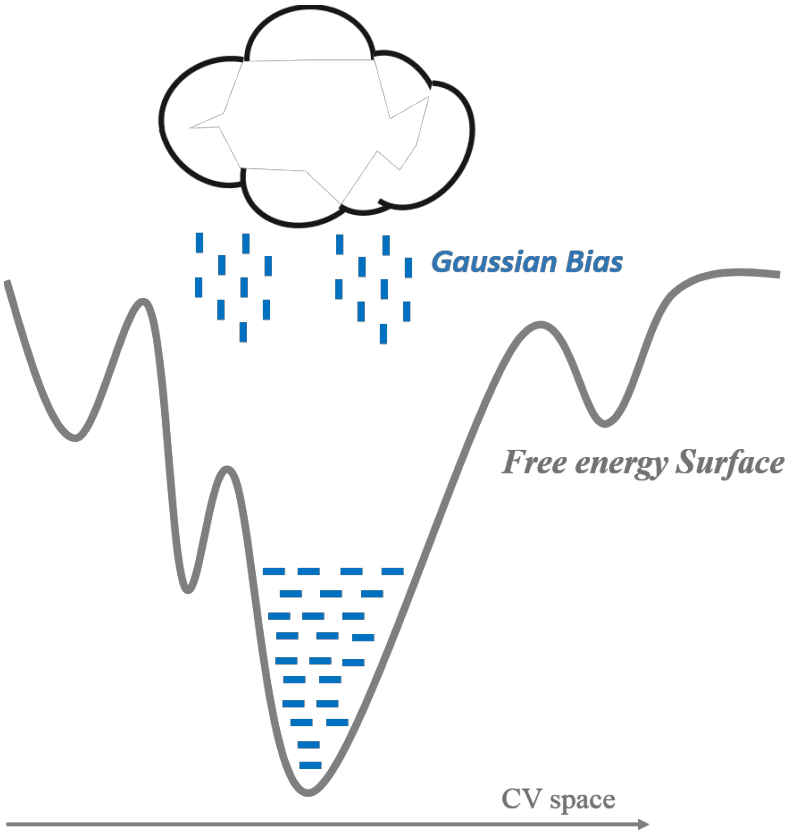


Figure 4. A cartoon representation of metadynamics.

During metadynamics simulations, Gaussian shape bias is added along specific coordinates (or CVs). Upon the convergence of the simulations, the "true" potential energy surface could be recovered through reweighting techniques.

Since the first algorithm of the metadynamics [61], many modified methods of metadynamics have been proposed, such as well-tempered metadynamics [69], the bias-exchange approach [70] and the parallel tempering metadynamics [71]. Besides, various new CVs as reviewed in reference [62], such as widely used path variables, principal component analysis (PCA) variables, alpha root mean squared deviation (RMSD), beta RMSD, and contact map variables, have accelerated the sampling efficiency vastly. In the same time, Plumed 2.1 [72] and METAGUI [73], the VMD plugin for metadynamics analysis, have been developed. These two powerful tools make the simulation and the post-processing of the data much more convenient. For a general understanding of the metadynamics, we take the well-tempered metadynamics [69] as an example, briefly introduce the rationales behind.

Supposing the microscopic coordinates as **R** at temperature *T* with a potential as *U(***R**,*t)*, the ultimate goal of metadynamics is to obtain the FES from the unbiased probability distribution *P(s(R))* along the CV *s(***R***)* by solving the following equation:

$$F(s) = -\frac{1}{\beta} log P(s) \qquad (28)$$

where the $\beta = \frac{1}{k_B T}$, k_B is the Boltzmann constant. For the unbiased $P(s)$, we could have:

$$\mathrm{P(R,t)} = \frac{\mathrm{e^{-\beta U(R)}}}{\int \mathrm{dR e^{-\beta U(R)}}} \qquad (29)$$

By adding a history-dependent bias potential as the following

$$\mathrm{V(s,t)} = \Delta\mathrm{T}\ln(1 + \frac{\omega \mathrm{N(s,t)}}{\Delta \mathrm{T}}) \qquad (30)$$

where $N(s,t) = \int_0^t \delta_{s,s(t')} dt'$ is the histogram of the CV *s*(**R**), ω is energy related term, ΔT is a temperature factor, which we would explain in detail. For this equation, $V(s,t)$ disfavors the frequently visited space in CV *s*(**R**). The first order derivate of the *V(s,t)* could be expressed as following:

$$\dot{\mathrm{V}}\mathrm{(s,t)} = \frac{\omega \Delta \mathrm{T} \delta_{\mathrm{s,s(t)}}}{\Delta \mathrm{T} + \omega \mathrm{N(s,t)}} = \omega \mathrm{e}^{-\frac{\mathrm{V(s,t)}}{\Delta \mathrm{T}}} \delta_{\mathrm{s,s(t)}} \quad (31)$$

Then we use τ_G to replace $\delta_{s,s(t)}$, and let w = $\omega e^{-\frac{V(s,t)}{\Delta T}} \tau_G$, where τ_G is the Gaussian height deposit time step. By incorporating the equation 29 into the standard metadynamics, where the height of Gaussian deposited is a constant, we could modulate the bias potential as a history dependent style.

As the time *t* becomes really large, $\dot{V}(s,t)$ could be determined by $\frac{V(s,t)}{\Delta T}$, which indicates at a position where $V(s,t)$ is large, the $\dot{V}(s,t)$ is close to zero to resemble a thermodynamic equilibrium. At this situation, one might get $P(s,t)ds \propto e^{\frac{-F(s)-V(s,t)}{T}} ds$, based on that, then one could have the following relationship between $\dot{V}(s,t)$ and *F(s)*:

$$\dot{\mathrm{V}}\mathrm{(s,t)} = \omega \mathrm{e}^{-\frac{\mathrm{V(s,t)}}{\Delta \mathrm{T}}} \mathrm{P(s,t)} = \omega \mathrm{e}^{-\frac{\mathrm{V(s,t)}}{\Delta \mathrm{T}}} \frac{\mathrm{e}^{\frac{-\mathrm{F(s)}-\mathrm{V(s,t)}}{\mathrm{T}}}}{\int \mathrm{ds}\, \mathrm{e}^{\frac{-\mathrm{F(s)}-\mathrm{V(s,t)}}{\mathrm{T}}}} \quad (32)$$

While the time $t \to \infty$,

$$V(s) = -\frac{\Delta T}{T + \Delta T} F(s) \qquad (33)$$

If we go a step further, we could have $F(s) + V(s) = \frac{T}{T+\Delta T} F(s)$. Until now, we have established the relationship between $V(s,t)$ with *F(s)* in a simpler way. Combining equation 32 with equation 33, we then can deduce the estimation of *F(s)*:

$$\tilde{F}(s,t) = -\frac{T+\Delta T}{\Delta T} V(s,t) = -(T + \Delta T)\, ln(1 + \frac{\omega N(s,t)}{\Delta T}) \qquad (34)$$

From this equation 34, we could estimate the final FES by calculating the histogram $N(s,t)$ of CV *s*(**R**). The modulation of the constant ΔT could define different behavior of the Gaussian deposit. For $\Delta T = 0$, the bias potential added in the system is always 0 along the simulation time *t*. Another

limiting case, if $\Delta T \to \infty$, then $-\frac{T+\Delta T}{\Delta T} \to -1$, thus we could give the conclusion that $\tilde{F}(s,t) \approx -V(s,t)$, which is the case in the standard metadynamics where the bias potential energy, added in the system along the CVs space, has the same absolute value as the FES along the same CVs space. What's more, a finite value of ΔT, therefore restricts the system from exploring all the physical space, and ensures a thermodynamic equilibrium at local minima in FES. Fine tuning of this ΔT could facilitate us to better explore the CVs space. By calculating the $V(s,t)$ in the following equation:

$$V(s,t) = w\sum_{t'=\tau_G,2\tau_G,3\tau_G,\cdots} e^{\left(\frac{-s(\boldsymbol{R})+s(R_G(t'))}{2\delta s^2}\right)^2} \qquad (35)$$

where w and δ have been defined already. Finally, we could recover the $\tilde{F}(s,t)$ from the biased probability distribution of s(**R**).

In practice, we often use the following bias factor $\gamma = \frac{\mathrm{T}+\Delta\mathrm{T}}{T}$ to set up the ΔT. τ_{G} is the Gaussian deposition stride, and another input value ω in $\mathrm{w} = \omega \mathrm{e}^{-\frac{\mathrm{V(s,t)}}{\Delta\mathrm{T}}}\tau_{\mathrm{G}}$, is the initial Gaussian height. And the δ is the width of the Gaussian for the CV s(**R**).

2.2.3 REMD

Quite long time MD simulations are required to recover the biological meaningful kinetics and large conformational changes It is relatively common that the simulation system would be trapped in local energy minima, which lead to low sampling efficiencies. To overcome energy barriers and search for the global energy minima, the REMD method, sometimes also called generalized ensemble method, is emerged as one efficient sampling technique [74-77] which has been demonstrated to be more powerful than the cMD [74, 78].

The idea was based on the hypothesis that at a higher temperature, the potential energy of the simulation system is much higher and is more easily to escape the local energy minima to explore global energy minima (Figure 5). The potential energies of the parallel systems (replicas) in different simulating temperatures may overlap, thus the low energy conformations from high-temperature replicas could be exchanged to lower temperature replicas and could be stabilized and sampled in lower temperature replicas. This way, in the normal temperature range, the replicas could sample the structures with low potential energy. The sampling efficiency, therefore, it suppresses the conventional single trajectory MD simulation.

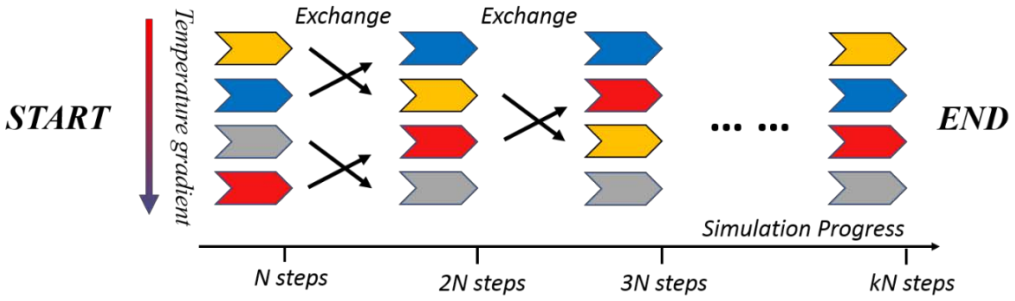


Figure 5. A cartoon representation of REMD methods.

In practice, a number of *N* replicas in the canonical ensemble are simulated with fixed temperature independently and spontaneously for some MD (or MC) steps, though the optimal number of steps between exchanges are controversial [79-81]. Then, the adjacent replicas (say *m* and *m*+1) could exchange their configurations based on the Metropolis criterion [82]. Once the exchange occurs, replica m adopts the configuration from replica *m*+1 and continues at the fixed temperature. The same goes for replica *m*+1 as well.

Here, we briefly illustrate the exchange method used in REMD. The Metropolis criterion specifies that the exchange between neighboring replicas *m* and *m*+1 at temperature T_m and T_m+1 could take place with an acceptance probability *P* to ensure the detailed balance,

$$P = e^{(\beta_m - \beta_{m+1})(F_m - F_{m+1})} \qquad (36)$$

where F_m and F_m+1 are the potential energies of replica m and replica *m*+1. And $\beta_m = 1 / (k_B T_m)$, where k_B is the Boltzmann's constant. To achieve an acceptable exchange ratio, certain temperature intervals between replicas should be selected to ensure enough overlapping of the potential energies of neighboring replicas [74, 83].

The convergence problem is one of the major issues to achieve a statistically meaningful REMD simulation [84-87]. It has been widely discussed in many types of research and reviews, therefore we are not going to illustrate it in detail here.

2.2.4 Hamiltonian REMD

The conventional REMD tends to sample low energy structure and then deposits it in the low-temperature replicas, thus it is blind sampling and requires no prior knowledge of the simulation system. During REMD simulation, the number of required replicas to ensure enough potential energy overlapping between adjacent replicas is exponentially increased along with the system size, therefore relays on a tremendous amount of computational power, and increases the difficulty of convergence, thus those limit the usage of REMD for huge simulation systems. Instead of exchanging between different temperatures, Hamiltonian could also be exchanged between replicas as long as there are

possible potential energy overlapping. Hamiltonian REMD is quite similar to REMD, where each replica samples independently in different environments (temperature for REMD, Hamiltonian for Hamiltonian REMD respectively). The exchanges every *N* step in Hamiltonian REMD also obey the metropolitan rule.

For Hamiltonian REMD, the Hamiltonian of each replica is scaled by a factor lambda, and ensure the first replica has a normal Hamiltonian, where the last one has the largest or smallest lambda. In the same time, enough overlapping of potential energies between adjacent replicas must be retained. There are several ways to modulate the Hamiltonian of simulation systems, for example, reducing the atomic changes of a molecule, increasing intra-molecular repulsions [88], decreasing interatomic attractions, or increasing molecular hydrophobic interactions. The combination of Hamiltonian simulation with REMD is proved to be a good practice. Using different Hamiltonian, we could bias the systems towards a different direction. Hamiltonian REMD thus is not blindly sampling, nor the CV based sampling, but something in between.

2.3 Molecular docking and virtual screening

Nowadays, the development of a new drug requires hundreds of million dollars and around 15 years, and a very high failure rate. The extremely high expenses of drug development are majorly composed of the money used in clinical trials, while the research and development part only takes ~20% of the costs. Although 20% is not a particularly large number, considering the giant size of the overall expenses, it is still a significant amount. Thus, it is worth trying to low down the costs in the R&D process, whereas the money spent in clinical trials is inevitable due to more strict regulations on new drug approval.

Computer-aided drug discovery is a good strategy to decrease drug discovery costs and accelerate the R&D process in meantime. And this chapter mainly focuses on this topic. The general types of drugs are small molecule drugs, peptide drugs, and polymer drugs. Their major targets are proteins. Thus, taking the small molecule drug discovery pipeline as an example, the first step is the identification of the target, generally, a protein. There are thousands of proteins in human cells, however, only part of them are suitable for drug targets. For currently approved drugs, their targets are mainly G-protein coupled receptors, nuclear receptors, ion channels. Proteomics and genomics approaches equip us the ability to identify key proteins as drug targets in an economic affordable manner. To verify the targets, gene knock-out and knock-in experiments, small inhibitors or agonists tests, or other biofunction analysis would not be enough. Further structure models from the NMR method or X-ray crystal methods of the target proteins are necessary for a detailed understanding of the dynamics and kinetics. The characterization of the target proteins thus is the essence of the target validation, which is the bottleneck of the drug discovery pipeline. The next step of drug discovery is lead discovery. Once we have a specific target, say a protein, then high binding affinity antibodies, or small molecules are

randomly screened, or designed specifically. To select useful small molecule ligands from large libraries of small molecules for further analysis, high throughput screening (HTS) could be performed using well-established platforms, with an efficiency of around several thousands of ligands per day [89, 90]. Though costly, HTS is proved to be efficient and helpful. Another random selection method, high throughput virtual screening (HTVS), by applying search algorithms, commonly called "docking" simulation, could speed up the screening process and low down the costs dramatically. Through HTS or HTVS, we find some potential "good-binders", however, which are not qualified as "leads". Further tests, modifications of the "good-binders" are necessary to have optimal properties, such as pharmacodynamic properties, pharmacokinetic properties, chemical optimization potential, physicochemical properties, and patentability. Once leads are identified, pre-clinical and clinical trials would be performed.

During the years-long drug development discovery process, the lead discovery stage could be accelerated through computational methods, which are the focus of the following sections.

2.3.1 Virtual Screening and multiple target docking

VS aims at predicting binding affinities of receptor-ligand complexes and enriching the druggable small molecules. There are majorly two strategies in VS: ligand-based virtual screening (LBVS) [91, 92] and structure-based virtual screening (SBVS) [93, 94].

For LBVS, known active binders of a receptor or target, whose 3-dimensional structures are generally not available, are used to screen "similar" molecules from large ligand libraries, based on the hypothesis that similar ligands may have comparable biological actives [92]. Though the hypothesis is held in most cases, there are situations that small modifications in a ligand may result in dramatic changes in function. In practices, similarity metric searching, substructure/superstructure search, quantitative structure-activity relationship (QSAR) models and other machine learning classifications are popular solutions in LBVS [92]. For current SBVS methods, large quantities (from some thousands to millions) of ligands are docked to receptor structures, which are obtained either through experimental methods or from homology modeling, sequentially or parallelly. The binding affinities between these ligands and the receptor are estimated using scoring functions and then are ranked. The top-ranking ligands thus are utilized as lead-like molecules. In SBVS, it is extremely difficult to accurately approximate the experimental binding affinities, therefore this raises the high false positive ratio problem. The trends towards reducing the false positive rate, could be majorly grouped into two directions, 1) incorporating ligand and receptor flexibility, and 2) developing more accurate scoring functions.

There are a lot of docking software available based on diverse docking and scoring algorithms. Most of these packages using rigid body docking, except several of them accounting for the ligand

flexibility using rotatable bonds, as well as receptor flexibility [95-97], which is extremely computationally intensive and time-consuming. Another solution towards receptor flexibility is using an ensemble of receptor conformations to perform virtual screening and choose the common good small molecules. The rationale behind lies on the hypothesis that multiple receptor conformations could represent a set of receptor states that binds the ligand well [98]. Either combining MD simulation snapshots [97, 98], Normal Mode Analysis (NMA) [99], essential dynamics analysis models (EDA) [100], the ensemble receptor docking accounting for receptor flexibility is proposed to increase enrichment rate and decrease false positive rate.

The scoring function is the core of docking simulations. The development of more accurate scoring functions is still a hot topic in SBVS. Several popular scoring functions have been developed, and there are majorly three types of scoring functions, the knowledge-based, force field based and empirical based scoring functions. For example, GoldScore [101], D-score [102], DOCK [102] and AutoDock [103] are representative force field based scoring functions, while PMF04 [104], DrugScore [105] and ASP [106] are typical knowledge-based scoring functions. And X-score [107], ChemScore [108], LUDI [109], and FlexX/F-score [102] are popular empirical scoring functions. Need to mention, though there are many scoring functions are available, these scoring functions are only ready to use in combination with their respective docking packages. Besides, machine learning based scoring functions have also been developed to improve predictive accuracy [110].

To access the success of a virtual screening trial, if we know the number of hits in the screening library, or could estimate the number, then we may use enrichment factor (EF) and receiver operating characteristic curve (ROC) to quantify it [111]. Here, *EF* is defined in this equation:

$$EF = \frac{Hits_{sel}}{Hits_{tol}} \frac{N_{tol}}{N_{sel}} \qquad (37)$$

We select a subset of ligands, generally top 0.1% ~ 10% performing ligands from the screening library with a total N_{tol} ligands, and N_{sel} is the number of the ligands in the subset. And $Hits_{sel}$ is the active compounds (we know these ligands before docking) in the subset, and $Hits_{tol}$ is the total number of active molecules in the screening library. *EF* is used to estimate the ability of the virtual screening finding a small fraction of active molecules from large "bad" binders.

ROC describes how the true positive rate increase as false positive rate rises. Generally, for a classification model, the true positive rate and false positive rate compose the y-axis and x-axis of ROC curve respectively. A reference line along the diagonal would represent a totally randomized classification model. The area under the ROC curve (AUC) is used to assess the fidelity of the binary classification model, the larger the value of AUC, the better the model is. ROC is a popular assessment in docking scoring function evaluation.

2.3.2 Overview of molecular docking

Making use of computational tools, such as docking, would vastly short the ligand library screening routine, since finding the binding pattern of a small molecule with its target takes less than a minute. Parallel docking simulations thus enable a large number of small molecules being examined using docking. The interactions between macromolecules, or between macromolecules and ligands, are of vital value to understand biological structure-function relationships, as well as to facilitate pharmaceutical researches. Prediction of protein-ligand interactions and binding affinities are key tasks in drug discovery. Although it has come to an agreement that other than the binding affinity, the drug residence time is more important in protein-drug dynamics and kinetics, it is still valuable to have the binding affinity as one of the key criteria of "good" drugs. Docking simulations target to solve the above two problems, the binding pattern, and affinity.

In general, docking simulations are composed of two steps, "docking" a molecule to its receptor (target) and "scoring" the receptor-ligand complex to estimate the binding affinity. The "docking" process, though based on diverse algorithms in different docking packages, aims at thoroughly exploring the conformational space of the ligand in the defined receptor binding pocket, or at the surface of the macromolecule. Currently, the most commonly used searching algorithms in "docking" process include Monte Carlo (MC) method, genetic algorithm (GA), matching algorithms, incremental construction algorithms (IC), and MD simulation algorithms.

Here, we briefly explain the basics of the searching algorithms, as well as their applications in docking packages. MC method is based on random searching. In each step, a random movement is made, and then it is evaluated based on the energy difference the new movement and the previous state. The new movement is accepted according to a probability calculated from the following equation:

$$P = e^{-\beta(E_{new} - E_{old})} \qquad (38)$$

The genetic algorithm was inspired from the evolution of organisms and mainly relies on "mutation" and "crossover". In practice, the configurational space of ligands in the receptor binding pocket is firstly randomly searched, and then this information is reformed to be a "chromosome", where the mutations are actually the movements of ligands. The steps are repeated until the convergence condition met. The several popular docking packages are listed in table 1.

Table 1. Commonly used protein-ligand docking packages

Package Names	Search Algorithm	Website	References
GOLD	GA	http://www.ccdc.cam.ac.uk/products/life_sciences/gold	[101]

AutoDock	MC/GA	http://autodock.scripps.edu	[103]
DOCK	IC	http://dock.compbio.ucsf.edu	[102]
Glide	Hybrid	http://www.schrodinger.com	[112]
AutoDock Vina	MC	http://vina.scripps.edu/	[113]
Surflex	IC	http://www.tripos.com/index.php	[114]
FlexX	IC	http://www.biosolveit.de/flexx	[102]
LigandFit	MC	http://accelrys.com/products/discovery-studio	[115]

2.4 Machine learning based classifications in SBVS

Recent fast advancements of the computational hardware enable large parameter optimization models more affordable. The training of supervised learning models, such as machine learning models and deep learning models, which generally could be treated as parameterization problems, for classifications and regressions, becomes more doable. Especially, the graphics processor units (GPU) show tremendous ability in numeric integrations and make them more suitable for machine learning or deep learning, by harnessing the computation power of multiple cores. However, the advantages of GPUs are not only for supervised learning calculations, but also for molecular simulations, and any other numeric calculation intensive methods. And the large parallel GPU clusters make it even easier to train supervised learning models in relatively short timescale. Therefore, the power of these supervised learning methods has been greatly intensified.

There have been studies applying machine learning models in computational drug discovery a long time ago. The successful applications of the machine learning models demonstrate the promising future towards automated, high-throughput, and accurate computation drug discovery. Here, in the following sections, the basic theories of commonly used machine learning algorithms are presented, and the combination of machine learning models and drug discovery is also reviewed.

2.4.1 Basic theories of common machine learning algorithms

Opposed to the machine learning methods, other methods, such as clustering methods, principal component analysis, t-distributed stochastic neighbor embedding (t-SNE), and LDA, could be applied to group data using unlabeled data, are then classified into unsupervised learning methods. The strength of machine learning methods, such as k-nearest neighbor (*kNN*), Bayesian classification, support vector machine (SVM), decision tree (DT), linear regression, artificial neuron network (ANN)

and so on, lies in the fact that they enable us the ability to predict unseen data using labeled training data. And due to this reason, these methods are also called supervised learning methods.

Supposing we have a cleaned dataset X with n columns (or called features) and m rows (or called cases), and a label vector Y representing the group information (for easy understanding, y=0 or 1), for simplicity, we use these simple datasets to explain the machine learning algorithms (Figure 6). And we have to predict the group information of another dataset Z, having n features and l cases, based on the previous dataset. Generally, we would divide the dataset X into a training dataset X_{train} and a testing dataset X_{test} by assigning cases into the two datasets randomly. Grouping vector Y are also split based on the same dividing scheme, to ensure the one-to-one pairing between X and Y. The variables (X, X_i, Y, P, and Q) represented by capital characters are vectors, or matrices, while those presented using lower cases are values (m, n, l, x_i, y_i).

$$X=\begin{bmatrix} X_1 \\ \vdots \\ X_i \\ \vdots \\ X_m \end{bmatrix} = \begin{bmatrix} x_{1,1} & \cdots & x_{1,n} \\ \vdots & & \vdots \\ x_{i,1} & \vdots & x_{i,n} \\ \vdots & & \vdots \\ x_{m,1} & \cdots & x_{m,n} \end{bmatrix} \quad \Longleftrightarrow \quad Y = \begin{bmatrix} y_1 \\ \vdots \\ y_i \\ \vdots \\ y_n \end{bmatrix}$$

Figure 6. The pairing relationship between X dataset and its labeling vector Y.

With datasets provided, one should train the model, and validate the model, then utilize the model to predict unseen data, Z. The goal of the training is parameter optimization to ensure that the loss of X_{test} is minimized. And the loss function is defined using mean square error *MSE*:

$$MSE = \frac{1}{m}\sum_{i=1}^{m}\left(y_i - predication(X_i)\right)^2 \qquad (39)$$

where X_i is the ith case in X_{test}, and y_i is the ith value of the grouping vector Y. After optimal parameters obtained, the mature model with the optimized dataset thus is used to predict the grouping information of the prediction dataset Z. Based on the above dataset and training process, a brief introduction of the machine methods is provided.

k-NN method is often regarded as the most easily understood machine learning methods due to its straightforward rational behind. For a case Q in X_{test} dataset, a distance matrices d, such as Euclidean distance, is used to assess the distances between this case with respect to all the cases in X_{train}. The grouping label of this query Q case is the most frequent group of the first k nearest sample cases in X_{train}. A general format of the Euclidean distance takes the form of the root squared distance between feature values:

$$d = \sqrt{\sum_{i=1}^{n}(q_i - p_i)^2} \qquad (40)$$

where q_i is a feature value in Q, and p_i is a feature value in a reference case P from X_{train}. In *k-NN* algorithm, only k is required to be optimized.

Linear regression algorithm is a popular method in machine learning for its simplicity. The linear combination of the original features creates a prediction Y':

$$Y' = b + W \cdot X \qquad (41)$$

while W is the weight vector with a size of n, and b is the bias. Training of the model towards small MSE would explore a suitable W and b. And a proper initialization of the W vector would also affect the final convergence of the parameter vectors. To obtain the optimal W vector, the MSE needs to be minimized through numeric methods, such as "stochastic gradient descent" (SGD) scheme, iteratively updating the W until the loss is converged. Meanwhile, to avoid overfitting, another regularization factor is added to the loss function in addition to *MSE* to restrict the size of the W values. L_2 regularization is often used, and is defined as following:

$$L_2 = \lambda \sum_{i=0}^{N} |w_i|^2 \qquad (42)$$

While λ is a constant which controls the contribution of L_2 to loss function.

SVM transforms the original dataset X into a higher dimensional phase space and find a hyperplane that would divide two classes as widely as possible. Maximizing the distance between any nearest points to a n-1 dimensional hyperplane thus could linearly separate data points into two classes. By implementing non-linear kernels for data transformation, SVM could perform non-linear classifications. Some commonly used kernels include rbf, linear, logistic, poly, and Gaussian. A detailed comparison of the kernels and their properties is reviewed in this reference [116].

The decision tree is a popular non-linear non-parameter binary classification model. It has two major advantages, easy to construct and easy to interpret. Starting from the most informative feature, a threshold is calculated, the dataset could be divided into two leaf nodes. For each of the leaf node, if all cases belong to one class, then the classification for this leaf is completed, otherwise the cases in this node will be further divided based on non-used most informative features, and this step is repeated until the samples in the last level of sibling node are all belonged to the same class. An ensemble of the DT could form a random forest (RF) model, thought the training dataset are randomly partitioned into small subsets to feed the RF. RF is rather robust and widely accepted classification model.

ANN methods are getting popular in recent years. They have been widely applied in image classification, speech recognition, and text processing. By simulating the biological "neurons", or perceptrons, the ANN model could be formed by a lot of neurons with multiple layer structures, thus sometimes they could be called multilayer perceptron (MLP). A perceptron receives an input (or multiple inputs) and emits an output (Figure 7A). The emission is determined through a kernel

function, which could be the arbitrary kernel, sigmoid, exponential or other types. The connections between perceptrons are edges, and weights are assigned to each edge to adjust the learning process of the MLP (Figure 7B). By gradually adjusting the weights little by little, MLP could achieve great classification accuracy. Due to the large parameters in an MLP, the learning process of the model is non-trivial. In one hand, we need to have strong predictive power, meanwhile, we also need to avoid overfitting problem, therefore, a cost function combined RMSE and regularization term would be necessary. The most popular learning method in ANN is SGD method, which greatly lowers down the learning time and resources consumption, and it could also find an optimal minimum, though may not necessarily be the global minimum.

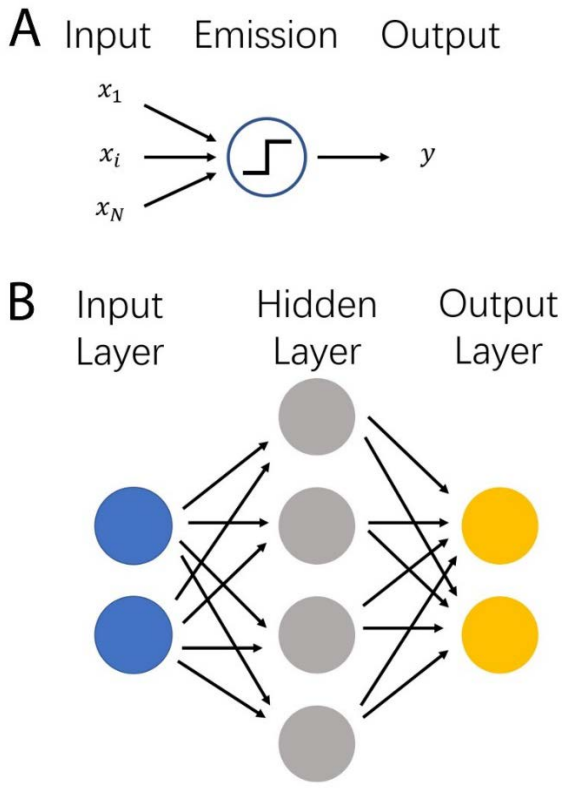


Figure 7. A simple cartoon representation of a perceptron and ANN.

2.4.2 Machine learning in VS

Data mining methods, including learning models, have long been applied in the drug discovery field, for an understanding structure-activity relationship, lead generation, lead optimization, SBVS, and drug toxicity [117-122]. In general, cheminformatic descriptors (or called features) of the ligands, or ligand-receptor complexes, are gathered to train labeled dataset to classify the molecules based on a property, such as bind or not-bind, toxicity, and binding affinity. The trained model thus could be applied to predict new molecules. Current HTVS has difficulties to accurately predict the real-world

binding affinity data and sometimes could result in high false positive rates for the binder and non-binder classifications. Therefore, the successful applications of machine learning methods in drug discovery, especially in SBVS, pave the way for low false-positive lead discovery. Applying machine learning based scoring functions, we could improve the accuracy of current docking-based virtual screening.

Early machine learning based scoring function tried to apply the nonlinear model to construct an additive form model to differentiate actives from the inactive molecules [110, 123, 124]. The earliest machine learning based scoring function adopted the Kernel partial least squares (K-PLS) model with RBF kernel to train two small dataset of protein-ligand complexes [125]. Later, the kNN model was also used to work a scoring function and achieved quite good results [126]. ANN model was introduced and coupled with detail interaction features (van der Waals interactions, columbic interactions, and cheminformatics features) to rescore protein-ligand complex poses [127]. Besides, SVM and RF were also proved to be effective in classifying actives and inactives [128, 129]. In 2010, NNScore [130], which is a neuron network-based scoring function, was introduced by Durrant and McCammon to rescore docking poses resulting from AutoDock Vina [113] docking results. RF-based models, such as B2BScore and SFC-Score, showed optimal enrichment efficiency of the PDB id benchmark dataset using both structural features and physio-chemical properties [131, 132]. The above studies all use diverse protein-ligand complexes, but these universe datasets could have trouble to predict unseen receptor-ligand complex [110]. The hypothesis that learning from a subset or a specific family receptor-ligand would perform much better for the unseen complexes of this subset or family of targets, though it is quite difficult to claim that family specific scoring function would out-perform the generic scoring function [110]. Decomposing the receptor-ligand interactions and physio-chemical properties is a general strategy for learning features, the molecular interaction energy components (MIECs) features in combination with SVM outperform all classical scoring functions. The MIECs features are composed of hydrogen bonds, desolvation energies, van der Waals interactions, electrostatic energies, and the nearest distance between the target protein atoms and the ligand atoms. This research achieves 0.99 AUC in two datasets [133].

The choice of the molecules' descriptors is critical for a successful machine learning based scoring function in SBVS [134]. The descriptors could represent the 3-dimensional structure information of ligands, or ligand-receptor complex quantitatively, using well-established packages and online servers, such as Xue descriptor set [135], Molconn-Z [136, 137], DRAGON [138], MICEs [133], and JOELib [139]. Thousands of these descriptors, which have been thoroughly discussed in this reference [140], include abundant information for ligand classifications, however not all of them are required for a machine learning model. Noise may arise from these descriptors, due to the high redundancy and overlapping of the descriptors. Thus, proper filtering and engineering of the descriptors are necessary, by applying feature selections using simulated annealing-based approaches [141],

recursive feature eliminations (RFE) [142], genetic algorithm based approaches [143], or simple but robust dimension reduction methods (such as PCA, tSNE and manifold learning methods). Contrast to the impression that model detailed features may contribute to higher predictive power, it was proved by Ballester et al, that more fine-grained features do not necessarily lead to a more accurate model, due to the systematic errors in the binding pose prediction, the modeling assumption, as well as the conformational heterogeneity in dataset [144].

The classification models, which show promising results with training sets, thus would be adopted for predicting unseen complexes, possibly generated using docking software, such as GOLD or Glide [133, 145]. The combination of machine learning methods with virtual screening is, actually in most cases, a re-scoring strategy which re-assess the binding pose generated from docking or MD simulations. There could be regression models predicting binding affinities given experimental binding affinity data, such as Ki values, or predicting the binding strength, such as weak, medium or strong [110, 133]. Fast development of the machine learning scoring functions (re-scoring methods) enables more precise prediction power in future in SBVS.

2.5 Applications of the computational methods

Molecular dynamics have widely been adopted in molecular modeling from organic chemicals, short peptides [146-151], protein-protein interactions [152-154], lipid-protein complexes [146-151, 155-158], even virus capsid [159, 160]. By harnessing the great computation power from the supercomputing facilities, the physical properties of sub-micrometer particles could be revealed. Using MD simulations, as well as other simulation methods, stable conformations, the folding process, energy landscape, transitions between macro-states, enzyme catalytic mechanisms, and aggregation states could be examined. Besides, MD simulations are also commonly used in computational drug discovery and design.

For example, the enhanced sampling methods, umbrella sampling, and metadynamics simulation could explore the free energy landscape of a system along with some specific collective variables. Bias potentials are deposited in simulations. The post-processing reweight techniques would then recover the true energy landscape of the simulation system along selected dimensions. The combination of metadynamics with REMD or parallel exchanges between replicas biasing difference collective variables are proved to be more efficient and converges faster than the single-trajectory metadynamics simulations.

By conducting multiple short MD simulations, Markov state model (HMM) could be applied to analyze the trajectories and compute the transition dynamics between macro-states of the system. [161-163] Starting from different conformations, it is expected that multiple short simulations starting in different directions are more powerful in conformational sampling than a single long trajectory.

Combining the short simulations, several macro-states of the system could be identified using different methods, such as clustering and independent component analysis. The transition probabilities between the microstate then would be derived from the trajectories. More importantly, the transition timescales could be estimated.

What's more, combining MD simulation with quantum calculation, more accurate enzyme activation mechanisms. [164-168] The statistic properties could also be calculated and compared to experiments.

2.5.1 MD Simulation and Molecular Docking

Molecular docking is a simplified form of MD simulation. It can be used on intervals to replace lengthy segments of MD simulation trajectories, especially in cases where certain domains undergo large translations, rotations, and conformation changes. A typical example is biological interactions that include large protein folding, like capsid or vesicle formation.

2.5.2 MD Simulation with Markov State Model (MSM)

The dynamics properties would be a great value to understanding the structure-function relationships of macromolecules. MD simulation method has its innate advantage over current experimental techniques in exploring macromolecules kinetics and dynamics. By post-processing of large quantity of MD simulation trajectories using MSM, or Hidden Markov State Model (HMM), the dynamic properties of proteins, as well as other macromolecules, could be estimated with near to experiment accuracy. [161-163, 169, 170]The protein-ligand binding dynamics could also be estimated using the MD-MSM strategy [171].

Chapter 3 Explorations of *apo*-form PR LBD conformations

The structure-function relationship is always a charming area in biological sciences. To have a better understanding the molecular machinery, the efforts of solving the structures of macromolecules, using NMR, X-ray crystallography, or the rather young method Cryogenic Electron Microscopy (CryoEM), have never been decreased. We are also quite interested in the structures of PR, especially the LBD, in a dynamics and biologically relevant context. However, currently, only limited snapshots have been captured using X-ray crystallography method, which in most cases leads an unphysical environment where free solvents are void. The crystal structures of PR LBD could be divided into two groups, the ligand-bound agonistic states, and the ligand-bound antagonistic states. Although these structures are important and informative, the *apo*-form PR LBD structure would be equally important. Current experimental methods are not suitable for the *apo*-form PR LBD structure construction, partially due to the potentially high flexibility of the domain. Therefore, we employed MD simulations to explore the *apo*-form PR LBD in a water environment with the biological relevant salt condition. The cMD and enhanced sampling simulations enable us a clear picture of the flexible instinct of the PR LBD, and the possible metastable conformations of PR LBD, which may be useful for SBVS in future.

3.1 Materials and Methods

3.1.1 cMD Simulations

cMD [172] simulation as a well-developed method, is widely used in understanding macromolecules' dynamics. cMD simulations were adopted in this study, to elucidate the dynamical behaviors of *apo*-form PR LBD starting from various initial conformations. The cMD simulations were completed with Gromacs 4.6.7 package [173]. The force field for PR LBD was Amber99SB-ildn [174].

After removing existing peptides (if there are any), ligands and crystal water molecules, the crystal agonistic (PDB ID 1A28, chain A) and antagonistic (PDB ID 2OVH, chain A) conformations were utilized as initial structures. Thus, two *apo*-form PR LBD systems were constructed (see Table 2). Also, two other systems of *holo*-form PR LBD short cMD simulations were performed. For ligands (progesterone and asoprisnil), the atomic charges were determined using Gaussian09 package [175] with HF/6-31G* basis set following by restricted electrostatic potential (RESP) charge fitting using Amber11 [176]. And tleap module in AmberTools 12 was used to generate ligands topology, while the bond, angle and dihedral angle parameters were adopted from the Amber general force field gaff [177].

For the *apo*-2ovh system, the missing heavy atoms, as well as the missing residues in the original PDB file, were added using SwissModeller web server (http://swissmodel.expasy.org/). The shortest

distances between the protein surfaces and the edge of the solvation box were set as 1.0 nm to avoid possible contacts with images, and the protein was fully solvated with TIP3P [178] water molecules. Na+ and Cl- ions were added to neutralize simulation systems and modulate the salt concentration at 0.15M. The deepest descent 1000-steps energy minimization was carried out, following by an heavy-atom position restraint *NPT* equilibration in Berendsen barostat [51] (1 ns at the 1atm pressure and 300 K) using force constant of 1000 kJ/(mol·nm^2). The v-rescaling [179] *NVT* ensemble simulations at 300 K were performed as production runs. The long-range electrostatic interactions were calculated using the PME algorithm [180] with a 1.0 nm cutoff. Meanwhile, a 1.2 nm cut-off was adopted for short-range van der Waals interactions. All the hydrogen/heavy-atom bonds were fixed according to SHAKE scheme [181], while non-hydrogen the covalent bonds were constrained with the LINCS method [182]. Three repeat runs were carried out by assigning random initial velocities.

Table 2. The setup of the 4 systems in CMD simulations of *apo*-form PR LBD.

S/N	System Name	Initial structure	Ligands	Co-peptides	Simulation Time
1	*Apo*-1a28	1A28, A	NA	NA	2000 ns
2	*Apo*-2ovh	2OVH, A	NA	NA	2000 ns
3	*Holo*-1a28	1A28, A	P4	NA	200 ns
4	*Holo*-2ovh	2OVH, A	Asoprisnil	2OVH, B	200 ns

The dPCA analysis [183] of the last 80% of the 3 repeats trajectories were performed. Low energy basins representative structures were extracted based on clustering analysis. All the frames located low energy basins were firstly extracted and then were clustered to discover the most populated groups, where the central structure was used as the representative structure. The gromos clustering method [184] implemented in Gromacs g_cluster utility was applied to grouping similar conformations with a 0.2 nm αC RMSD cutoff. A π-cation interaction [185] is defined such that the side-chain centroid distances within 0.6 nm and planar angles within 60˚ to 120˚ between an aromatic residue and a positive charged residue.

The helical crossing angle between two α-helices was formed by the helical axis. The vector connecting the center of mass (COM) of starting residues and ending residues defines a helical vector. In this study, residues 883-886 (the start end), residues 894-897 (the end point) form the helix 11 vectors. Similarly, COM of residues 712-715 and COM of residues 730-733 was selected for helix 3 vector. Potential of mean force (PMF) analysis [186], was computed based on the probability distributions of some pre-selected reaction coordinates. The PMF is defined as following: $PMF = -k_B T \ln \frac{P_i}{P_{max}}$, where T is the temperature and k_B is Boltzmann constant, while P_i is the probability distribution along some reaction coordinates and P_{max} is the normalization factor.

The cross-correlation coefficient was calculated using Wordom 0.23 [187, 188]. The cross-correlation coefficient (ccc) C_{ij} is defined here:

$C_{ij} = \frac{\langle \Delta\vec{r}_i \cdot \Delta\vec{r}_j \rangle}{\sqrt{\langle \Delta\vec{r}_i^{\,2} \rangle \langle \Delta\vec{r}_j^{\,2} \rangle}}$, where $\Delta\vec{r}_i$ is the displacement distance of residue *i* with respect to its average position. When C_{ij} = 1, residue *i* and *j* are fully correlated; when C_{ij} = -1, then residue *i* and *j* are totally anti-correlated. The correlation networks between residues thus were drawn based on ccc information using VMD [189, 190]. Neighboring residues within 10 (|*i*-*j*| <= 10) residues were ignored to remove unnecessary high correlated interactions. In this study, we only calculated the ccc from last 40% of the cMD trajectories. In each trajectory, a 10 ns blocks strategy was used and the average ccc of all blocks was used as the final correlation network. Protein ligand binding pocket detection and size calculation were completed with Fpocket 2 [191] with default parameters.

3.1.2 Umbrella sampling

Umbrella sampling relies on reaction coordinates to obtain free energy profiles. We chose $\Delta RMSD$ as the reaction coordinate in umbrella sampling. $\Delta in\ u$ was defined as the *RMSDs* (*RMSD1* and *RMSD2*) differences, calculated based on the two crystal reference conformations. *RMSD1* and *RMSD2* were updated every 1 ps with the optimal superimposition scheme in Plumed 2.1 [72]. Each instantaneous frame was firstly aligned to the antagonistic reference residues 683-902 backbone atoms, and residues 903-932 αC atoms *RMSD* was computed as *RMSD1*. Similar procedures were carried out for *RMSD2*. The rational of aligning LBD on H1-H10 part but updating *RMSD* based on the C-terminal residues, was maximizing and biasing the flexibility of LBD C-terminal part. A transition pathway between the two crystal references, not considering the ligand was generated with iMOD [192]. The transition pathway was consisted by 30 equally separated intermediate states, forming the starting structures for 30 umbrella sampling windows. The $\Delta RMSD$ differences between the successive windows were 0.1 nm. For each window, harmonic potentials were deposited

The simulations were completed using Gromacs [173] patched with Plumed 2.1 [72]. For each and every of the 30 simulations in umbrella sampling of *apo*-form PR LBD, a 1000 step energy minimization with 1000 kJ/(mol·nm^2) position constraints on the non-hydrogen atoms was performed, following by 1 ns heavy atom position restraint *NPT* equilibrium, where the force constant was gradually decreased from 1000 to 10 kJ/(mol·nm^2). For product runs, the harmonic bias potentials *E* was applied on the $\Delta RMSD$ with force constant k = 10 kJ/(mol·nm^2). The force constant in production runs was chosen such that during simulations PR LBD secondary structure would not be distorted. Other parameters (the force fields and simulation setup) were adopted from cMD runs. The product trajectories were 2 ns since longer simulation did not significantly affect the results. The last 75% of the trajectories were collected for the WHAM analysis [67] to construct the free energy profile alone

the Δ*RMSD* coordination The representative structures of the low energy minima were clustered and extracted, as discussed in the cMD method part.

3.1.3 Metadynamics simulation

Metadynamics [71, 193-195], as an enhanced sampling method, could explore the free energy landscape along predefined collective variables (CVs) [196]. By depositing gaussian-like bias potentials along the selected CVs, it could accelerate sampling process to capture rare events, such as the protein folding. Post-processing (reweighting) of metadynamics could reconstruct the free energy surface (FES) along the biased CVs.

The meta-stable states of *apo*-form PR LBD are the major aim of the metadynamics in this study. We thus performed the well-tempered metadynamics simulation to recover the FES along two CVs. After some trials and errors, the general CVs, such as distances or angles, are not good enough for exploring the complexity of our PR LBD system. For relatively large protein or domains, the optimal CVs should be able to describe the folding motions directly. For protein folding related problem, there are some frequently used CVs, such as the number of contacts (NC), the hydrogen bonding network, the α-helix RMSD [197], anti-β RMSD [197], and the most efficient one, the path CVs [193], which enable one to explore the FES along a possible transition path and are proved to be efficient and accurate [198-200].

In this study, three CVs were defined. CV1, the number of contacts (NC1) between αC atoms of helix 12 (residues 907-922) and helix 3 (residues 712-734), and CV2, the number of contacts (NC2) between αC atom of helix 12 (residue 907-922) and helix 11 (residue 881-898), were chosen for gaussian bias deposition. To obtain smooth transitions for NC1 and NC2 during the simulation, a switch function was employed to evaluate NC1 and NC2: $NC_{CA} = \sum_{i=1}^{N_1}\sum_{j=1}^{N_2}\frac{1-\frac{r_{ij}^8}{r_0^8}}{1-\frac{r_{ij}^{12}}{r_0^{12}}}$, where the r_0 is the distance cutoff close contacts. In this study, r_0= 0.85 nm, while N_1 and N_2 are the total number of the αC atoms in respective helices. In addition, wall-bias potentials were applied on CV3, the α-helix RMSD [197] of helix 12 (residues 907-922). The α-helix *RMSD* is a measure of the α-helical content of selected peptides, with an idealized α-helix as the reference structure. This CV was defined as: $s_\alpha = \sum_{i=1}^{N}\frac{1-\frac{r_i^8}{r_0^8}}{1-\frac{r_i^{12}}{r_0^{12}}}$, where r_0= 0.08 nm and r_i is the *RMSD* between the interesting peptides and a standard α-helix. The wall bias could constraint CV3 within the given range to avoid the deformation of helix 12, thus, facilitating the convergence within computational affordable simulation time scale. The square-well bias potentials were deposited along CV3: $E_{wall} = \sum_i^N k_i((x_i - a_i + o_i)/s_i)_i^e$, where k_i= 300 kJ/(mol·nm^2), the rescaling factor s_i=11, the exponential factor e=2, the offset term o_i=0.

Only if the instantaneous CV3, x_i, is out of the range a_i=10 and a_i =5, the bias potential is calculated and applied to the simulation system. In this equation, i is the number of CVs, here in this metadynamics simulation, i=1.

The crystal antagonistic structure (2OVH, chain A) was used as the initial structure, though the original co-peptides, ligands, and solvents were removed. The system setup and procedures of energy minimization, *NPT* equilibration, and the product run were consistent with the previous cMD simulations. The Gaussian width of the two biasing CVs (NC1 and NC2) was 1.0. This Gaussian width was chosen such that the simulation time and accuracy could be balanced [70]. And the Gaussian potentials height for NC1 and NC2 was set as 0.2 kJ/mol and was updated and deposited every 1 ps (τ_G= 1 ps). The bias factor, defining in the biasing degree of well-tempered metadynamics [69], was 10.

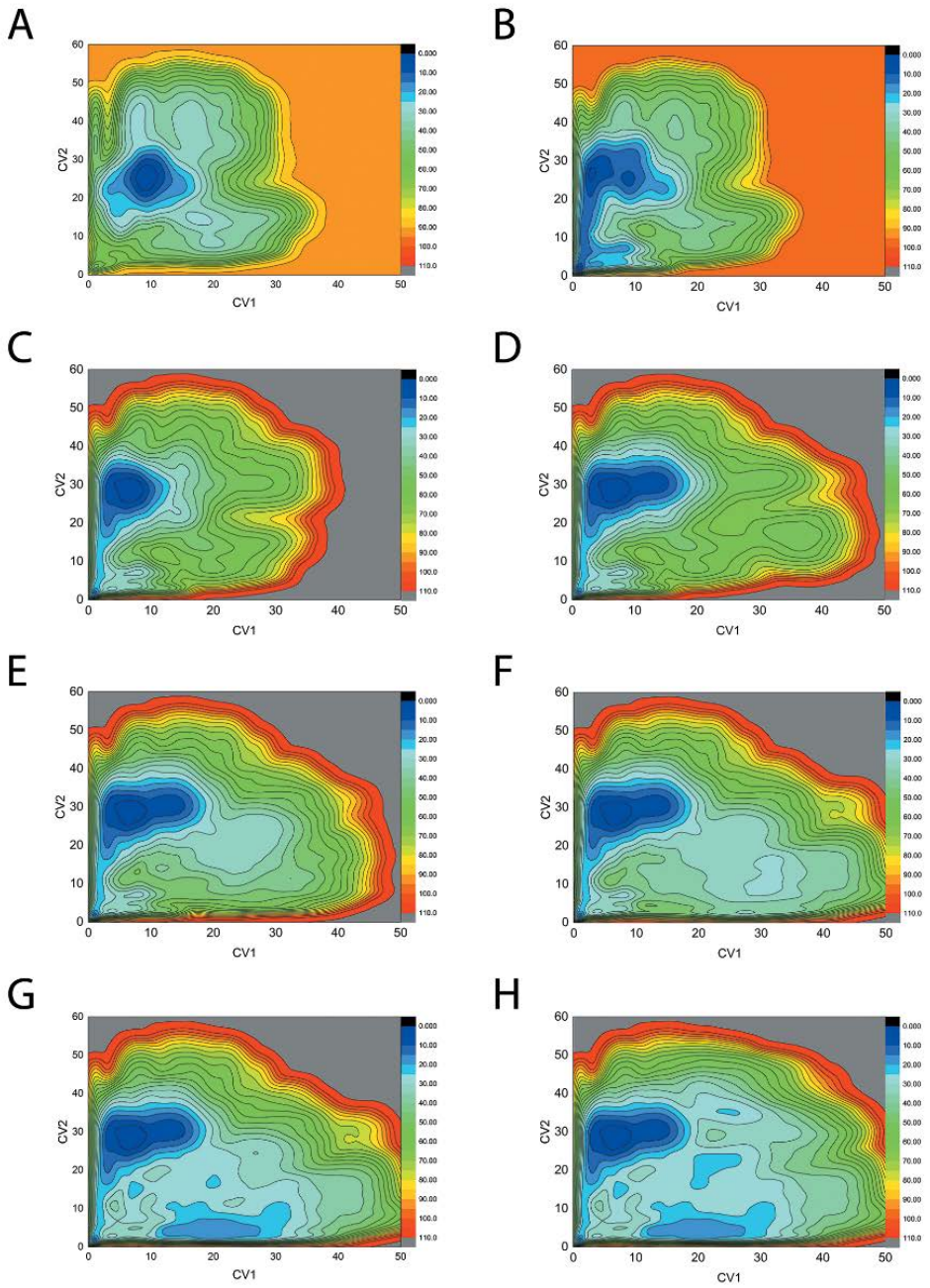


Figure 8. Free energy surfaces (FESs) constructed in different time windows during metadynamics simulation.

The FES maps are constructed based on the Gaussians added from start of the simulation to a specific time point, such as 200 ns (A), 300 ns (B), 400 ns (C), 500 ns (D), 600 ns (E), 700 ns (F), 800 ns (G) and 900 ns (H). The color scales given at the right side of the figures indicate the free energy levels in a unit of kJ/mol, while the iso-lines are drawn every 10 kJ/mol. The FESs for 0~700 ns (F), 0~800 ns (G) and 0~900 ns (H) are quite similar from a globular view, thus they could be an indication for the convergence of the metadynamics simulation.

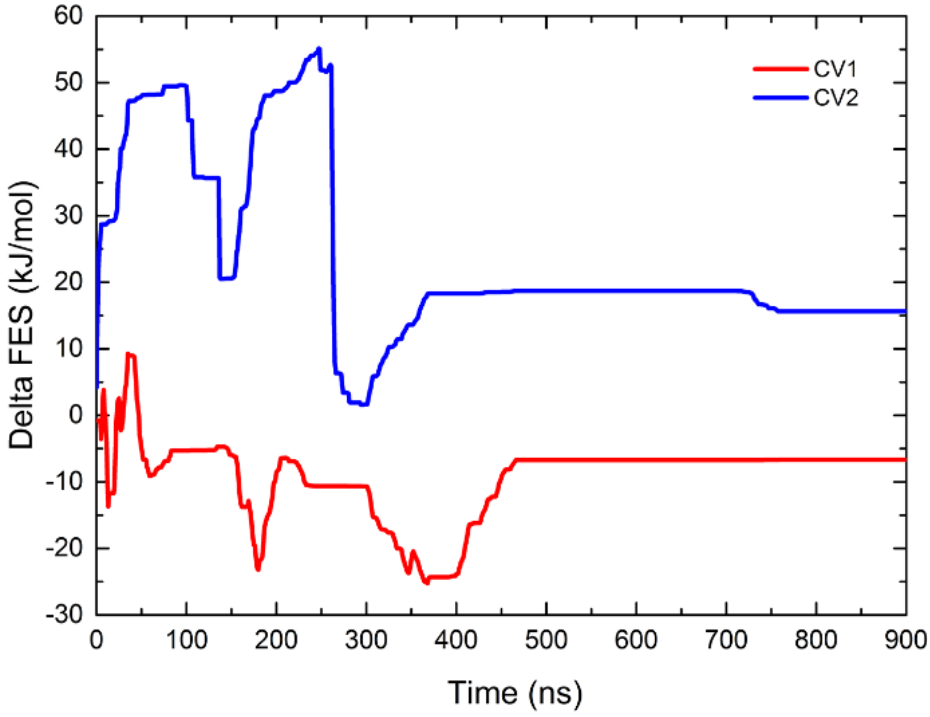


Figure 9. The delta FES changes of 1D CV space along simulation progress.

The delta FES changes of NC1 (CV1) and NC2 (CV2) during the whole simulation are presented by the red line and the blue line, respectively. Towards the large part (200 ns) of the simulation, the delta FES tends to be stabilized within a rather small range (1 kJ/mol), therefore it indicates the convergence of the metadynamics simulation.

The convergence of the simulation was checked. Several FES maps, recovered along NC1 and NC2 during the different periods of the metadynamics simulation, were shown in Figure 8, the FES map does not change much during the last few periods. Meanwhile, the difference of free energies (ΔFES) between two local minima among one of the CVs (NC1 or NC2) was calculated and plotted against simulation progress, according to the tutorial in Plumed 2.1 website (http://plumed.github.io/doc-v2.2/user-doc/html/belfast-6.html). Along the first CV1 (NC1), the one-dimensional free energy difference between two local minima (5 <NC1<9 and 12<NC1<15) was used as ΔFES, while for NC2 (CV2) the ΔFES was computed based on the free energy values of the two local minima (0<NC2 <3 and 27<NC2<30). During the last 200 ns of the metadynamics simulation, the ΔFES curves for both NC1 and NC2 fluctuate within a quite small range (1 kJ/mol) (Figure 9). Therefore, the metadynamics simulation has achieved the convergence criterion.

The representative conformations of the low energy basins in the FES map (during 0-900 ns) were extracted. For each local minimum, the representative conformation of the largest populated cluster was further adopted as the initial structure for another round of 50 ns cMD simulation to confirm the stability of these conformations.

3.2 Results

3.2.1 The *apo*-form antagonistic conformation is intrinsically unstable

The αC atoms RMSDs of the *apo*-2ovh system go up quickly and fluctuate frequently during the 2000 ns cMD simulation, while the RMSDs of *apo*-1a28 simulation remain lower than 0.25 nm (Figure 10A). Clearly, the *apo*-form LBD could only maintain the compact agonistic conformation, but not the more loosely patched helix 12 antagonistic conformation. Consistently, the RMSFs of the apo-form agonistic system *apo*-1a28, are smaller than those of the antagonistic *apo*-2ovh system. Especially, the helix-loop-helix region show quite large RMSF values, indicating a rather flexible instinct of this region and it is the most flexible region in the *apo*-form PR LBD (Figure 10B).

The first two components of αC atoms' coordinates PCA shows that *apo*-1a28 system and the *apo*-2ovh system sample totally diverse space (Figure 10C). The sampling area of *apo*-1a28 simulation is rather narrow and localized, whereas, for apo-2ovh simulation, the sampled conformations are sparser and more separated. Therefore, from the PCA, we could conclude that apo-form antagonistic PR LBD conformations are more diverse. The clustering analysis of the two simulation systems was performed. There is only one major conformation of *apo*-1a28 simulation, whereas more clusters were observed in *apo*-2ovh simulation. The representative structure of *apo*-1a28 simulation is similar (with 0.25 nm RMSD) to crystal agonistic conformation. However, the representative conformations of the first 20 clusters could not be well aligned, especially the helix-loop-helix region (Figure 10D).

Overall, the antagonistic conformation of *apo*-form PR LBD could not be well maintained during cMD simulations, whereas the agonistic state *apo*-form PR LBD is rather stable. The finding here is consistent with the experimental result that, in the antagonistic crystal structure models, the atomic coordinates of the loop 895-908 are void, while the b-factors of the helix 11 and helix 12 are relatively higher than other regions in PR LBD [25].

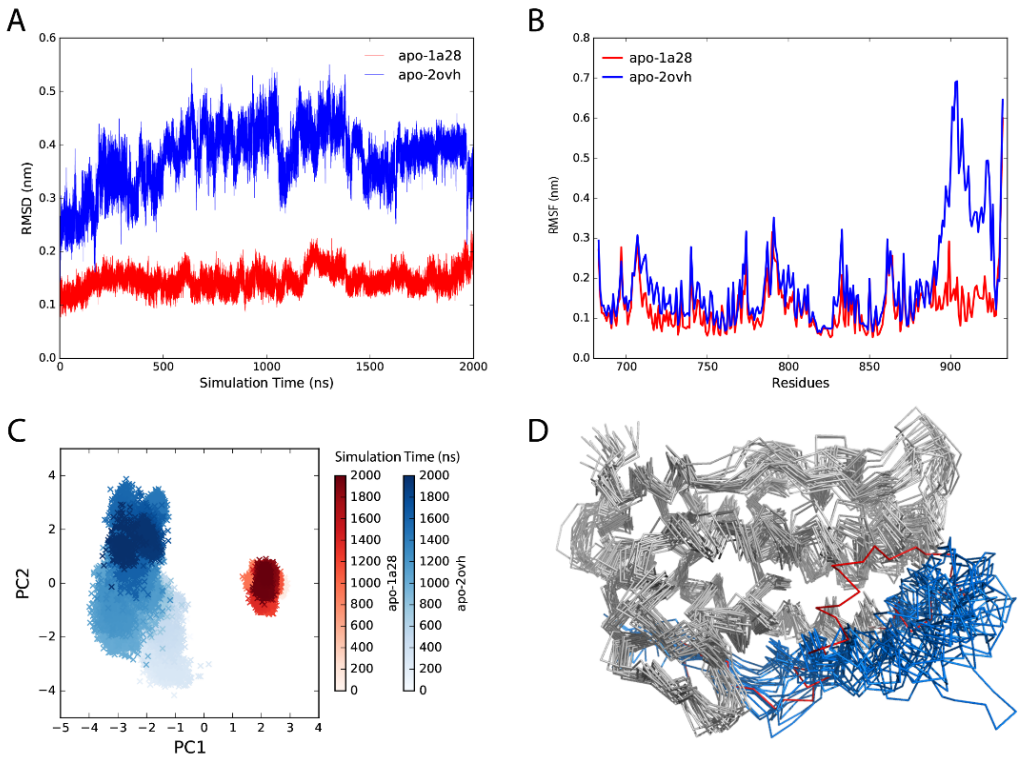


Figure 10. The antagonistic conformation of apo-form PR LBD is unstable.

A, RMSDs of PR LBD simulations from different initial states. B, PR LBD residues RMSFs starting from different initial structures. C, PCA of αC atom coordinates of combined trajectories of *apo*-form PR LBD. D, representative structures of *apo*-form PR LBD simulations. In panel D, the N-terminal bulk end of PR LBD are shown as gray, the red ribbon is the representative structure of H11-H12 part *apo*-1a28 system and the blue ribbons are the representative structure of H11-H12 part *apo*-2ovh system.

3.2.2 Population distribution of LDB between agonistic and antagonistic conformations

The umbrella sampling in this study is designated to explore the free energy profile along the $\Delta RMSD$ coordinate (Figure 11A). The more negative value the $\Delta RMSD$ indicates the PR LBD is more similar to the crystal agonistic conformation. The free energy as a function of the $\Delta RMSD$ recovered in the umbrella sampling indicates two major energy basins (labeled as umL1 and umL2) separated by 6.5 kJ/mol energy barrier, and multiple apo-form intermediate states exist. For the umL1 ($\Delta RMSD$= -1.3 nm), only one major cluster was discovered, and the representative structure resembles the crystal agonistic model (Figure 11B). However, for umL2 ($\Delta RMSD$ = 0.3 nm), multiple states co-exist. The representative structures of the first 10 largest clusters are superimposed with the reference structures. The helix-loop-helix region of umL2 structures are quite diverse (Figure 11C). Helix 12 in umL2 is not totally extended, but rather restricted by the salt-bridge formed between

Arg899 (loop 895-908) and Glu723 of helix 3 in several representative structures (Figure 11D). However, in some representative structures, loop 895-908 points outward from the binding pocket and does not form interactions with bulk LBD (Figure 11E).

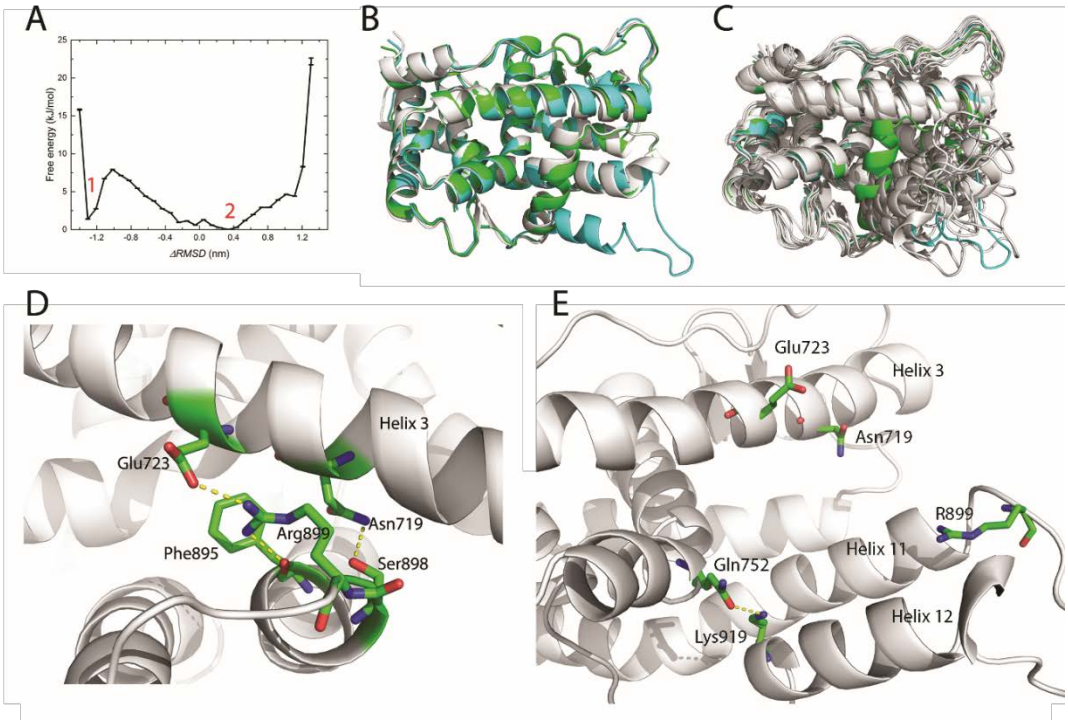


Figure 11. Free energy along the $\Delta RMSD$ coordination and the low free energies representative structures.

A, the free energy along the one-dimensional coordination together with error bars. B, the representative structure (gray color) in the local minimum 1, where $\Delta RMSD$ = -1.3 nm. C, 10 representative intermediate structures in a local minimum 2 (the $\Delta RMSD$ range 0.3 nm to 0.5 nm). In panel B and C, the PR LBD crystal agonistic conformation (green) and the crystal antagonist conformation (cyan) are also superimposed with the representative structures.

3.2.3 Free energy surface and meta-stable states of apo-form PR LBD

There may not be a straightforward transition from the antagonistic to agonistic conformation for *apo*-form PR LBD, thus the one-dimensional CV $\Delta RMSD$ may be too simplistic to form a complete dynamic picture of PR LBD. Therefore, two CVs have been chosen to enhance the sampling of the *apo*-form PR LBD with the aid of metadynamics.

The FES recovered along the two NC CVs (NC1 and NC2) in metadynamics simulation is shown in Figure 12. Eight local minima have been identified and labeled as 1 to 8, where the local minimum 1 (L1) is the dominant basin in this FES map. The crystal antagonistic conformation (NC1=0.99, NC2=34.08) is not located near any low energy basins, while the position of crystal agonistic conformation (NC1=16.69, NC2=31.24) in the FES map almost resides in L2. This further suggests

that, as revealed in cMD simulations, the antagonistic conformation of *apo*-form PR LBD is indeed unstable, and crystal agonistic *apo*-form PR LBD is relatively stable but is not the dominate meta-stable structure. The representative structures of these 8 local minima are generated using clustering analysis. Except for the helix-loop-helix region, the other parts of PR LBD could be well-aligned with agonistic and antagonistic models.

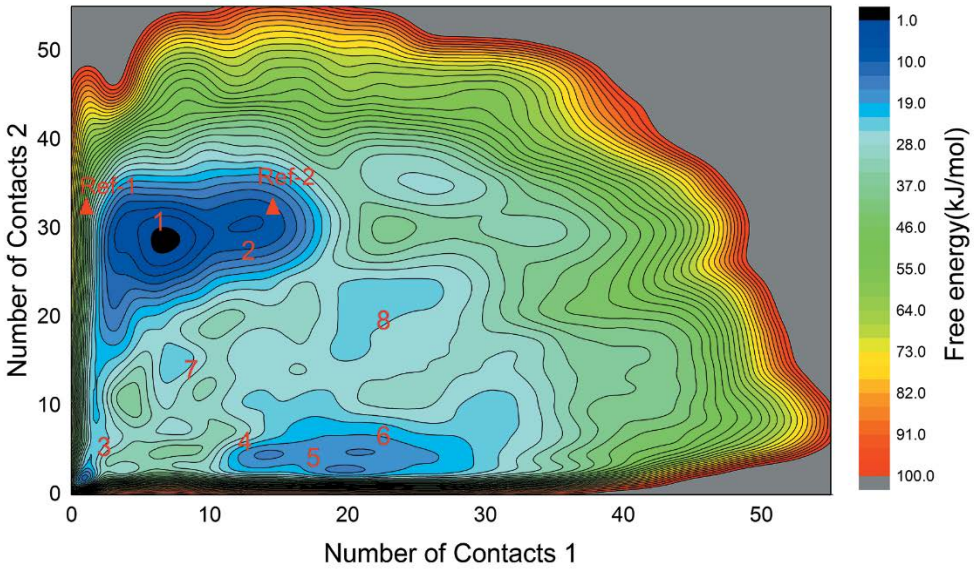


Figure 12. The free energy surface (FES) map constructed by metadynamics simulations.

8 major basins of local minima are labeled together with the positions of two crystal structure models (Ref-1: 2OVH; Ref-2: 1A28) indicated by red triangle marks. The color key at the right side of the figure indicates the free energy scales in a unit of kJ/mol. The iso-surface lines are given every 10 kJ/mol.

Table 3. Interaction pairs involving residues in the helix-loop-helix segment in metadynamics low energy structures.

Structures	Interaction pairs around helix-loop-helix region
mL1	Asn719-Ser898, Glu723-Lys919, Glu723-Arg899
mL2	Asn719-Ser898, Asn719-Arg899, Glu723-Arg899, Glu723-Lys919, Trp755-Pro918*
mL3	Asn719-Ser898, Asn719-Arg899, Glu723-Arg899
mL4	Glu723-Ser902, Glu723-Gln916, Glu723-Lys919
mL5	Glu723-Gln916, Asn719-Ala900*
mL6	Glu723-Gln916, Glu723-Lys919, Asn719-Ala900*, Arg899-Ala900*
mL7	Glu723-Gln916, Glu723-Lys919, Trp755-Lys919*
mL8	Glu723-Gln916, Glu723-Lys919, Arg899-Glu904, Ser898*-Thr716

* indicates the interacting atoms are on the backbone, whereas all other interaction pairs are on side-chains.

For the representative structures, in the most flexible helix-loop-helix region, some common interaction patterns have been discovered (Table 3). For example, Glu723 (helix 3) forms hydrogen

bonds with Lys919 and Gln916 to stabilize helix 12 C terminal end, thus helix 12 adopts a quite flexible N terminus conformation, with only one exception (structure mL3). Besides, the salt-bridges or hydrogen bonds formed by Arg899-Glu723 and Arg899-Gln719 are also observed inmL1, mL2, and mL3 structures. These interactions were also observed in cMD simulations and umbrella sampling simulations (Figure 11).

Taken mL1 conformation as an example, the overall forming of PR LBD is quite distinct from the agonistic or antagonistic models (Figure 13A and 13B). The N-terminus of helix 12 is pointed outwards leading to a totally exposed ligand binding pocket. The C-terminus of helix 12 is stabilized by a hydrogen bond formed between Glu723 and Lys919, while residues in of helix 12 N terminus are free of contacts, therefore enabling a flexible of helix 12 N-terminus. And helix 11 shifts more towards helix 3, where the cross angle between helix 11 and helix 3 is around 75˚, about 20˚ more than the those of crystal agonistic and antagonistic conformations. The interactions (Arg899-Glu723, Arg899-Asn719, and Ser898-Asn719) would attract helix 11 towards the middle of helix 3, thus they may contribute to the larger cross angle in this mL1 structure. For loop 895-908, it is somehow disordered. The majority residues in loop 895-908 do not form any contacts with other residues in LBD, except a possible hydrogen bond between Ser902 (in helix 12) and Gln916.

Using metadynamics simulation, we have sampled a low energy model close to the agonistic conformation. mL2.3 structure (the representative conformation of the 3rd most populated cluster at L2) (Figure 13C and 13D), resembles the crystal agonistic structure, with an αC RMSD around 0.3 nm. For this mL2.3 conformation, the hydrogen bonds or salt-bridges between Asn719-Arg899, Glu723-Arg899, Trp755-Ala914, Glu911-Glu723, and Glu911-Arg899, render helix 12 in a configuration similar to the crystal “closed” agonistic state.

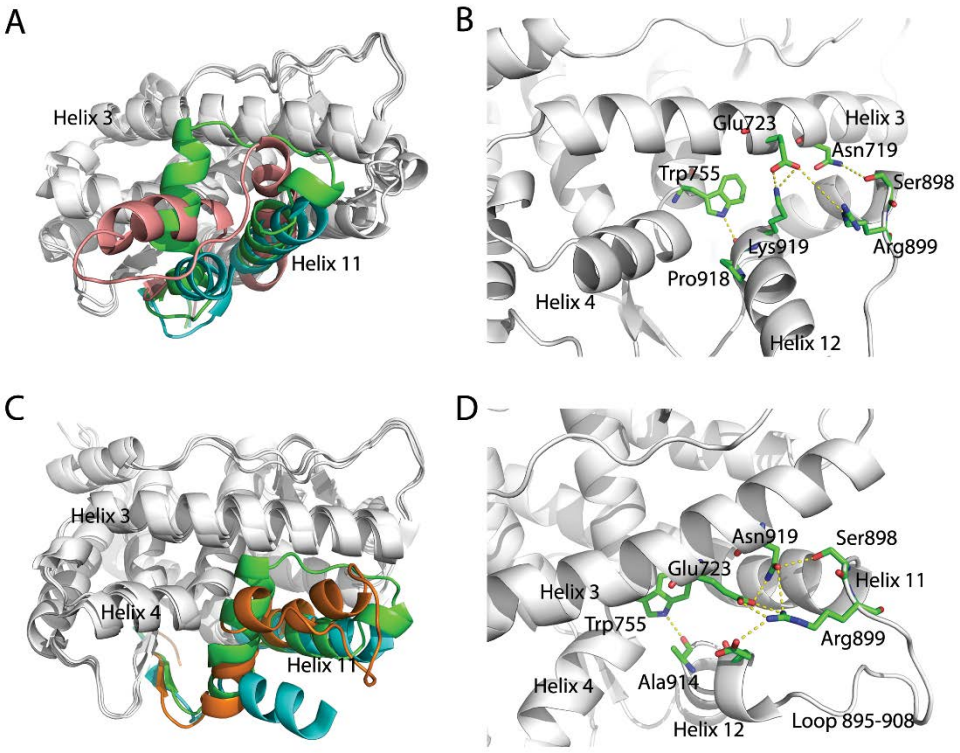


Figure 13. Representative structures of local minima 1 and 2 sampled in the metadynamics simulation.

A, the representative structure mL1 (pink) superimposed with crystal agonistic and antagonistic models. B, the detailed view of interaction networks around the helix 11, loop 895-908 and helix 12. C, the representative structure mL2.3 superimposed with crystal agonistic and antagonistic models. D, the detail interaction network between residues in helix-loop-helix segment. The crystal agonistic (PDB ID 1A28) and antagonistic (PDB ID 2OVH) conformations are colored as green and cyan in the helix-loop-helix region. The yellow dashed lines indicate close contacts between residue atoms.

As far as we know, no experimental *apo*-form PR LBD model is available currently to compare with sampled models directly in this study. To verify the stability of our models, each representative structure of the local minima is recruited as the initial structure for a cMD simulation. Most of the initial structures (except mL3) are relatively stable during the 50 ns cMD simulations (See section 3.3). These stable structures thus may represent the different *apo*-form states and could be adopted for future docking studies as ensemble receptor models to search for potential ligands [98].

3.3 Discussion

3.3.1 Helix 12 is not likely to adopt a totally extended conformation in PR LBD

In both *apo*-form ERα LBD [201] and RXRα LBD [35] X-ray structure models, helix 12 adopts a totally extended conformation dispatching the core LBD, thus exposes the ligand binding pocket widely open (Figure 14A). Therefore, upon ligand binding, helix 12 undergoes large conformational

changes to form functional or the agonistic conformations and exposed the respectively co-activator or co-repressor binding surface. However, several studies question the existence of this extended-helix 12 model of NR LBD [36, 202, 203] in a biological context.

The fast transition from the extended conformation to an intermediate state in *apo*-form ERα LBD has been observed in a biophysics study [204]. The quick folding behavior of helix 12 further indicates the extremely flexible instinct of this helix. Batista et al [36] explored the dynamics of the *holo*-form PPARγ LBD using time-resolved fluorescence anisotropy decays by attaching a fluorescent probe at the extreme C-terminal tail. They also modeled the ERα LBD and RXR LBD using cMD simulations. In their work, only rather local motions of helix 12 in different NR LBDs had been detected. From our enhanced sampling simulations, it turns out that only relatively local motions of helix 12 have been sampled. Besides, the PR LBD with helix 12 totally extended has never sampled in all the simulations, thus it may further indicate that, unlike ERα LBD, the extended PR LBD with an extended helix 12 is not a high population conformation, which is out of reach for simulations with limited time-scale.

In ERα LBD, the shorter helix 12 (about 8 residues) may adopt the extended conformation, partially due to the loss of hydrophobic interactions between the extreme C-terminus with the LBD. Other NR LBDs, such as PR LBD, AR LBD, and GR LBD, have longer helix 12 and an extreme C-terminal tail (12 residues in PR LBD)., For these LBDs, the extended conformation (helix 12 totally dispatched), has not been observed experimentally. In PR LBD, the extreme C terminal tail forms a shortβ-sheet, similar to that in GR LBD and AR LBD. Besides, in PR LBD agonistic and antagonistic structures, the π-cation interaction (Arg845 and Phe930) was observed. and this interaction is well maintained in different simulations. What's more, it turns out that these two residues are highly conserved NR LBDs, such as AR LBD and GR LBD [40] and their counterparts could also form π-cation interactions (Figure 14B). According to a computational study, the energy contribution of one π-cation residue pair contacts (a phenylalanine and an arginine) is around -2.9 ± 1.4 kcal/mol [205], therefore the Phe943-Arg845 π-cation interaction in PR LBD further stabilizes the extreme C terminal tail, which restrains the helix 12 mobility. In summary, we propose that for *apo*-form PR LBD, helix 12 could not adopt a fully extended conformation as the helix 12 of *apo*-form ERα LBD.

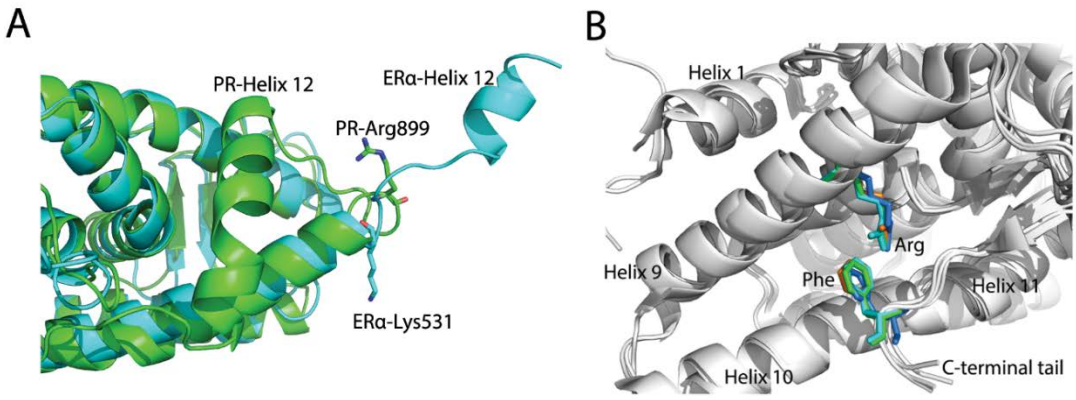


Figure 14. Overlay of *apo*-form ERα LBD with PR LBD and the π-cation interaction in NR LBDs.

A, ERα LBD (cyan, PDB ID 1A52) has a relatively shorter helix 12 and doesn't have an extended C terminal tail, whereas PR LBD (green, PDB ID 1A28) has a longer helix 12 and a C terminal tail. B, conserved arginine (in helix 9) and phenylalanine (in C terminal tail) are shown as sticks, whereas other parts are shown as cartoon. Green, PR LBD agonistic conformation (PDB ID 1A28); cyan, PR LBD antagonistic conformation (PDB ID 2OVH); marine, AR LBD agonistic conformation (PDB ID 4OEA); orange, GR LBD agonistic conformation (PDB ID 3K23).

3.2.2 Simulating *apo*-form PR LBD targeting novel inhibitor discovery

From our cMD simulations, water molecules never enter the binding pocket in *apo*-1a28 and *apo*-2ovh simulations. These observations, together with fact that the binding pocket forming residues are mainly non-polar residues, indicate a rather hydrophobic binding pocket, as previously reported [2, 25, 32, 40, 206, 207], which explains why known active molecules are majorly steroid derivatives or other hydrophobic molecules.

The cMD simulations and enhanced sampling were all initiated from-the ligand-bound crystal models. This is a common practice in computational simulation studies, such as the MD simulations of HIV protease, where new allosteric pockets were detected [208]. In another research by Miao et al [209], several allosteric druggable sites were identified from *apo*-form M2 Muscarinic Receptor simulations. The non-classic binding sites are useful for future novel ligand designing. Similarly, four major druggable sites were formed during cMD simulations (Figure 15A). With pharmacophore searching (with online server) around the PR LBD surface, the probes were iteratively docking to PR LBD to detect the location where multiple probes could bind. The largest pocket is the common steroid binding site and has a pocket-size ranging from 300 to 1400 $Å^3$ (Figure 15B), consisting with sizes of crystal structures [2], while the other three sites have small pockets. During the cMD simulations without a ligand in this major site, side-chains around the pocket are freely rotatable and therefore the smaller binding pocket forms. The binding pocket could fit a variety of molecules which may be dramatically diverse from the known architecture. Previous screening studies [210, 211] against PR LBD were solely based on the crystal ligand-bound structures. Making use of the newly discovered druggable pockets would possibly lead us to the discovery of a totally different category of ligands.

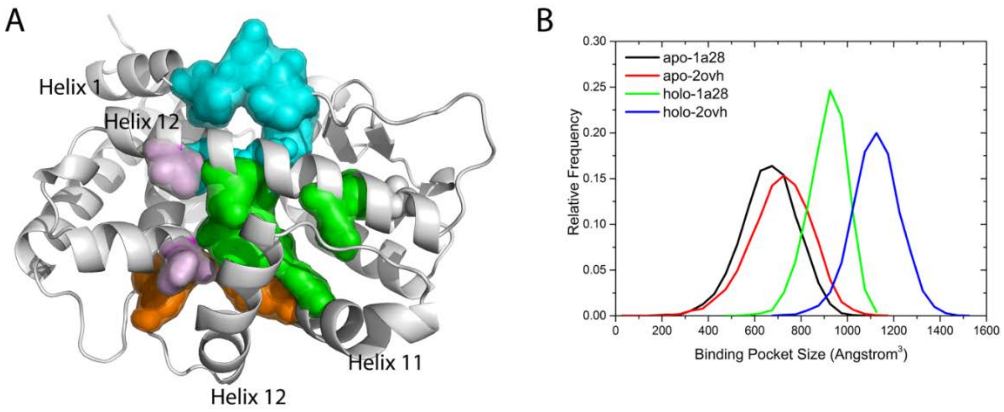


Figure 15. Druggable sites in PR LBD detected and binding pocket size.

A, the four druggable sites detected based on the representative conformations from normal MD, umbrella sampling, and metadynamics and plotted on crystal agonistic conformation; their outside residues are shown in different color, whereas the major binding pocket is surrounded green residues. And the other three druggable sites are formed by cyan, orange and magenta residues, respectively. B binding pocket size relative frequency distributions in normal MD simulations; black, red, green and blue lines represent the relative frequency of the binding pocket size in *apo*-1a28, *apo*-2ovh, *holo*-1a28, and *holo*-2ovh normal MD simulations respectively.

3.4 Summary

Through large-scale MD simulations, we uncover the flexible instinct of *apo*-form PR LBD in solvent, and the possible ligand-free conformations have been sampled. The "closed" conformation, observed in crystal ligand-bound PR LBD structures, have been re-visited, indicating the efficiency of our advanced sampling methods, and also one of the major stable structures. While the crystal "open" state with helix 12 exposed in solvent, however, are rather flexible and is not one of the meta-states of the ligand-free PR LBD. In addition, the flexibility of *apo*-form PR LBD is majorly governed by the orientation of helix 12, helix 11 and the loop between these two helices. The formation of several salt bridges (such as Arg899-Glu723) facilitate the motions of helix 12. We also found that the ligand binding pocket in PR LBD is also high resilient and could accommodate ligands with different sizes. Based on the simulation trajectories, we discovered two alternative binding pockets, which could be future PR LBD drug screening sites.

Chapter 4. Conformation determination of PR LBD

The conformational changes are required for recruiting activators or repressors to PR homodimer or PR/ER heterodimer. The most typical conformational changes of PR LBD are the orientation of the AF2, the helix 12. Need to mention, that the NTD domain experiences large conformational changes upon a ligand binding, or co-activator and co-repressor binding. But since the defined structures of PR NTD are not available, it is very difficult to explore the ligand affected conformational changes of the NTD, therefore we still focus on the LBD related conformational changes.

It is hypothesized that ligand-induced conformational change is a general mechanism for biological functions in NR. Meanwhile, the ligand-induced LBD conformational population shift could also be a possible strategy in PR induced gene expressions. However, it is quite difficult to measure the induced fitting or population shift directly. Here, we applied cMD simulations to study the possible mechanism of the conformational changes of PR LBD, and possibly the ligand-induced transition pathways.

4.1 Methods

4.1.1 Simulation protocols

Molecular dynamics simulations are useful tools to explore molecular behavior and mechanism in atomic level. In this study, we harness the power of GPU cards to perform micro-second range cMD simulations to understand the determinants of the PR LBD conformations. Several systems were built to consider not only the existence of ligands but also co-peptides, as listed in the following table 4.

Table 4. Simulation systems and setup

Name	LBD initial	Ligand	Co-peptides	# of Repeats	Simulation Time (ns)
S1	1A28, chain A	P4	None	1	2000
S2	1A28, chain A	None	None	1	2000
S3	2OVH, chain A	asoprisnil	2OVH, chain B	1	2000
S4	2OVH, chain A	asoprisnil	None	3	2000
S5	2OVH, chain A	None	None	1	2000
S6	2OVH, chain A	None	2OVH, chain B	1	2000
S7	2OVH, chain A	P4	None	3	1000

Initial structures of the LBD, ligands, and co-peptides were downloaded from the RCSB protein data bank (PDB) (See table 4). Original water molecules in the initial structures were removed, missing atoms and residues of LBD were modeled in SWISS-MODEL online server (https://swissmodel.expasy.org/) [212]. For simulation system S7 and S8, the initial LBD-ligand complexes were modeled using GOLD docking package [213]. The docking calculations of P4 and asoprisnil were based on the 2OVH chain A, without co-peptide binding, using the default parameters

for GOLD. The amber99sb-ildn force field [214] was used for LDB and co-peptides. As for ligands, AM1-bcc charges together with amber gaff force field were used [177]. The LBD or the LBD in complex with ligand or co-peptides were solvated in abundant TIP3P [215] water molecules as well as 0.15 M NaCl salt condition. All simulations were completed with Gromacs 5.1.2 package [216] together with GPU acceleration on national super computation center (NSCC), Singapore.

Each simulation system was first energy minimized using the deepest descendant algorithm, and then equilibrated with protein (as well as ligands if exists) heavy atom position constrained using a 1000 kJ/mol·nm^2 force constant under *NPT* ensemble at 300 K and 1 bar for 10 ns. Production runs were performed under *NVT* ensemble at 300 K. The time step for all simulations was 2 fs, while the coordinate data were stored every 2 ps. Thermostats and pressure coupling were achieved using velocity rescaling [179] and Berendsen pressure coupling [217] method respectively. Bonds between heavy (non-hydrogen) atoms were restricted with LINCS algorithm [182], while bonds between hydrogen atoms and heavy atoms were fixed according to SHAKE algorithm [181]. Particle mesh Edwald (PME) scheme [180] algorithm was adopted for the long-range electrostatic potential calculation. A 1.2 nm distance cutoff for both long-range electrostatic and van der Waals interactions was used.

4.1.2 Analysis methods

Δ*RMSD*, used in one of our previous studies [218], is defined by the difference between two RMSD values, RMSD1 and RMSD2, which are the real-time RMSD calculated using antagonistic and agonistic LBD conformations as references. The PCA analysis was performed using python sklearn package and numpy package [219]. The αC atom distance matrices were used as datasets for PCA analysis. For a simulation trajectory, the conformations of every 100 ps were collected, and the distances between all αC atoms were calculated. The dataset thus was fed into the sklearn decomposition module, and the eigenvalues and transformed dataset were plotted. Based on the PCA analysis, the eigenvectors per αC atom were determined and adopted for the essential dynamics of each residue and visualized using VMD 1.9 [190]. Clustering analysis was performed using Gromacs g_cluster tool with the gromos algorithm and a 0.2 nm RMSD cutoff. Quasi-Harmonic conformational entropy [220] calculations of PR LBD were performed using Gromacs g_covar and g_aneig tools. The *Home Sapiens* PR, AR, αER, MR and GR sequences were downloaded from NCBI protein sequence database, and the alignment of these sequences was performed using T-Coffee alignment server (http://tcoffee.crg.cat/apps/tcoffee/index.html) [221], and the alignment figure was generated using BoxShade server (https://embnet.vital-it.ch/software/BOX_form.html).

The first hydration shell according to S. Sinha et al [222] is around 5 Å, therefore we used this distance as the cutoff for calculating the coordination number between the protein (with ligand where necessary) and solvent water molecules. The water residence time was calculated based on the

distance between a molecule and a protein residue within 5 Å during a continuous time, using home-made python package (http://www.github.com/zhenglz/dockingML). Bridging water molecule was defined if a water molecule oxygen atom is within 0.35 nm distance range of both the Trp755 sidechain nitrogen atom and the Val912 backbone nitrogen atom. The residence time of the bridging water molecule is calculated if it is continuously remaining close to the two residues for a period.

The electrostatic potential energies are calculated based on the following formula:

$$E = \frac{q_1 q_2}{4\pi\varepsilon d^2} \qquad (43),$$

where ε = 80 in this equation. And the electrostatic potential surface of the LBD were generated using Chimera [223] using default parameters.

4.2 Results

4.2.1 P4 binding decreases agonistic PR LBD conformation flexibility

The deposited crystal structures of PR LBD in RCSB PDB all adopt a typical 11 helical folding scheme as other LBD in NR family [2]. All the structures are *holo*-form, either with a ligand (or together with a co-peptide), no *apo*-form PR LBD conformations have been determined using X-ray or NMR methods. Our previous theoretical study [218] has proved that, for *apo*-form PR LBD, there exist multiple intermediate states, which includes the agonistic conformation (the "closed" state) where helix 12 is hooked by helix 3 [40, 41] by electrostatic interactions, forming an isolated binding pocket.

Our long-scale MD simulation (simulation systems S1 and S2, see table 4) indicate that the initial "closed" state (agonistic conformation) could be well maintained with or without a ligand bound throughout the whole simulation (Figure 16A), though the residues around the binding pocket show large fluctuations (Figure 16B) when no ligand bound. From the RMSF plot, residues of three regions are more flexible in *apo*-form (simulation S1) than P4-bound LBD (simulation S2). These three regions are located around helix 5, helix 6, helix 7, helix 11, helix 12 and the loops around these helices, which forms the binding pocket of P4. The configurational entropy of the binding pocket residues in these three regions are much higher in *apo*-form LBD (1529.86 $Jmol^{-1}K^{-1}$) than P4-bound LBD (1230.43 $Jmol^{-1}K^{-1}$), thus it also indicates that residues around binding pocket are rather mobile in *apo*-form LBD than P4-bound *holo*-form LBD.

From *apo*-form (simulation S1) and P4-bound *holo*-form (simulation system S2) simulations, overall dynamics of LBD are also diverse with or without P4 binding. αC atoms distance matrix based PCA analysis indicates that the configurational space of *apo*-form LBD and *holo*-form LBD simulations could be well separated (Figure 16C) since the first principle component (PC1, counts for 44.0%

dynamics) of system S1 and system S2 are located in different area. And the data points along PC2 for *apo*-form LBD is sparser than P4-bound *holo*-form LBD, also suggesting a more flexible instinct of *apo*-form LBD. With or without P4, LBD samples divergent essential dynamics spaces (Figure 16D). H1-H3 loop (the loop between helix 1 and helix 3) in *apo*-form LBD shifts towards helix 3 and C-terminal of helix 11, whereas it binds tightly with region 1 and region 2 in P4-bound LBD majorly through electrostatic interactions and hydrogen bonds (Figure 16E and 16F).

In conclusion, the loss of P4 binding in agonistic LBD would increase residues flexibilities, especially residues around the binding pocket. Meanwhile, loop H1-H3 has different conformational orientations in *apo*-form and P4 bound simulations (simulation S1 and S2, respectively), though the biological significance of the dynamics of this loop remains unknown.

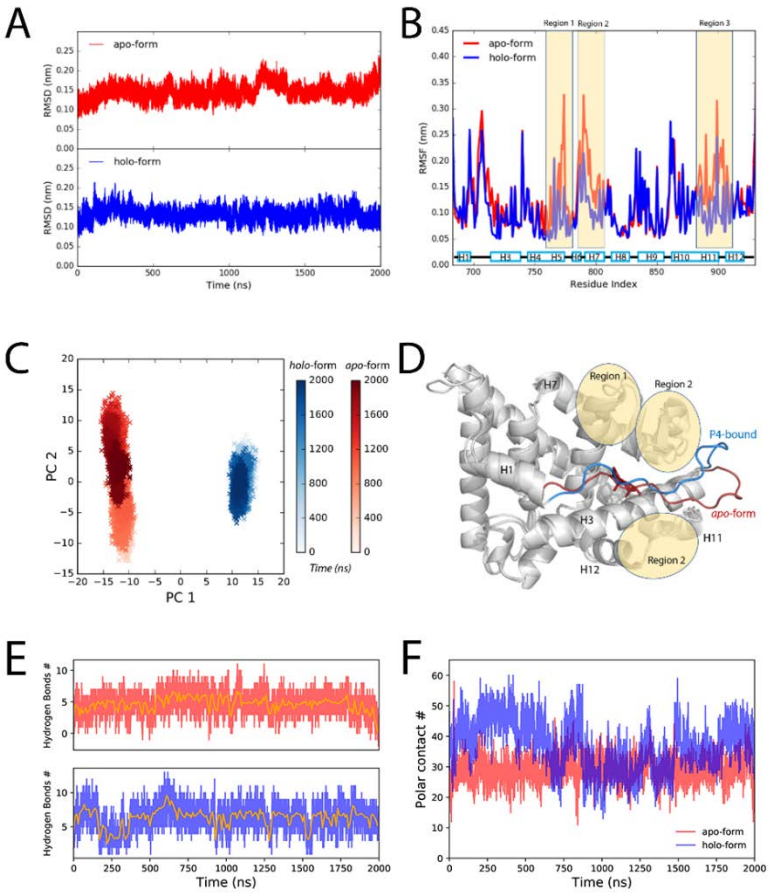


Figure 16. P4 binding stabilizes agonistic PR LBD conformation.

A, the αC atoms RMSDs of *apo*-form and P4-bound LBD. The large rise in the late stage of the RMSD of apo-form LBD indicates the system is more flexible. B, the residue RMSFs of *apo*-form and P4-bound LBD. From the RMSFs of residues, 3 regions are more flexible in *apo*-form LBD than P4-bound LBD. C, PCA projections of the contact matrix of αC atoms of *apo*-form and P4-bound LBD. The sparser distribution instinct of *apo*-form LBD suggests that it is more flexible. D, structure alignment of the representative structures from *apo*-form and P4-bound LBD simulations. The most

flexible regions in the LBD have been highlighted with yellow circles and blue/red colors. E, the hydrogen bond numbers between regions 1,2 and H1-H3 loop as a function of simulation time for *apo*-form (red color) and P4-bound (blue color) systems. F, the number of polar contacts between regions 1,2 and H1-H3 loop as a function of simulation time for *apo*-form (red color) and P4-bound (blue color) systems. The *apo*-form and P4-bound LBD simulations are systems S1 and S2 respectively. The helices are labeled as "H", for example, H12 in panel D (and all other figures where necessary) represents helix 12.

4.2.2 P4 binding induced PR LBD conformational change is a multiple-stage process

It has been reported that induced fit is quite a general mechanism in NR LBDs [2], however, it is still unclear whether agonists would induce the folding of PR LBD. Our simulations indicate that the binding of P4 would result in the orientation rearrangement of helix 12, as well as the H11-H12 loop to form the agonistic-like conformation.

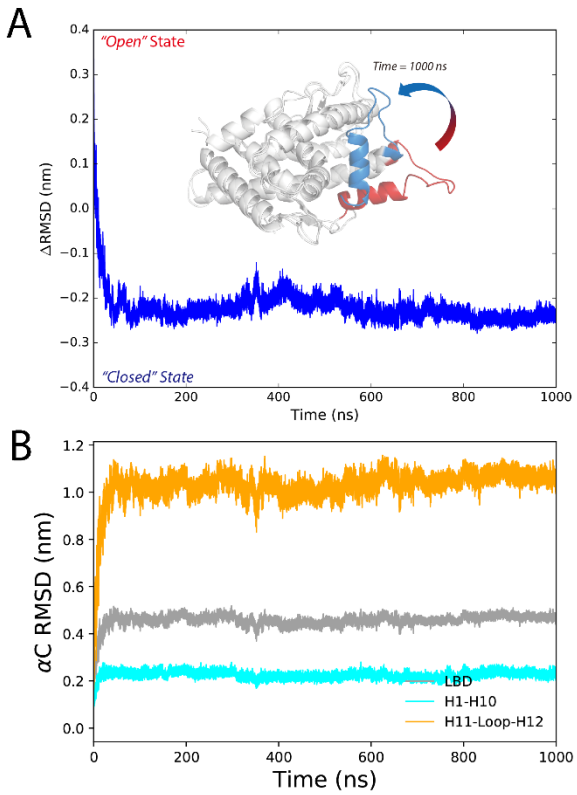


Figure 17. Conformational changes of P4 bound PR LBD.

A, the $\Delta RMSD$ along the trajectory (simulation system S7 in table 4, repeat #1). B, the αC atom RMSDs of the PR LBD (gray), the helices 1-10 (cyan) in LBD, and the helix11-loop-helix 12 region (orange) in LBD (simulation system S7 in table 4, repeat #1).

We simulated the PR LBD initiated from the "open" state (antagonistic conformation), where helix 12 is dispatched from core LBD, and with P4 docked into the binding pocket (simulation system S7).

From the $\Delta RMSD$ changes along the simulation S7 repeat 1, the P4-docked antagonistic PR LBD quickly drops below -0.1 nm within the first 50 ns (Figure 17A). The decreasing of the $\Delta RMSD$ suggests that instantaneous LBD conformation has lower *RMSD* against the agonistic reference structure, but high *RMSD* against the antagonistic reference. The decreasing of this $\Delta RMSD$ is majorly caused by the large conformational dynamics of helix 11, helix 12 and H11-H12 loop (Figure 17B). Detail examination of the trajectories reveals that helix 12 of PR LBD rapidly folds back and facilitates the formation of the "closed" conformation, which resembles the crystal antagonistic conformation with an αC atoms RMSD within around 0.2 nm upon the end of the 1000 ns simulation (Figure 17A). And the key interactions between P4 and LBD residues are also recovered and stabilized from MD simulations (Figure 18).

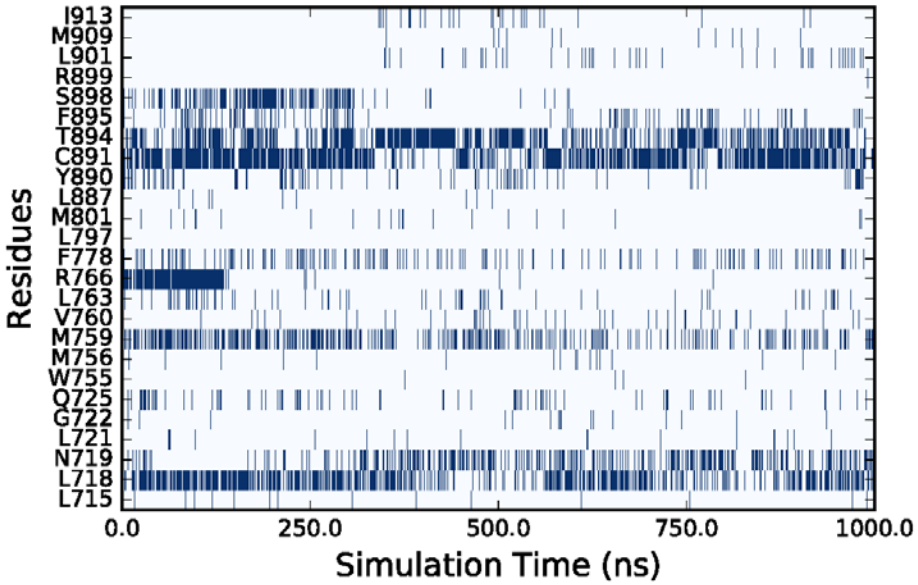


Figure 18. Close contacts between LBD residues and P4 along the simulation system S7, repeat #1.

The contacts between P4 and residues in LBD were calculated along the simulation, if the contacts between a residue and P4 is higher than 1, a dark blue bin will shown in the figure, while a white bar indicates that at a specific time point, a residue does not form any contact with P4.

Previous researches have proved that Met909 plays a vital role in stabilizing the agonistic state of PR LBD [41], and Arg899/Glu723 hydrogen bonds may affect the conformational dynamics of *apo*-form PR LBD [218]. We believe that these interactions would be good features to depict the conformational transition of P4-bound PR LBD.

Initially, Arg899 is quite flexible and fully exposed in the solvent, and helix remains the antagonistic state orientation (Figure 19, time = 0 ns). Several clear stages are recorded for these interactions during the folding of helix 12. During the early folding process (from 0 ns to ~30 ns), Arg899 quickly approaches Asn719 and it forms transient hydrogen bonds with Asn719, and then it forms hydrogen

bonds with Glu723. Then, Val912, Ile913, Trp755 and P4 gradually form a hydrophobic core, thus restricting the flexibility of helix 12. As for the second stage (from 30~50 ns), after the formation of electrostatic interactions and hydrogen bonds between Glu723 and Arg899, Met909 sidechain starts inserting into the hydrophobic core and forms a hydrogen bond with Glu723 sidechain, resulting in the dropping of distances between Met909 and Glu723 within the first ~30 ns and remaining under 0.5 nm throughout the whole simulation, patching N-terminal end helix 12 tightly against helix 3 (Figure 19, time = 30 ns). After some oscillations of LBD residues, Met908 sidechain also inserts into the hydrophobic core. For the third stage, (after ~50 ns), Arg899 moves far apart from Glu723 and thus frees the restrictions around helix 11 C-terminal end (Figure 19, time = 400 ns). In this third stage, the re-orientation of Arg899 dismisses the space clash around helix 12 N-terminal end and enables the conformational adjustment of H11-H12 loop in later simulation, and helix 11 would be relatively more stable with the hydrogen bonds forming between Arg899 and Ser712 (Figure 19, conformation at time = 1000 ns). Therefore, MD simulations reveal the multi-stage transition of PR LBD from “open” state to “closed” state. Besides, the distances, as well as the hydrogen bonds forming, between Glu723-Met909 and Glu723-Arg899, as well as other residue pairs such as Trp755-Val912 and Trp755 and Ile913, are highly correlated with the helix 12 folding events.

P4-bound PR LBD could form stable agonistic “close” state conformation. The transition from antagonistic to agonistic conformation may involve three clear steps: 1) fast hydrophobic core forming and helix 11 localization, 2) fast patching of helix 12 and 3) slow adjustment of H11-H12 loop, highly coupling with key hydrogen bonds and distances of Met909/Glu723 and Arg899/Glu723.

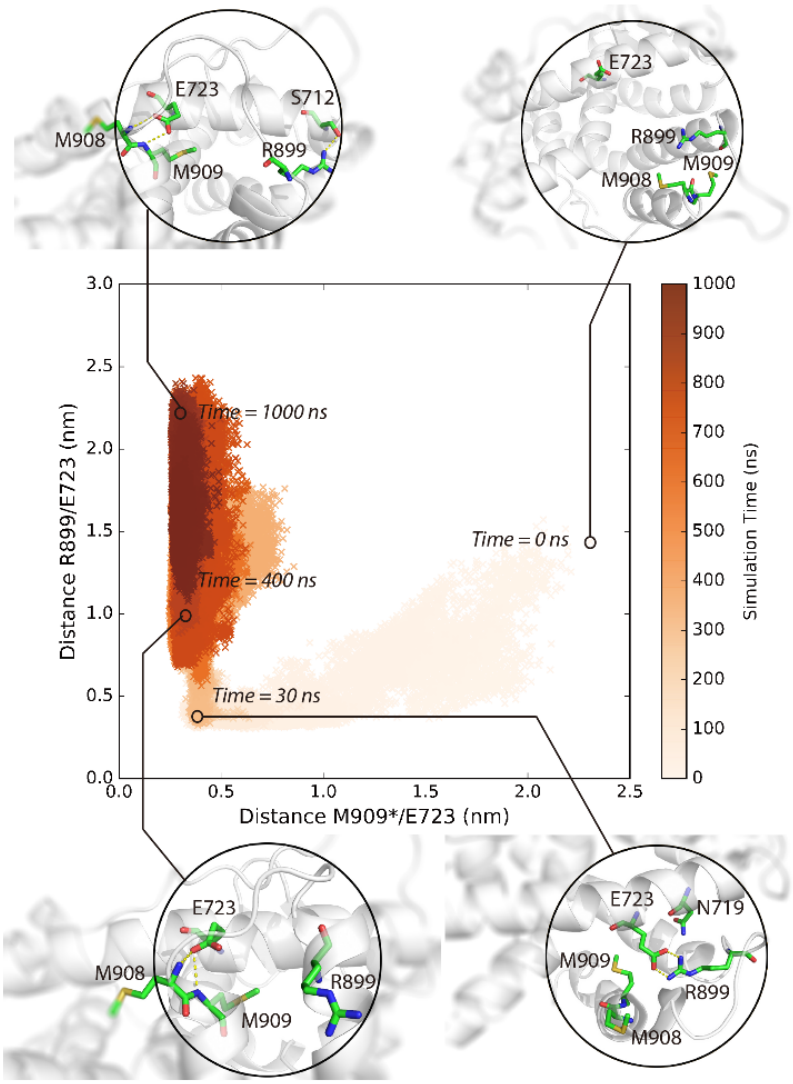


Figure 19. The conformational adaptation of antagonistic PR LBD with agonist P4 binding.

P4 was firstly docked into antagonistic PR LBD binding pocket. The MD simulations (simulation system S7 in table 4) were then performed starting from the docked complex conformation. Several snapshots are presented here using cartoons.

4.2.3 PR LBD forms the "closed" conformations upon SMPR binding

From available crystal structures, PR LBD may not adopt the antagonistic "open" state upon antagonists or SPRMs (such as RU486, Asoprisnil) binding [42, 43]. Soaking technique [42, 224] was used in producing the antagonist-bound agonistic conformation with helix 12 partially destabilized [43]. It was hypothesized that the agonistic state PR LBD with antagonist bound may not be the most stable conformation, but it is also biological relevant. Starting from the "open" state LBD with asoprisnil bound (simulation system S3), we sampled the transition from antagonistic to

agonistic conformations. The SMPR or antagonist binding induces the conformational rearrangement of helix 12 from the "open" state towards the "closed" state.

At the beginning of S3 simulation, LBD adopts the antagonistic conformation, thus the $\Delta RMSD$ (see method part) is quite large, after 60 ns, it drops below -0.1 nm and the LBD conformation more resembles the X-ray crystal agonistic conformation. Therefore, starting from the antagonistic conformation (the "open" state), asoprisnil-bound LBD quickly shifts towards the "closed" state, though slightly different from the crystal P4 bound conformation. And are the detailed asoprisnil induced folding patterns are similar to P4 induced folding? Yes, and no. Similarly, during the early quick folding process, Arg899 finds its binding patterner Asn719 within a very short time scale (2 ns), their distance quickly drops from 6.5 nm to 0.3 nm. Then Arg899 forms electrostatic contacts with the Glu723 sidechain. Although the interactions are not very stable, and they facilitate the formation of a hydrophobic core by Leu893, Trp755, Val912, Ile913, and asoprisnil. After that, Met909 sidechain, as well as Met908, inserts into the hydrophobic core, leading to the patching of helix 12 and closing of the LBD. The differences are Arg899 could also interacts with the β13 Oxygen of asoprisnil, and the interactions are rather stable and create space clash for H11-H12 loop conformational optimization. Besides, Met909 never forms stable hydrogen bonds with Glu723 during the simulations.

As helix 12 remains relatively stable in the later simulation stage (after 100 ns), H11-H12 loop is still in the disordered state, and helix 11 occupies a position quite diverse from the crystal agonistic state. αC atom distance matrix based PCA calculations indicate the disordering of H11-H12 loop (Figure 20A). From the first PC (PC1) of PCA, it is clear that the system goes through a quick transition from "open" to "closed" state globally (Figure 20B). However, the diffusive behavior of PC2 of the system indicates that PR LBD has local structure arrangements. The essential dynamics analysis reveals that the major differences of the residues' motions along PC2 are lie in the H11-H12 loop region and the C-terminal end region of helix 11 (Figure 20C and 20D).

These dynamics differences thus echo the experimental finding that the crystal structure model of asoprisnil-bound LBD reveals unstable conformation.

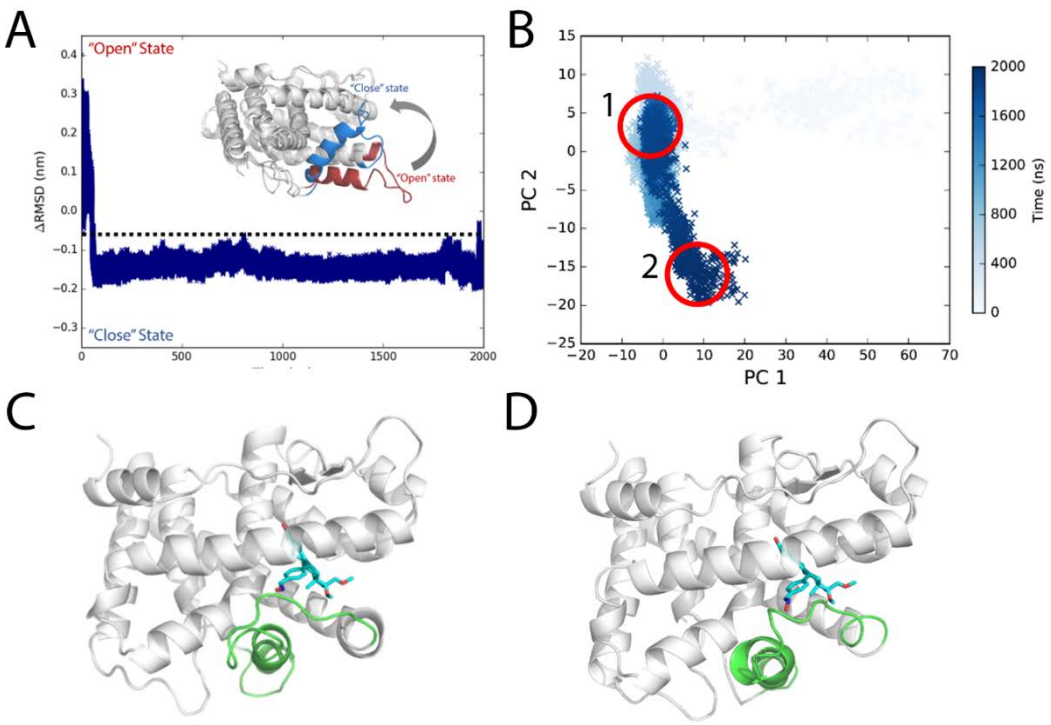


Figure 20. Conformational transitions of Asoprisnil bound PR LBD.

A, *Δ RMSD* of simulation system S4. B, the PCA projections of the first two PCs using the coordinates. C, the representative structure in PCA map region 1 in panel B. D, the representative structure in PCA map region 2 in panel B.

4.2.4 Helix 12 deforms in co-peptide bound PR LBD simulation

Starting from the antagonistic state (simulation system S6), with co-peptide bound, PR LBD goes through large dynamical changes in helix 11, H11-H12 loop and helix 12 regions. The αC RMSD of PR LBD structure with reference to the antagonistic state rises within the first 500 ns. And the conformation of PR LBD also diverges from the agonistic state, where Trp755 and Lys919 sidechains remain close to each other.

For this complex, Trp755 remains quite close to Val912 for rather a long simulation time, and Glu723 does not form any hydrogen bonds with Met909 or Met908. These evidences indicate that helix 12 C-terminal half part is restricted by the Trp755/Val912 interactions and it takes a similar position as that in agonistic conformation, whereas N-terminal half remains flexible and exploration in the solvent. However, at a late stage of the simulation (after 1500 ns), Leu2263 in co-peptide gradually forms hydrophobic interactions with Trp755 and create space clash, and therefore Leu2263 block the accessing of Trp755 by Val912 in helix 12 (Figure 21). During the simulation, helix 11 shifts towards the binding pocket and make contacts with a middle part of helix 3 and is restricted by hydrogen bonds formed between Arg899 and Glu723 sidechains (Figure 21). Thus helix 11 occupies the space

and prohibits PR LBD from forming an active agonistic state. Helix 12 is distorted and bends outwards (Figure 21).

From the simulations of LBD together with co-peptide (simulation S6), we sampled a "semi-open" state of PR LBD (Figure 21), where the conformation of helix 12 C-terminal part of is almost identical to the "closed" state, where there exist no hydrogen bonds between Trp755 and Lys919, which rotates over 90° and points towards the solvent. However, the N-terminal part of helix 12 adopts the "open" state and is still quite flexible and the binding pocket is totally exposed to solvent.

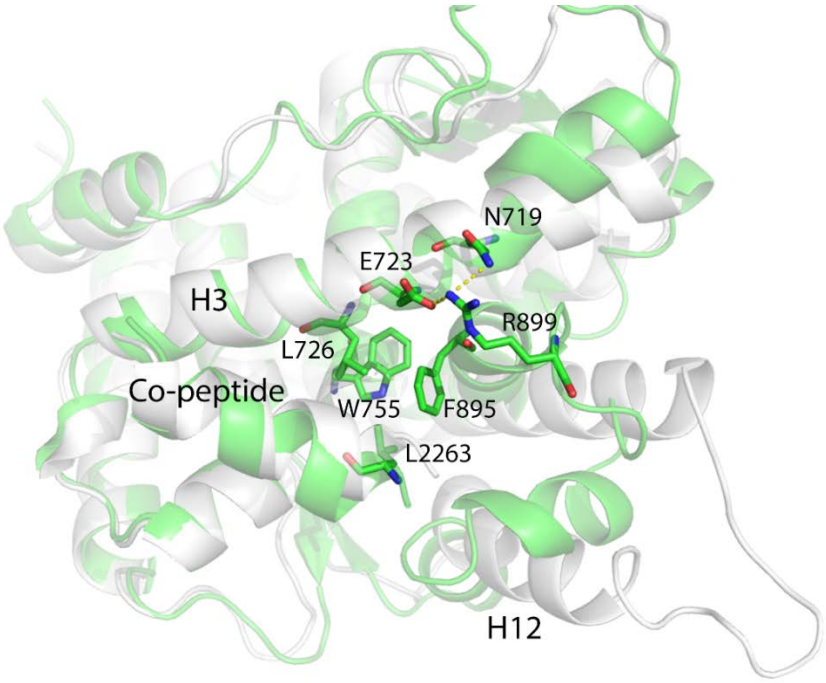


Figure 21. The representative conformation (green) of the largest population cluster superimposed with antagonistic PR LBD conformation (white).

4.3 Discussion

4.3.1 Hydrophobic effect facilitates helix 12 re-packing

The hydrophobic effect is recognized as the dominant driving force for protein folding [225-228]. Non-polar residues thus tend to be buried inside the protein core, inaccessible to solvent water molecules. With a ligand binding, the protein would become more or less stable and the conformational entropy decrease would not be favorable for ligand-induced folding [229, 230]. However, it has been proved that the translational entropy gains of solvent water molecules escaping from protein surface would at least compensate that [231, 232]. The free energy changes of the dehydration process, where stabilized water molecules surrounding protein surface are released to the bulk solvent environment, thus are also quite significant [233]. The steroids or steroid-like ligands

are hydrophobic [40, 41, 224] in nature and the binding of these ligands in not only PR LBD, but also other NR LBDs [234-236], would induce the forming of hydrophobic core and dehydration of the binding pocket, as well as PR LBD, by re-orientating non-polar residues surrounding the pocket (Figure 22A and 22B), therefore also facilitate the conformational changes of helix 11, helix 12 and H11-H12 loop. In the early stage of P4-bound induced conformational adaptation of PR LBD, the non-polar residues from helix 12 (Met908, Val912 and Ile913), helix 11 (Cys890 and Phe895), helix 5 (Trp755), as well as helix 3 (Leu727), together with the hydrophobic ligand P4, forms a hydrophobic cluster, marking as the early event of the helix 12 re-orientation. Similarly, in asoprisnil-bound induced folding of helix 12, these residues, except Cys890 and Phe895, are closing to each other and exclude the solvent molecules near the binding pocket. Sequence alignment of several other NR LBDs indicates that these non-polar residues are quite conserved (Figure 22D). Especially, Trp755 is highly conserved in all five NR LBDs, suggesting the vital importance of this residue. While the other non-polar residues in PR LBD have their equivalent non-polar counterparts in αER, AR, MR and GR LBDs.

We observed the rapid loss of water molecules in the first hydration shell upon ligand-binding induced conformational changes. Coupling with the hydrophobic cluster formation, the numbers of 1st layer hydration water outside PR LBD surface are decreasing in the early stage of the "open-to-closed" transition (Figure 22C). With ligand binding, the hydration water molecules are released from LDB surface, and there are more fixed water molecules around the *apo*-form LBD than P4-bound or asoprisnil-bound LBD. These statistics of the first hydration shell around LBD thus clearly suggest that the dehydration around LBD surface would be correlated to helix 12 repositioning.

Briefly, the dehydration of protein surface and hydrophobic cluster formation drives the conformational adaptations of PR LBD. The loss of stabilized water molecules within the first hydration shell would contribute to large favorable free energies to the helix 12 "open-to-closed" transition. Meanwhile, the ligand-induced formation of the hydrophobic cluster by several conserved non-polar residues around the binding pocket and helix 12, is a key driving force for the conformational adaptation as well.

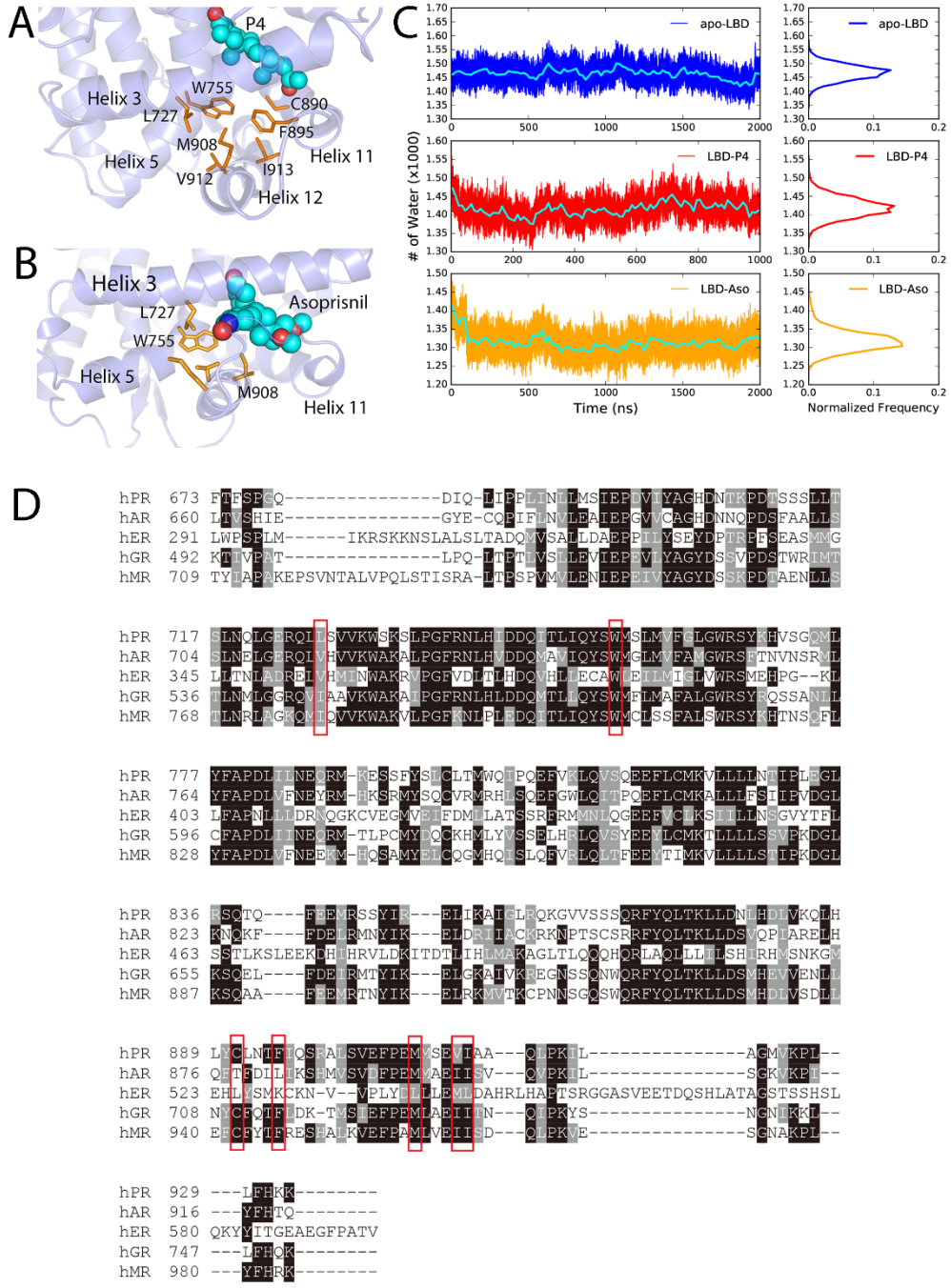


Figure 22. The hydrophobic cluster that may facilitate helix 12 patching.

A, t=30 ns the hydrophobic cluster formed during simulation. The residues forming hydrophobic core are shown as orange sticks, while the ligand atoms are shown as spheres. B, asoprisnil-bound PR LBD hydrophobic core t=60 ns. The residues forming hydrophobic core are shown as orange sticks, while the ligand atoms are shown as spheres. C, number of first hydration shell water molecules for *apo*-form LBD (MD system S1), P4-bound LBD (system S7), and asoprisnil-bound LBD (system S4) recorded during the MD simulations as a function of simulation time. The blocked averages of the time series water molecule numbers around the LBD are shown in cyan lines for every 1 ns. D, the amino acid sequences alignment of PR, AR, αER, GR and MR LBDs.

4.3.2 Electrostatic interactions contribute to helix 12 patching

Short-range electrostatic interactions are important in protein folding and stability [225-228]. Statistic analysis indicates that around 80% of the ion-pairs or polar-interactions of proteins are exposed to solvent. In PR LBD, we also identified several key short-range electrostatic interaction pairs or hydrogen bonds pairs.

Cancer genome sequencing studies discovered that Glu723 mutation was recorded in breast cancer samples [237], indicating the vital role of Glu723, though whose role has never explained in atomic level. In αER, an equivalent residue Asp351 also stabilizes helix 12 in agonistic conformation LBD, the D351Y mutation and other artificial mutations thus abolish the agonistic effect of αER [43, 234, 238, 239]. Our previous simulation study also suggests that Glu723 is actively involved in the conformational dynamics of PR LBD [218]. Here, in this study, the highly frequent electrostatic interaction and hydrogen bonds between Glu723 and Arg899 are recorded in the early "open" to "closed" transitions of P4-bound LBD and asoprisnil-bound LBD. The distance decreasing of Glu723/Arg899 from around 1.5 nm to 0.3 nm would contribute to a 2.5 kJ/mol electrostatic energy change. Another hydrogen bond forming by His888/Gln916 also stabilizes helix 12 in the "open" to "closed" state transition of PR LBD.

Bridging water molecules in protein folding process ares rather commonly seen [229]. These water molecules would accelerate protein folding by taking a structural role, hydrogen bonding between two hydrophobic residues through their hydrophilic backbone atoms. However, in a later stage, these water molecules would be expelled from the hydrophobic core. These collective emptying of water molecules from the hydrophobic core are quite common and the process is called dewetting (drying) effect, where the water molecules escape and large vapor bubbles form [229, 240]. We also notice the existence of the structural water in our simulations (Figure 23A) around Trp755 and Val912. Interestingly, Trp755 captures Val912 backbone oxygen atom in helix 12 indirectly, bridging by a conserved water molecule (Figure 23B), as observed in several crystal structures (PDB ID: 1A28 [40], 1SQN [241], 2W8Y [42], 3D90 [41], 3G8O [242], 3ZR7 [224] and 3ZRB [224]). The residence time of this interior water molecule is around 2~20 ns, which is well correlated with the experimental residence time of buried water molecules [222]. The resampling of the important structural water not only shows that our simulations could capture key features of PR LBD dynamics but also indicate that the conserved structure water could facilitate the formation of hydrophobic cluster formation by ligands and Trp755, as well as other residues.

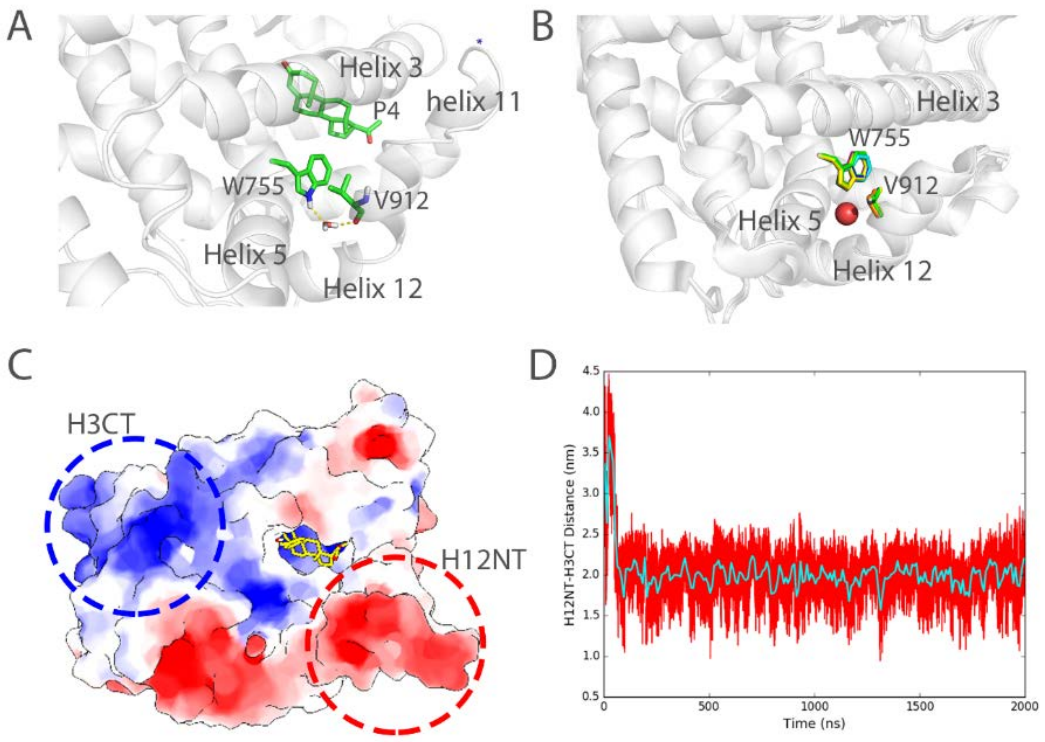


Figure 23. Structural water bridging helix 12 stabilization and long range electrostatic interactions in PR LBD simulations.

A, stable interaction formed between Trp755 and Val912 bridging by a stable water molecule in P4-bound LBD simulation (t=100 ns). B, structural alignment of several crystal structures of PR LBD with a conserved water molecule (red sphere) bridging Trp755/ Val912 interactions, the colors of the PDB structures are green for 1A28, blue for 1SQN, yellow for 2W8Y, meganta for 3D90, cyan for 3G80, orange for 3ZR7, and grey for 3ZRB. C, a cartoon representation shows the two highly charged centers in PR LBD surface. D, the distance between H12NT and H3CT during asoprisnil-bound LBD simulation, while the cyan line shows the time-blocked average of the distance.

Furthermore, long-range electrostatic interactions may also contribute to the helix 12 patching process. The N-terminal end of helix 12 (H12NT) is negatively charged (the negative charge center), mainly due to the existence of several polar residues Glu904, Glu907, and Glu911. The 3 residues mentioned above, as well as the whole H11-H12 loop, were missing in the crystal antagonistic structure model, and thus it indicates that this region and the loop are rather mobile [25]. Another region (positive charge center), the C-terminal end of helix 3 (H3CT), however, is highly positively charged due to two polar residues, Lys733 and Arg740. The energetic favorable long-range electrostatic attractions between the positive charge center and the negative charge center thus would further contribute to the "open" to "close" transition. The distance between the positive charge center and the negative charge center shrinks from about 3.6 nm to 2.0 nm, and it would contribute to a 2.5 kJ/mol electrostatic potential energy change. These electrostatic interactions, hydrogen bonds, and indirect water bridging, are crucial for PR LBD ligand-bound conformation reconfiguration and structural ability.

4.3.3 Bulk 11β group blocks stable helix 12 hooking and H11-H12 loop patching

There has already long been discussed that Met909 sidechain would create sterically impede with the bulk 11β group on steroid scaffold of PR antagonists (mifepristone and) and SPRM asoprisnil [42, 224, 243, 244]. The space clashes would then explain the partial agonism, or the loss of agonist response, of the molecules [25, 41, 224, 236].

From the asoprisnil-bound simulations, it is clear that the bulk 11β group works as a double edge sword for PR LBD agonistic LBD formation and stabilization. In one way, it would facilitate the positioning of helix 12 to the agonistic state, by stabilizing the electrostatic interaction network forming by Arg899, Glu723, Asn719, and SPRM asoprisnil, and further restricting the motions of helix 11 C-terminal region. On the other way, however, it makes clashes with the ligand and inhibits the close contacts between Met909 and Glu723, therefore prohibits the hydrogen bonds formation and helix 12 conformational adaptation. As Met909 plays a vital role for the agonistic conformation stability as revealed in several studies [25, 41], the Met909/Glu723 hydrogen bond contributes to a favorable free energy to the "closed" agonistic state, it also shields the ligand binding pocket from solvent access. Thus, the inability to form this critical hydrogen bond would result in an unstable helix 12, and sequentially contribute to the partial agonism of asoprisnil-bound PR LBD [41, 42, 236]. The bulk 11β group also destabilize the agonistic contacts between H11-H12 loop with helix 3.

Comparing to the P4 induced conformational adaptations, there are two key differences in asoprisnil induced conformational changes. Firstly, Met909 forms hydrogen bonds with Glu723 in P4-bound LBD, but not in asoprisnil-bound LBD simulations. The loss of hydrogen bonds between helix 12 and helix 3 thus renders helix 12 more flexible and mobile comparing to the crystal "closed" state. Secondly, in P4-bound LBD simulations, H11-H12 loop is rigid and tightly aligned with helix 3 and form contacts with helix 3 through polar residues, but in asoprisnil-bound LBD simulations, H11-H12 loop is rather mobile and does not form stable interactions with helix 3. However, the interactions between helix 3 and H11-H12 loop is vital to maintain the "closed" conformation of PR LBD.

We believe that the bulk group in the 11th position of steroid ligands would hinder the occupation of agonistic position for helix 12 and H11-H12 loop. Though SPRM molecule (asoprisnil) also induces the conformational adaptation of helix 12, it also contributes to the instability of this helix, as well as helix 11 and H11-H12 loop. The SPRM-bound PR LBD dynamics thus correlates with the finding of partial agonism of the ligand.

4.3.4 Conformational adaptions of PR LBD upon ligand binding

It has been proved that *apo*-form PR LBD is flexible and there exist multiple intermediates, among which the agonistic "closed" conformation is a rather stable state. Ligands (and/or co-peptides) binding, however, may induce conformational changes or stabilizes one of the intermediates. From our simulations, we recorded the evidence of ligand and co-peptides binding-induced conformational adaptations and observed some of the interaction patterns formation using extensive MD simulations.

Based on the simulation results, we proposed a possible roadmap of PR LBD conformational adaptation induced by ligands or co-peptides binding (Figure 24). Firstly, both agonist and SPRM would induce the formation of agonistic-like "close" state PR LBD, which helix 12 covers the hydrophobic ligand binding pocket. However, in SPRM-bound LBD, the "close" state complex is not stable, helix 12 N-terminal could not be hooked by Glu723, and H11-H12 loop could not be re-orientated to tightly patched against helix 3 in a short period. These conformational adaptations thus explain the partial agonism of SPRM [243, 245].

Secondly, helix 12 in the LBD with co-peptides binding in the cleft formed by helix 3 and helix 5, would bend outwards, thus the PR LBD samples the semi-antagonistic conformations, which would be favorable for both agonists or antagonists binding, but resisted to co-activator binding [246]. Thirdly, with both co-peptides and SPRM ligand binding, the antagonistic PR LBD conformations are well maintained. This is not surprising, since the SMRT co-peptides could form a ternary complex with RAR in complex with RAR-response element [247, 248]. The agonist binding related releasing of SPRM co-peptides, from RAR and thyroid-hormone receptor (TR) [247], or possibly from PR, may result from the spatial clashes between helix 12 and the SMRT co-peptides [25], created by the conformational changes of helix 12 from the antagonistic state towards the "close" state. And it is highly possible that the antagonistic effect of SPRM is the result of competition binding [25, 246, 248] of the co-repressor peptides against co-activator peptides, since only with both SPRM and co-repressor peptides binding, PR LBD could adopt the antagonistic conformations. Meanwhile, an antagonist or SPRM ligand binding with LBD would facilitate the co-repressor peptides association and suppress the co-activator binding [25, 246].

However, there are still missing parts in this roadmap, such as how would the *apo*-form LBD react to the pure antagonistic binding, though we do observe the destabilization of agonistic LBD with an antagonist (RU486) binding? Meanwhile, what could be the transition timescale of *apo*-form LBD to stabilized agonistic-like "close" state LBD with P4 or SPRM binding? And how antagonism of antagonists is related to conformational changes of PR LBD? Further analysis of the conformational transition kinetics, as well as an antagonist binding, would be required to further enhance our understanding of this PR LBD conformational adaptation roadmap.

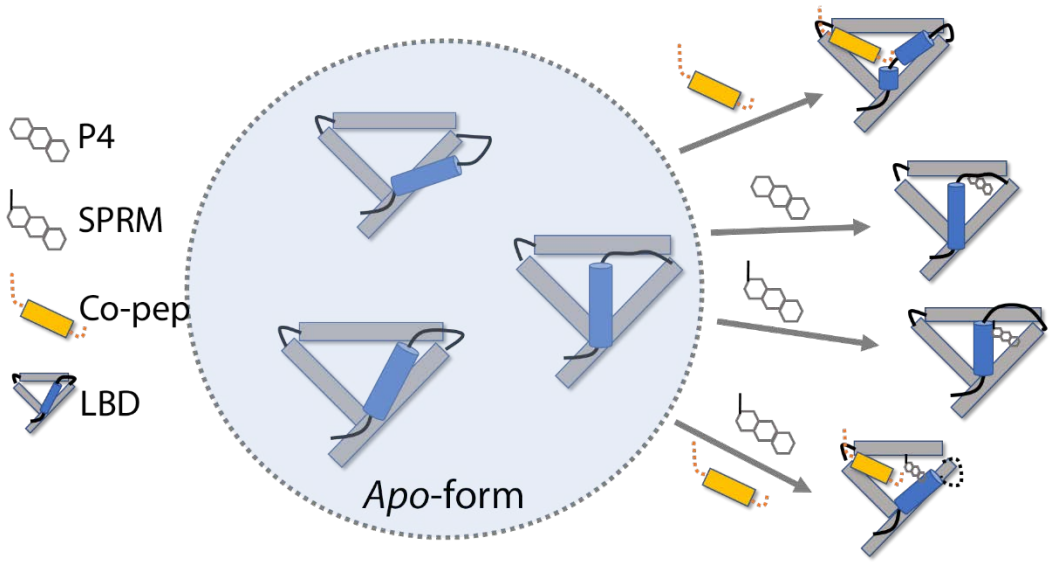


Figure 24. PR LBD conformational adaptations. SMRT co-repressor peptides are shown in orange. Only 4 helices (helices 3, 5, 11 and 12), as well as some loops, in the LBD are presented, whereas helix 12 is in ice-blue color. Key loops and tails in the LBD are shown as black lines.

4.4 Summary

The dynamics study of PR LBD is extremely important for a thorough understanding of the structure-function relationship. Current methods are not effective to carter this demand. In this chapter, we introduced extensive cMD simulations to explore the possible ligand induced conformational changes and adaptations. We observed the conformational transitions from “open” state to “closed” state accompanied by the large re-orientation of helix 12, with either a progesterone (agonist) or the asoprisnil (SPRM) binding. The conformational changes of the LBD are majorly marked by three states, 1) hydrophobic core formation, 2) local salt-bridges formation, 3) minor structure adjustment. The driven forces of the re-orientation are majorly the hydrophobic interactions between the ligand and its surrounding residues, as well as the electrostatic interactions between helix 12 and helix 3. In contrast, with both the SPRM and the co-repressor peptides binding, the “open” state PR LBD structure could be well maintained. Lastly, we composed a ligand/co-peptide included conformation adaptations of PR LBD with various binding patterners.

Chapter 5 Virtual screening study towards PR LBD drug discovery

PR, especially the LBD, is an important target for a drug. There have been several drugs targeting this receptor for abortion, hormone replacement treatment, breast cancer treatment and so on. However, the available drugs of PR all somehow bring side effect. The continuous development of PR LBD drugs is beneficial for female health. Besides, both pure agonists, antagonists, and the SPRM molecules are useful in some aspect, as we discussed in Chapter 2. Although there were hundreds of new lead molecules have been identified and having quite high binding affinities, the mature drugs targeting PR LBD are still limited. Here, starting from the cMD sampled *apo*-form and *holo*-form PR LBD conformations, we performed large-scale VS against this domain and selected several high potent lead-like molecules for future verification.

5.1 Methods

5.1.1 Docking and Virtual Screening

The crystal structures of the PR LBD were downloaded from the RCSB protein data bank. If there are a homodimer of LBD in a structure, only one chain is harnessed, otherwise, both chains are kept as receptor conformations for re-docking simulations. The original ligands co-crystalized in the LBDs were docked back into the binding pocket using GOLD [101]. The RMSD of the ligand poses were compared with their native positions. If the RMSD of a docking pose with respect to its native state is less than 2 angstroms, it is recognized as an accurate redocking process.

It is clear that GOLD could successfully recover the native conformations of the ligands. The docking simulations were performed using GOLD with an ensemble docking strategy where multiple target (receptor) conformations were applied to include receptor flexibility in virtual screening [96]. Three receptor conformations are chosen to screen a combined library database composed by Chembridge premium set, Maybridge screening collection, Otava prime screen 15, Enamine diversity drug-like set 2016, FDA approved drugs, 147489 ligands in total. The 3 receptor conformations are summarized in the following table 5. The receptor conformations were processed using Discovery Studio, and the ligand database were loaded into discovery studio to assign protonation states at pH = 7.0. In Gold, default parameters were chosen and the Goldscore scoring function was used. A 12 Å distance cutoff was set to define the active region.

Table 5. the receptor conformations for SBVS.

SN	Description	RMSD1 (nm)	RMSD2 (nm)	Pocket Size
C1	Representative of LBD/Asoprisnil simulation dPCA L7	0.45	0.36	830

C2	Representative of *apo*-LBD US sampling usL2 minima	0.43	0.39	685
C3	Representative of *apo*-LBD metadynamics simulations	0.40	0.32	702

The RMSD1 is calculated using the agonistic conformation as a reference, while the RMSD2 is calculated using the antagonistic conformation as a reference. Two ligands, P4, and asoprisnil, have been selected as positive controls.

Although many studies have indicated that docking scores or energies are notoriously bad for binding affinity predictions, they are quite commonly used for ligand binding affinity ranking in practice due to their easy to implement. Here, we adopted another ligand ranking strategy, using molecular weight normalized scores to rank the ligands, to minimize the molecule size bias in binding strength estimation in molecular docking.

5.1.2 Simulation of docking poses with PR LBD

The topologies of the selected ligands were processed using AmberTool 16 [249] and converted into Gromacs topology file format. Firstly, the atomic charges of a small molecule ligand were determined using antechamber module in AmberTool 16 [249] by AM1-bcc charge model. Amber gaff force field [177] parameters were supplied to estimate the bonded and non-bonded parameters for a ligand. The amber format topology files of the ligands were converted into Gromacs format using Acpype python model [250]. The protein (LBD) force field was Amber99SB-ildn [174], and the explicit solvent model TIP3P [178] was used for waters, and ions concentration was set as 0.15M to mimic the biological relevant environment. Equilibration and product runs were under NVT under 300 K and NPT ensembles at 1 bar pressure, respectively. The thermostats and pressure coupling were maintained by velocity rescale and Berendsen algorithms.

After 1000 steps of energy minimization, a 1 ns equilibration was performed for each ligand-LBD complex. Then 50 ns production runs were carried out using Gromacs 5.1.2 [216] on a GPU cluster. The same simulation parameters were selected, as described in section 3.3.1.

5.1.3 Binding Free energy estimation of ligand-LBD complex

Binding energy is a fundamental quantity to assess the binding strength of molecules. However, the determination of the absolute binding energy is rather troublesome, if not computational accessible. In order to compare the binding energies of the simulation systems, we used the approximated binding energy method molecular mechanics (MM) Poisson-Boltzmann (PB) surface area (SA) (MMPBSA) to calculate the binding energy between the LBD and the ligand with the g_mmpbsa tool [251] for the various simulation systems. The theory of MMPBSA has been thoroughly covered in many researches [251-253], and we are not going to explain in detail here. In the process, we ignored the entropic contribution to the binding energy, and only included the MM, PB and SA parts. For calculating MM and PB, the dielectric constant was set as 4.0. Other parameters were chosen as the

default values. Only the last 40 ns trajectory of each system with a stride 100 ps was adopted for obtaining binding free energies.

The machine learning based protein-ligand binding energy predictions were also performed, using the online server TML-BP (http://weilab.math.msu.edu/TML/TML-BP/) [254]. The prediction models were trained using known crystal structures in BindingDB. And the authors extracted the binding related topological information like features, element-specific persistent homology, following by machine learning model training. Although the prediction models were constructed based on a diverse set of protein-ligand complexes, their ability to predict the binding energies has been proved to have a higher than 80% correlation with experimental values.

5.2 Results

5.2.1 Antagonistic conformations are more suitable for SBVS

There already around 20 crystal structures of PR LBD bound with ligands deposited in RCSB PDB database. Among those structure models, only 2ovh, 2ovm, and 4oar capture the antagonistic conformation of PR LBD, although with additional co-peptides binding. Except for the displacement of the helix 12, antagonistic PR LBD conformations differ from the agonistic conformations also in the size of the binding pocket (Figure 25). The binding pocket size of the PR LBD ranges from 500 $Å^3$ to 1200 $Å^3$. The removal of the SPRM (asoprisnil) from the antagonistic PR LBD conformation results in a decrease of around 600 $Å^3$. Therefore, PR LBD could accommodate its binding pocket to a large variety of various sizes of small molecules.

To screening large libraries for both antagonists and agonists, it is desirable to start with the PR LBD conformations whose binding pockets could be large enough or opened for ligand binding. Meanwhile, we performed a series of docking simulations applying the crystal structure models as docking receptor conformations (Figure 25). The original LBD-bound ligands (control set) in the crystal structures were also re-docked into PR LBD binding pocket. It is clear that for the three antagonistic PR LBD crystal structure models, the enrichment of the top 1000 and top 2000 ligands is maximized to include all the control set ligands. In conclusion, we believed that antagonistic conformations could better enrich the active molecules, due to the feasibility of pocket closing than pocket opening.

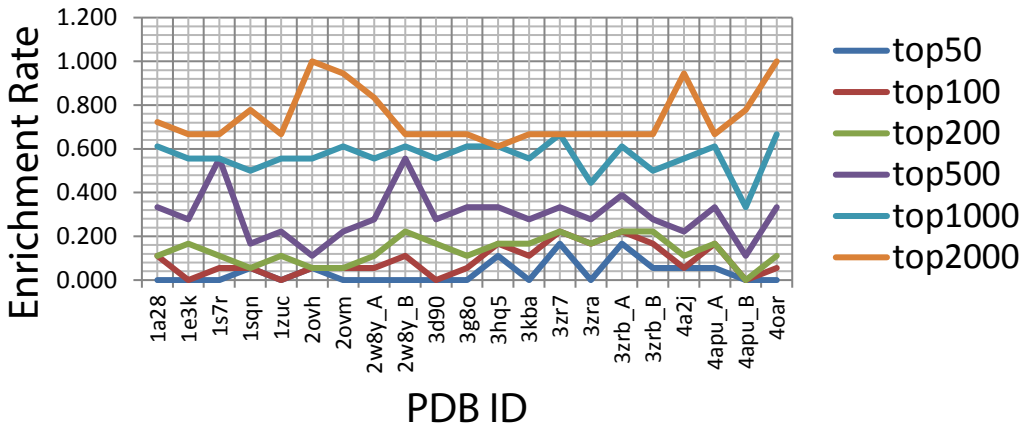


Figure 25. Enrichment rate in the cross-docking simulations.

5.2.2 Common top-ranking ligands screened using GOLD

We used GOLD software to identify potential potent ligands, based on the hypothesis that higher docking scores represent stronger binding affinities. Secondly, it was proved that incorporating receptor sidechain flexibility would lower down the false positive rate during SBVS. Therefore, we selected three initial LBD conformations for VS using GOLD. If a molecule could be docked into the three conformations with mutually high scores, then we call the molecule a common top-ranking ligand. The ligands, which show up in the 1000 top-ranking scores lists of both three conformations' docking simulations, and their docking scores as well as molecular weight normalized scores, are listed in Table 6.

Table 6. The docking scores of the common top 1000 ligands

Molecules	C1		C2		C3		Molecular Weight
	Normalized Score	Gold Score	Normalized Score	Gold Score	Normalized Score	Gold Score	
M3	11.72	71.59	12.01	73.38	11.47	70.08	449.46
M6	12.43	76.97	11.69	72.40	11.43	70.79	489.47
M7	11.45	70.94	11.79	73.05	12.78	79.23	491.33
M9	11.51	69.13	11.88	71.35	12.28	73.77	406.32
M12	11.41	70.52	11.82	73.06	11.33	70.03	482.4
M17	11.14	66.51	11.65	69.53	11.56	69.03	392.31
M20	11.73	72.68	12.24	75.84	11.63	72.08	490.4
M21	11.26	69.05	11.66	71.51	11.54	70.78	461.41
M22	11.35	69.16	12.84	78.18	12.81	78.03	441.77
M31	11.69	71.29	11.95	72.86	11.46	69.90	444.8
M32	11.37	69.74	11.66	71.53	11.85	72.69	462.35
M34	11.37	67.55	12.13	72.07	11.62	69.04	381.31
P4	8.95	51.49	9.27	53.28	9.41	54.08	314.21

Asoprisnil	11.23	68.57	11.33	69.21	9.86	60.25	449.29

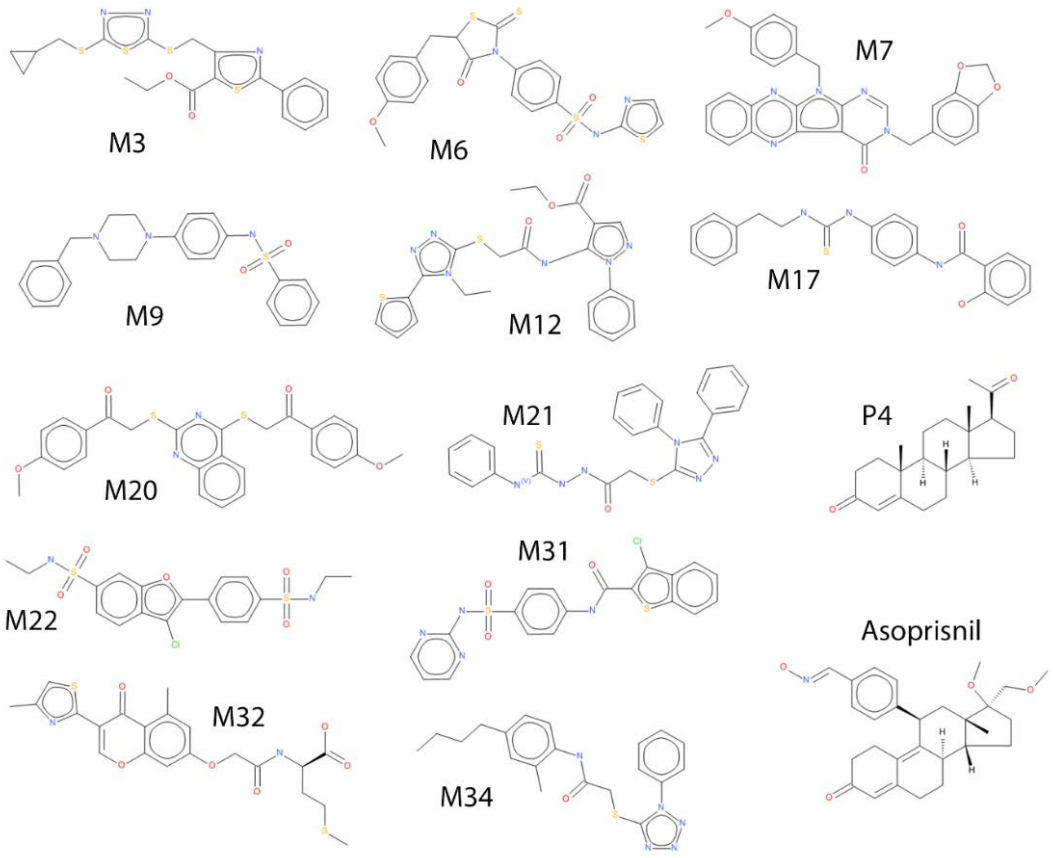


Figure 26. The top common ranking molecules, screened from an aggregated library, as well as the two control molecules.

The two control molecules (P4 and asoprisnil) could be docked into PR LBD antagonistic conformation C1 binding pocket, with 51.49 and 68.57 scores respectively. The PR LBD conformations were sampled from antagonistic structure simulations. Thus, the antagonists or SPRMs would be favored during molecular docking. And in order to decrease the effect of large molecule weights, the Gold docking scores would be normalized by the algorithms of the molecular weights.

The docking algorithm has identified abundant high score ligands, of which a large amount of false positive ligands exists. In general, we need to identify a molecule, which has a high docking score, and it is truly a binder with a strong binding affinity. The two control molecules, P4, and asoprisnil, are not in the common 1000 top-ranking molecule list. Thus, it may suggest the necessity of developing methods to reduce the false positive rate.

5.2.3 MD simulation of the common top ligands

The common 1000-top-ranking molecules were chosen for short MD simulations in explicit water environment as well as 0.15 M NaCl salt condition. The stability of the ligand-LBD complex could be simply checked using the RMSD of the LBD backbone. Similarly, the number of hydrogen bonds, number contacts, as well as the binding free energies between a ligand and its receptor could be used to access the binding strength. The key hydrogen bonds between a ligand and the LBD would be of crucial value.

The known active SPRM molecule asoprisnil and agonist P4 both could form at least two hydrogen bonds with PR LBD, as detected in MD simulations, though with different residues. P4 forms two hydrogen bonds with the LBD at its two sides (Figure 27A). The sidechains of Arg766 and Thr894 work as a hydrogen donor and the two oxygen atoms in P4 are hydrogen acceptors. Therefore, P4 is anchored by helix 11, and it also stabilizes the hydrophobic pocket. As for asoprisnil, the hydrogen bond formed between PR LBD Asn725 and β11 oxygen resembles the hydrogen bond formed between Arg766 and P4 β11 oxygen atom (Figure 27B). Glu723 in helix 3 also forms a stable hydrogen bond with 17 beta bulk group hydroxyl hydrogen. The large asoprisnil thus clamps helix 3 to ensure a tight binding with PR LBD. The critical hydrogen receptor or donor resides, such as Arg766, T894, Asn725 as well as Glu723, are potentially responsible for strong ligand binding, thus they are worth to be treated carefully in ligand selection.

We calculated the hydrogen bond formation frequencies during the simulations (Table 7). Clearly, asoprisnil could form at least one hydrogen bond with PR LBD during over 75% of the simulation trajectory. This frequency indicates the stable hydrogen bonds could be formed, as also detected in our previous simulation starting from crystal antagonistic conformation LBD/asoprisnil simulations. As for P4, however, the hydrogen bonds formed between LBD/P4 are barely maintained partially due to the fact that the P4 induced PR LBD conformation adaption is not observed in this limited simulation. Nevertheless, the hydrogen bonding frequencies could be a good indicator of high potent ligands.

Several ligands also show high hydrogen bond formation frequencies (>0.3). For example, M6, M17, M22, and M31 both have over 1 hydrogen bond with PR LBD during over 50% of the simulation time. While M20 also forms at least 1 hydrogen bond with PR LBD in around 40% of the simulation time.

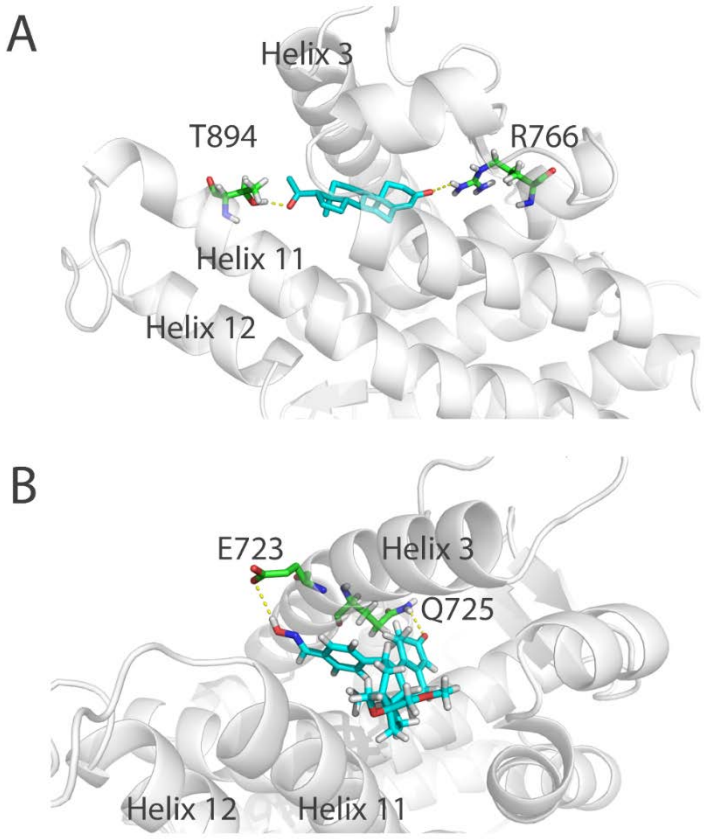


Figure 27. Hydrogen bonds formed between PR LBD and its active binding molecules.

Table 7. Hydrogen bonds formation frequencies between a ligand and PR LBD.

Name	n=1	n=2	n=3
M3	0.015	0.000	0.000
M6	0.506	0.250	0.019
M7	0.078	0.000	0.000
M9	0.000	0.000	0.000
M12	0.045	0.000	0.000
M17	0.504	0.177	0.002
M20	0.394	0.002	0.000
M21	0.145	0.020	0.000
M22	0.102	0.673	0.165
M31	0.409	0.157	0.072
M32	0.135	0.027	0.010
M34	0.002	0.000	0.000
M40	0.200	0.025	0.000
M50	0.476	0.287	0.002

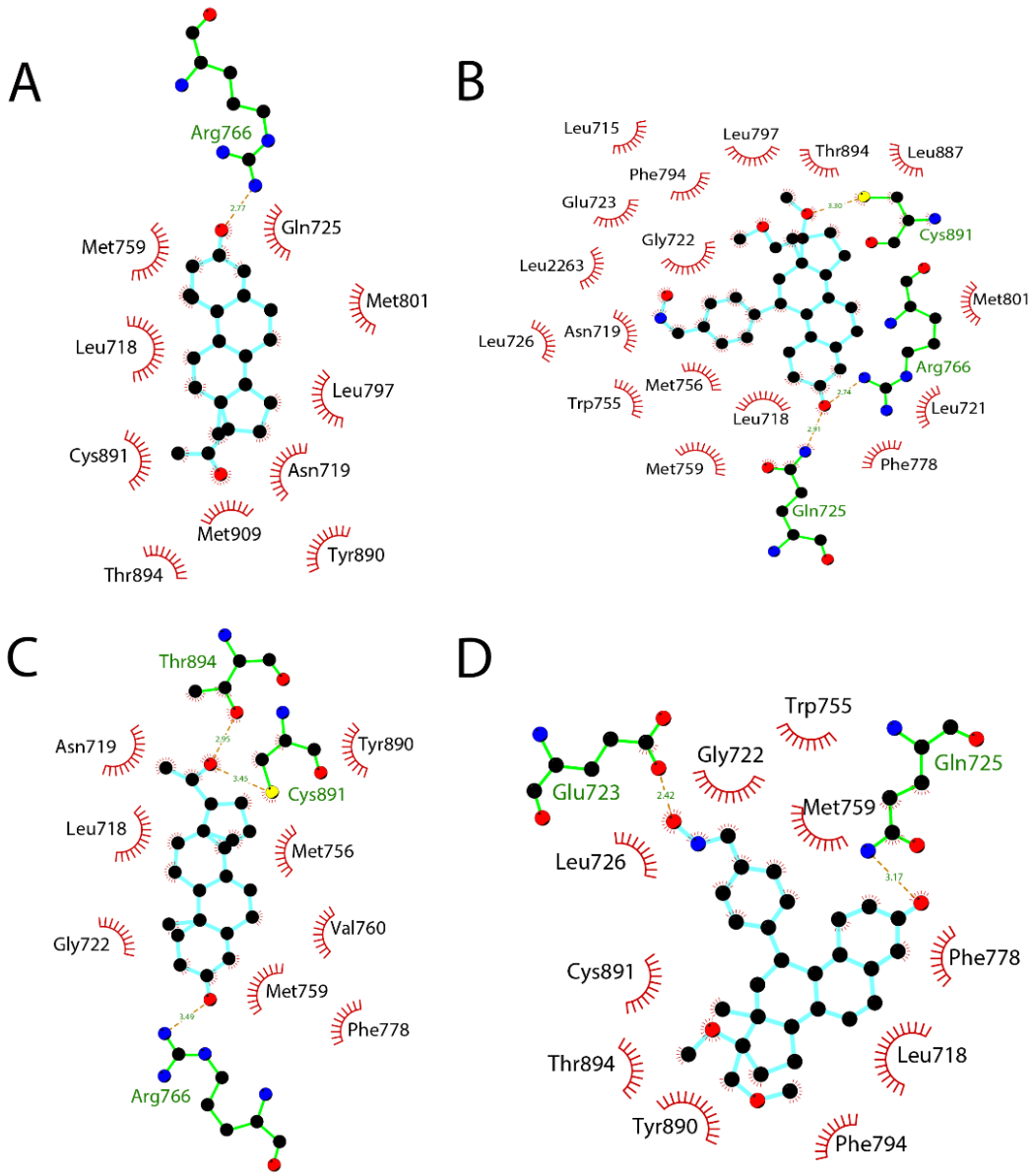


Figure 28. The interaction patterns between LBD and its active compounds.

A, the interactions between the agonistic LBD and P4, conformation adopted from PBD ID 1A28. B, the interactions between the antagonistic LBD and asoprisnil, conformation adopted from PBD ID 2OVH. C, the interactions between the antagonistic LBD and P4, conformation adopted P4-LBD re-docked MD simulated late frame. D, the interactions between the antagonistic LBD and P4, conformation adopted asoprisnil-LBD re-docked MD last simulated frame. In all panels, the ligand atoms are linked by cyan lines, while the bonds in polar contacting residues are shown as green lines, and the carbon, oxygen, sulfate and nitrogen atoms are in black, red, yellow and blue respectively. The van del Waals interactions are shown as brick red spikes; the polar contacts are shown as orange dashed lines.

5.2.3 Binding free energies of the common top ligands

A great amount of time and resources is required to obtain the absolute binding free energy analytically, alternatively, it is affordable to using MMPBSA binding free energy estimation to approximate the binding strength of a receptor-ligand complex. Here, the binding energy of the receptor-ligand complex was calculated using MMPBSA method, while the entropic contribution was ignored. With the aid of MD simulations, we, therefore, could get the more realistic estimations of binding energy by averaging a number of frames in a simulation trajectory, based on the hypothesis that MD force field could more precisely model the complex than the docking scoring function.

The binding free energy of the SPRM asoprisnil/LBD complex is relatively low, indicate a strong binding of the molecule. Whereas, the binding free energy of the P4-LBD complex is 35 kJ/mol higher than asoprisnil/LBD complex (Table 8). The van der Waals term of the asoprisnil/LBD complex is much lower than that of the P4/LBD complex, partially due to the bulk size of asoprisnil. In the meanwhile, the polar solvation term of the asoprisnil/LBD complex is much higher than that of the P4/LBD complex, indicating the unfavorable desolvation during the binding process. Overall, the large molecular weight of asoprisnil, in one way, would contribute the favorable van der Waals energy. In another way, its polar groups and large size would lead to the increase of the polar solvation term.

Table 8. Binding free energies of common top-ranking molecules towards PR LBD

Ligand Name	Vdw		Electrostatic		Polar solvation		Sasa		Binding Free Energy	
	Aver.	s.d.	Aver.	s.d.	Aver.	s.d.	Aver.	s.d.	Aver.	s.d.
M3	-236.13	15.23	-3.60	2.33	95.24	8.97	-23.69	0.95	-168.18	16.69
M6	-231.56	26.33	-13.52	6.70	138.75	19.32	-24.32	0.95	-130.65	12.30
M7	-227.30	12.26	-5.35	2.86	101.62	8.74	-23.15	0.81	-154.19	11.17
M9	-225.14	9.81	-1.71	2.42	96.08	5.85	-21.49	0.82	-152.26	10.50
M12	-251.48	14.87	-10.67	3.00	110.18	7.04	-23.87	0.89	-175.83	14.50
M17	-230.44	12.88	-7.84	8.76	131.66	13.88	-22.01	1.04	-128.62	12.03
M20	-214.96	19.40	-7.73	5.70	96.49	17.11	-22.80	1.51	-148.99	14.55
M21	-179.06	35.49	-8.31	9.97	137.46	33.47	-20.47	2.42	-70.38	15.56
M22	-228.42	14.50	-103.45	10.99	229.14	14.27	-22.64	0.82	-125.37	14.78
M31	-216.54	20.22	-13.67	8.02	111.38	12.36	-20.28	1.08	-139.10	17.88
M32	-258.75	14.36	-17.56	7.34	125.16	17.11	-23.82	0.99	-174.98	14.69
M34	-252.52	19.59	-5.80	3.05	89.96	10.26	-23.02	0.94	-191.38	16.16
P4	-183.61	13.51	-10.40	4.89	63.66	8.70	-17.89	0.69	-148.24	12.31
Asoprisnil	-234.72	14.22	-20.99	9.51	94.94	17.45	-23.08	0.79	-183.85	17.89

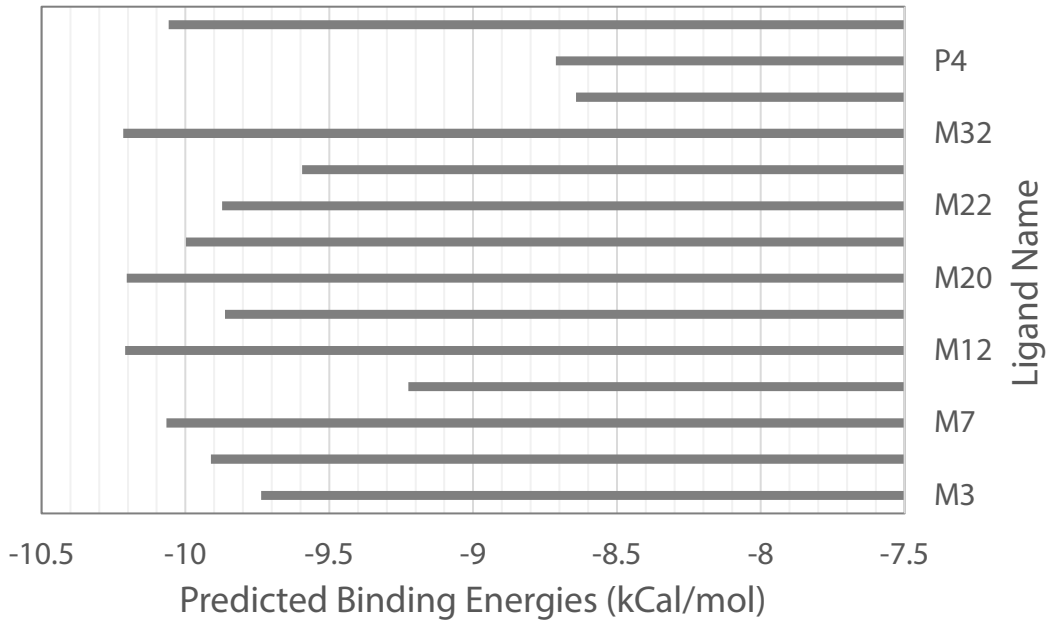


Figure 29. The TML_BP predicted binding energy of the ligands when bound to PR LBD.

The predictions were performed using online server TML_BP, using the MD generated the last frame as input ligand pose and receptor structures.

We noticed that the binding free energy M34/LBD complex is the lowest among the other molecules/LBD complexes. The van der Waals term of the M34/LBD complex is even lower than the asoprisnil/LBD complex, while the other terms are similar. Given the fact that the molecular weight of M34 (MW=449.29) is much smaller to that of asoprisnil (MW=381.31), it is quite surprised to have a much lower van der Waals term for M34/LBD complex. Therefore, it is worth trying to further uncover the high binding strength of M34/LBD complex.

And the binding free energies of M3/LBD, M7/LBD, M9/LBD, M12/LBD, M20/LBD and M32/LBD are not too high comparing to P4/LBD complex and asoprisnil/LBD complex. Among these molecules, M20 and M32 form transient hydrogen bonds with LBD during MD simulations (Table 6). Though M17 forms stable hydrogen bonds with LBD, the binding energy of M17/LBD is not favorable when compared to P4/LBD and asoprisnil/LBD.

Topological based machine learning method showed quite good binding affinity prediction results [reference here] [254]. Using this binding energy prediction method, we got another set of binding energies (Figure 29). These binding energies are quite different to values predicted using the MMPBSA method, with a very low correlation coefficient R^2 = 0.0271. Considering that the MMPBSA method did not include the entropic effect, it is not surprising to have a low correlation with the ML-based method. When comparing to the hydrogen bond formation ratio during the simulations, it seems like that the molecules form more frequent hydrogen bonds have higher TML_BP binding energy scores. M32, M20, M17, and M12 could bind to PR LBD with comparing

to asoprisnil/LBD binding energies, and both of them could form some hydrogen bonds during the simulations. However, although M3 could form very stable hydrogen bonds with PR LBD, the binding energy predicted by PML_BP method is much higher than the control molecule asoprisnil. As for M34, the TML_BP method considers the M34 is a bad binder for PR LBD. Meanwhile, M34 could not form any stable hydrogen bonds with PR LBD through the MD simulation. Thus, we do not take M34 as a potential PR LBD ligand.

Through MMPBSA based binding free energy calculations, as well as the machine learning based binding energy predictions, we identified several possible LBD strong binders, such as M12, M17, M20, M32.

5.3 Discussions

In this study, we identified 4 high potential ligands, which forms a relative stable complex with PR LBD, as suggested through hydrogen bonds frequencies, MMPBSA binding free energy calculation and the machine learning based binding energy prediction. The RMSDs of the PR LBD αC atoms for these ligand/LBD complexes are relatively stable during the first 50 ns simulations, except M17/LBD complex. Besides, some common van del Waals interactions are conserved among the ligands, as well as the control molecules (P4 and asoprisnil). As observed in crystal structures, Leu718, Asn719, Gln725, Met759, Arg766, Leu797, Met801, Tyr890, Cys891, Met909, as well as other residues, form the agonistic ligand binding pocket. While for antagonistic ligand binding pocket, additional residues are involved, such as Leu726, Met756, Trp755, Phe778, Phe794, Leu887, Thr894 and so on. Meanwhile, there are also some residues which are interacting with the ligand in the agonistic state would not form contacts in the antagonistic states, such as Met909, which resides in helix 12. Here, we noticed that most of the residues are conserved and remain contacts with those ligands. For example, in M20/LBD complex, nearly all non-polar residue-ligand contacts in asoprisnil/LBD complex have been recovered. Though, the hydrogen bonds formed between asoprisnil and LBD are not detected. Combining the hydrogen bonding frequencies, as well as the MMPBSA binding energies and the TML_BP binding energies, we identified the 4 potential PR LBD good binders, M12, M17, M20, and M32. However, we need to admit that the experimental binding affinity measurement of the ligand with PR LBD is still useful.

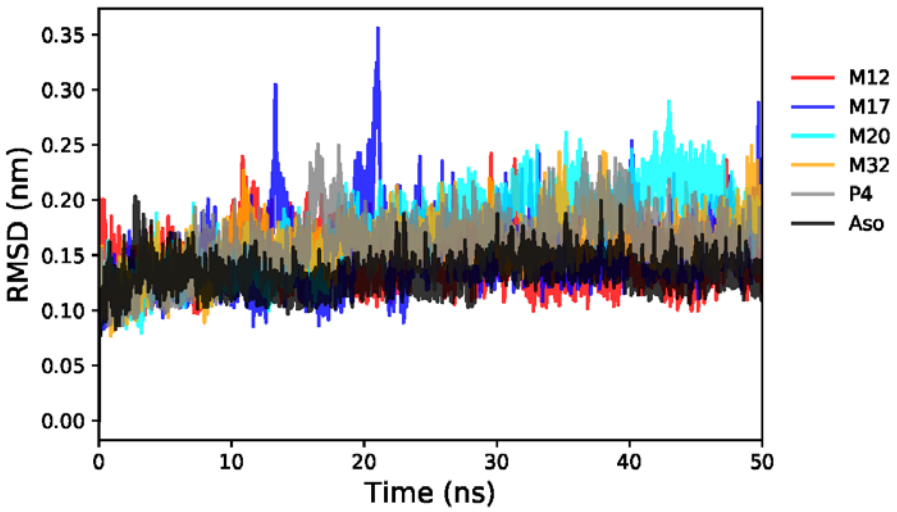


Figure 30. RMSDs of the ligand/LBD complex during the 50 ns simulations.

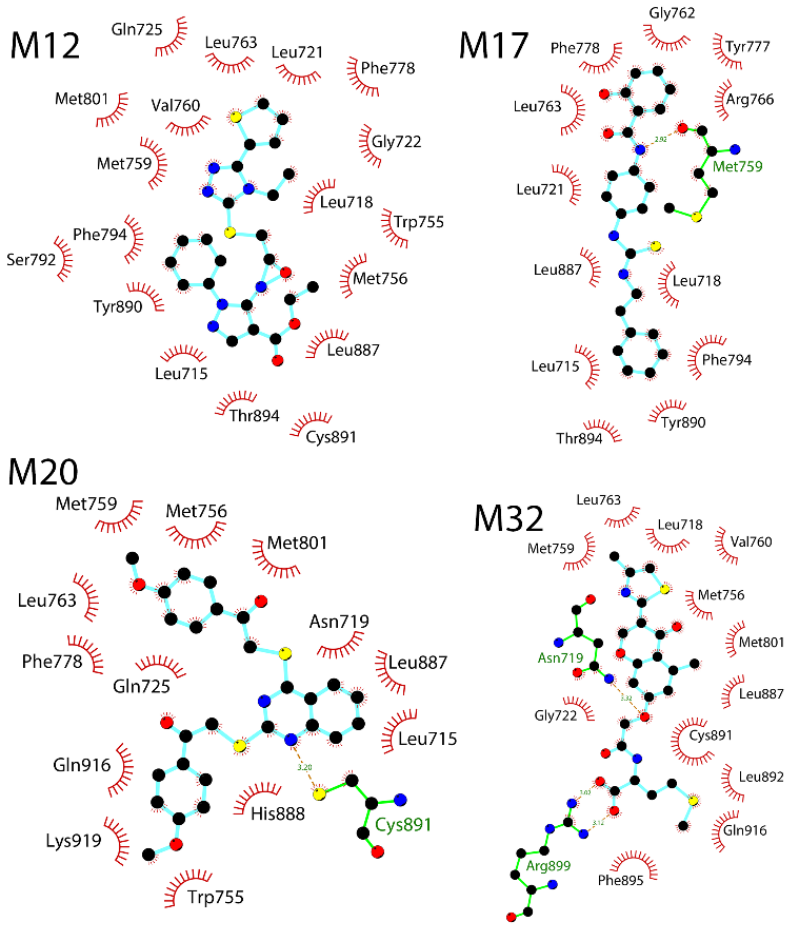


Figure 31. Ligand-LBD interaction patterns.

The carbon, nitrogen, oxygen, and sulfate ligand atoms (or electrostatic interaction involved LBD residues) are in black, blue, red and yellow respectively. The hydrogen bonds or electrostatic

interactions are represented through red dashed lines, whereas the van del Waals interactions are represented using brick red spikes. The complex structures are taken from the last frame of the MD simulation trajectories.

Previous researches [98, 255] have demonstrated that the common-top ranking strategy, incorporated with multiple receptor conformations to include receptor flexibility, is an effect tool in lead discovery. Although receptor flexibility is important in the common-top ranking process, it is not desirable to have too large divergence between the receptor conformations. Since when the search space of a ligand is similar, the performance of the docking pose could be relative stable and easily be selected as good ligands. Meanwhile, the ligand-receptor binding process could obey a population shift theory. The "good" receptor conformations for a strong binder could be sparse in the whole configurational space of a receptor. That means a real active molecule would not have a strong binding affinity with all receptor conformations. Therefore, we could not guarantee that initial receptors are all favorable for an active molecule. Keep this in mind, then it is natural to accept the fact that for a limited number of receptor structures, the high similarity between the conformation would be necessary to select potential lead molecules mainly for these conformations. We would notice the importance of the "scale" of docking. It is always helpful to screen as many as possible molecule libraries with a lot of receptor conformations. But due to the shortage of computational power, we have to narrow down our modeling and docking based on known knowledge, to get two most important information: the binding poses of a ligand and the binding affinity of the ligand/receptor complex.

However, the raw scoring function used in docking simulations are not accurate enough to compare with experiments. Given the fact that more accurate binding affinity prediction method does exist, they are not largely applied in the SBVS, at least not the initial stage of SBVS. Both MMPBSA, absolute binding energy calculate, and ML-based binding affinity methods [256-258] would greatly enhance our ability to estimate the real binding affinity of a receptor/ligand complex. The reason for the void of more accurate estimation of binding affinity is majorly the computation efficiency: the more accurate methods come with dramatically increased resources consumption. So, firstly, we could perform a raw selection using large-scale SBVS to rule out the surely not suitable molecules to narrow down our selections. For the enrichment process, more sophisticated docking simulations, or MD simulations would further kick out the obviously impossible small molecules. The MD simulations of the ligand/receptor complexes are based on force field, which is carefully optimized with and compared to experimental and quantum parameters. It is thus reasonable to believe it would be more reliable than the random sampling and scoring in docking methods. Together with MMPBSA method, and the TML_BP prediction, we could roughly estimate the binding affinity the ligand/receptor complexes. This step could be viewed as a re-score process in addition to the docking score ranking. We believe that this strategy would at least more accurate than the direct docking/scoring based computational drug discovery method.

5.4 Summary

PR, as a member of NR superfamily, involves a lot of important signaling pathways. The value of identifying new PR inhibitors, especially molecules targeting the LBD, lies in the fact that the dysfunction of PR often relates to different types of cancers, and current PR drugs have severe side-effects. Here, in this study, we adopted a VS strategy by considering the receptor flexibility to screen lead-like molecules for PR LBD. Firstly, several rounds of VS have been performed against three PR LBD representative structures (sampled in MD simulations). The molecules with high scores in each round of the VS thus were selected as the so-called "common-top" molecules. Then the first 12 molecules (in complex with the PR LBD), were subjected to MD simulations to assess the stability of the complexes. Additionally, the binding free energies between the molecules with PR LBD were also estimated using the popular MMPBSA method. It is clear that the 12 molecules selected from receptor flexibility enhanced VS were "good" binders to PR LBD due to the stability of the complexes as well as the relatively low binding free energies comparing to the control molecules. The molecule selection workflow applied in this work thus would be also useful for lead discovery for other targets.

Chapter 6. Machine learning based rescoring of PR LBD virtual screening

PR is a key protein in charging female reproduction system stability and development. It involves almost every aspect of female heath. Developing PR drugs are of great interests to scientists. We have already used SBVS, cMD and binding energy prediction combined methods to select several ligands. Continuous study of the PR LBD systems indicates that its highly flexible instinct would hinder the accuracy of our screening and re-scoring strategy. Therefore, we tried to adopt the same VS-Rescoring strategy, but including a different rescoring method, the target specific machine learning based rescoring. Firstly, we obtain active and decoys PR molecules from DUD database and dock all these molecules into PR LBD binding pocket. With the complex structures, decomposing the energetic terms, as well as the contact terms, we could construct a dataset containing the ligand-receptor interaction fingerprints. The dataset, which contains both active ligand and decoys information, could be applied for machine learning based classification models. We believe that these models could greatly decrease the false positive rate and increase the enrichment ability when comparing to the previous methods.

6.1 Methods

Active and decoy ligands were downloaded from DUD database for PR LBD. Given the DUD database, we firstly docked all ligands into LBD pocket, and the docking poses were used to generate binding pattern features. The docking simulations were completed using GOLD with default “slow” mode, defining as 1.2 nm pocket range. The scoring function of the docking simulations was GoldScore.

The dataset thus was processed and trained, to construction an “aggregated” model, which was further utilized in the rescoring of the prediction dataset (Figure 32).

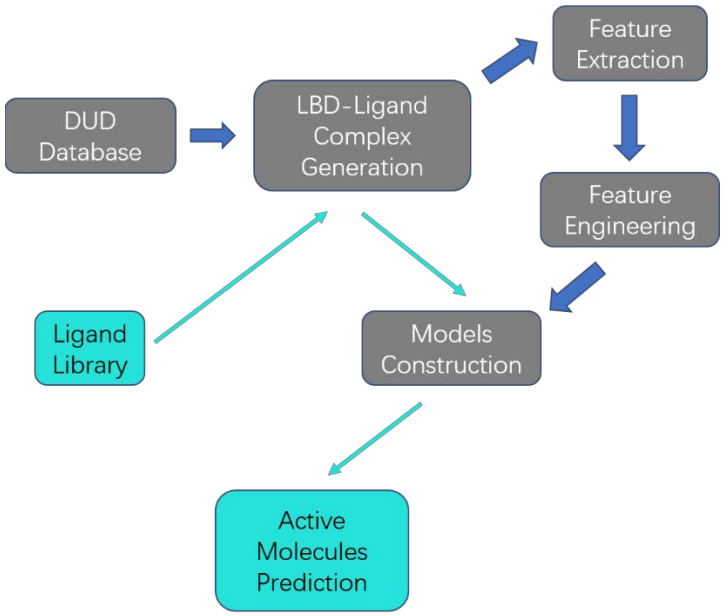


Figure 32. The workflow the machine learning aided rescoring of the docking data.

6.1.1 Docking protocols

The receptor conformation for docking is collected from the largest cluster in previous P4-bound PR LBD cMD simulations. The training and testing ligands are from the active ligands and decoys from DUD database (http://dud.docking.org/r2/) for PR. The predicting ligands are from a combined dataset we described in section 5.4.1. The procedures of docking the ligands in libraries to PR LBD ligand binding pocket using GOLD are adopted from those used in 5.4 Methods part.

The binding poses are collected only if the GOLD docking scores are higher than 50. The ligand-LBD complexes for all ligands whose GOLD score higher than 50 are then prepared. In total, we have 391 active ligand-LBD complexes and 6982 decoys. The relatively large number of decoys comparing to active molecules thus mimic the real situation that active molecules are only a very small portion of the large screening library. A large screening library (see section 5.3.1) containing 147489 molecules was docked to PR LBD (the largest cluster representative structure in LBD/P4 complex simulation). After removing the ligands with a lower than 50 GoldScore, we obtained 87334 ligands. The ligand poses were then combined into the PR LBD binding pocket. The complex structures were then used for binding feature generation and subsequent prediction.

6.1.2 Protein-ligand interaction dataset preparation

For each ligand in either DUD database and predicting library, the ligand-receptor complex is utilized to generate 3-dimensional interaction features, considering the per-residue and per-atom type interaction energies (van del Waals and columbic energies) and contacts.

A total number of contacts between a residue sidechain or mainchain and the ligand is calculated. For residue *i*, the sidechain or mainchain contact number C_i is defined as below:

$$C_i = \sum_s \sum_t c_{s,t} \quad (44)$$

$$c_{s,t} = \begin{cases} 1, & d_{s,t} \le d_{cutoff} \\ 0, & d_{s,t} > d_{cutoff} \end{cases} \quad (45)$$

where *s* and *t* are atoms in residue *i* and the ligand respectively, and the d_{cutoff} = 0.5 nm is the distance cutoff.

The van del Waals interactions between each LBD residues (250 in total) sidechain and mainchain with the ligand are quantified according to following equations 46-50:

$$E_{vdw,i} = \sum_{N_i}^{k=1} \left(\frac{C12}{d_k^{12}} - \frac{C6}{d_k^6} \right) \quad (46)$$

$$C12 = 4\xi_{s,t} \cdot {\delta_{s,t}}^{12} \quad (47)$$

$$C6 = 4\xi_{s,t} \cdot {\delta_{s,t}}^{6} \quad (48)$$

$$\xi_{S,t} = \sqrt{\xi_s \cdot \xi_t} \quad (49)$$

$$\delta_{s,t} = \frac{1}{2}(\delta_s + \delta_t);\ \xi_{s,t} = \sqrt{\xi_s \cdot \xi_t} \quad (50)$$

C6 and *C12* are atom-type specific constants, which are further defined by $\xi_{s,t}$ and $\sigma_{s,t}$, while *s* is an atom from an LBD residue sidechain or mainchain, *t* is an atom from the ligand. The combination rules from atom specific ξ and σ parameters are adopted from the amber99SB force field. N_i is combination pairs of atoms between a residue *i* sidechain (or mainchain) and the ligand. Thus, there are 500 van del Waals interaction features.

$$V_i = \sum_{N_i} f(q_s q_t) \,/\, (\varepsilon d_{s,t}) \quad (51)$$

$$f = \frac{1}{4\pi\varepsilon} \quad (52)$$

The summed columbic interactions between residue *i* sidechain (or mainchain) and the ligand. q_s and q_t are the atomic charges of atom *s* from residue *i* either sidechain or mainchain and atom *t* from the ligand. The atomic charges of the receptor atoms and ligand atoms were determined using OpenBabel [259]. $d_{s,t}$ is the distance between atom *s* and atom *t*, and ε is the dielectric constant. N_i is combination pairs of atoms between residue *i* sidechain (or mainchain) and the ligand. In this study, we had *f* = 138.935485. Therefore, there are 500 features for electrostatic interactions.

The contacts between atoms element types are also considered. A list of chemical elements E_L was considered, they are: C, O, H, N, S, P, Cl, I, F, Br, IB and DU, which represents any other elements not listed here.

$$E_L = [C, N, O, H, P, S, Cl, F, I, Br, IB, DU] \qquad (53)$$

The contacts between residues' sidechain or mainchain against the ligand were counted only considering the above element types.

For residue *i*, the sidechain or mainchain based element type contacts $EC_{i,\,Tk,\,Tj}$ is defined as following:

$$EC_{i,T_k,T_j} = \sum_k \sum_j c_{k,j}\,,\ if\ T_k, T_j \in E_L \qquad (54)$$

$$c_{k,j} = \begin{cases} 1, & d_{k,j} \le d_{cutoff} \\ 0, & d_{k,j} > d_{cutoff} \end{cases} \qquad (55)$$

For element type T_k (of LBD) and element type T_j (of the ligand), the element type pair contacts thus could be calculated by summing all the possible contacts if atom *k* (in LBD) and atom *j* (in ligand) are element types T_k and T_j respectively. Therefore, there were 121 features for element type contacts.

In total, we have 1650 features and 7373 samples. The feature generation calculates were completed using home-made python package dockingML, which could be found here: https://github.com/zhenglz/dockingML.

6.1.3 Feature engineering

Before we performed machine learning classifications, feature engineering was performed to extract informative features. In real-world dataset, there are a lot of features are text data, or categorical data, both of which are no numeric features and are not suitable for direct usage. Besides, noise and redundancy are also the common pitfalls in data analysis. And in order to obtain the weights of a model, it is generally the practice to normalize the dataset to ensure smaller weights. Here, in our dataset, all features are numeric vectors, and redundancy could exist. Therefore, we adopted the following procedures to remove noise, remove redundancy and reduce feature numbers. The all-zero features were removed firstly. Some residues in PR LBD are quite far away from the binding pocket, thus the van del Waals interactions, contacts, and columbic interactions are treated as zero values. And some element types could not be found in LBD or the ligand, thus the contacts are also void. Column-based scaling of the dataset was performed using sci-kit learn python module. The normalized dataset thus has a mean of 0.0 and a standard deviation as 1.0. PCA and isomap embedding were used to reduce the dimensions of the preprocessed dataset in the last step. The python sci-kit learn module was used to perform the PCA and isomap feature dimension reduction. Before the calculations, dataset normalization was carried out. In this study, we constructed two datasets: the

engineered dataset and the original dataset. The original dataset includes the original features, but with all features normalized. The original dataset thus was processed with the above procedures and the resulting dataset was named as the engineered dataset.

6.1.4 Machine learning model training

The processed dataset now contains 7373 samples and 50 features. It was randomly split into two subsets, a training-testing subset, and a validation subset, according to an 8:2 ratio. The training-testing subset was used for 5-fold cross validations using SVM, k-NN, DT, NB and MLP methods.

The relatively large size of the negative samples would bias the models more towards the negative side. In order to avoid the dataset imbalance, we further divided the negative samples in training set into n non-overlapping small subsets randomly. Each of the negative subsets was combined with the positive samples in the training set. The commonly shared positive samples thus would be treated more importantly than the learning model using the whole imbalance training set. Therefore, we constructed n subsets of training samples. For each subset, we applied 5-fold cross validation learning models to gain knowledge from the subset. For the training part, the average scores of all the n*5 models would be considered as the performance of the ensemble model. In the prediction mode, the majority voting of the n*5 models would be the final classification result. For SVM, multiple kernels were tested and compared. The most accurate kernel was selected for following studies. finally, "*rbf*" kernel was used. And the hyperparameters, such as C and gamma, were optimized using sklearn GridSearch module. The final parameters for C and gamma were selected if the combination of C and gamma hyperparameters offer the highest prediction score (accuracy) (Figure 33A). Thus, we used C = 100 gamma = 0.01 for the SVM models training.

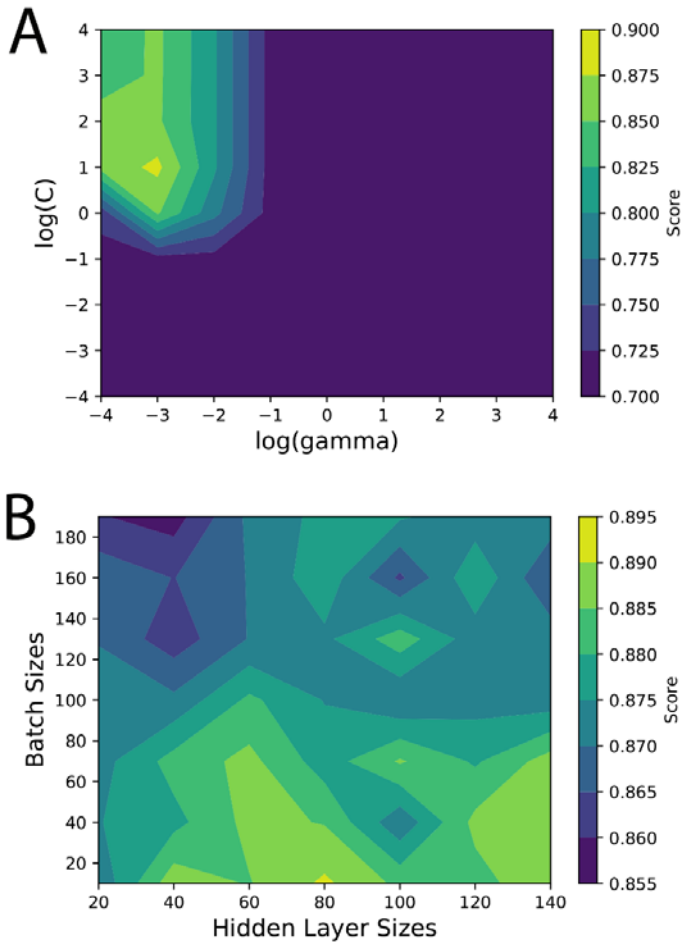


Figure 33. The grid search of hyperparameters of the 5-fold SVM and MLP models.

A, the prediction accuracy of the different hyperparameter combinations of the 5-fold SVM model with *rbf* kernel. B, the prediction accuracy of the different hyperparameters (hidden lay sizes and batch sizes)

MLP classifiers were trained using sklearn neural_network module, with a stochastic-gradient-decent based method for weights optimization. Hidden layer sizes and batch sizes were carefully screened using the grid-search method. The other parameters were set as default in sklearn neural_network MLPClassifier model (Figure 33B). The best combination is batch size = 10 and hidden layer size = 80.

6.1.5 Model evaluation and ligand prediction

The accuracy, specificity, and MCC values were evaluated. The above parameters could be determined from a confusion matrix. A confusion matrix (Table 16) is composed of 4 parts, true positive (*TP*), true negative (*TN*), false positive (*FP*) and false negative (*FN*). For a general binary classification model, the 4 values (*TP*, *TN*, *FP* and *FN*) could be counted based on frequencies of the sample cases.

Table 9. The confusion matrix

TP	*TN*
FP	*FN*

Sensitivity (SEN) is used to evaluate the true positive rate, thus it is sometimes also called recall. The definition of the SEN could be expressed as following:

$$SEN = \frac{TP}{TP+FN} \qquad (56)$$

Accuracy (*ACC*) of a model could be expressed as the truly classified sample size ratio with respect to total sample size:

$$ACC = \frac{TP+TN}{TP+TN+FP+FN} \qquad (57)$$

Specificity (*SPC*) is the true negative rate, expressed as the following:

$$SPC = \frac{TN}{TN+FP} \qquad (58)$$

MCC is a balanced factor to evaluate the quality of the classification model:

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \qquad (59)$$

F1 Scores were calculated to have an index of balanced recall and precision estimation. F1 score is defined as following equation:

$$F1 = 2 \cdot \frac{TP}{(2TP+FP+FN)} \qquad (60)$$

ROC and AUC were also computed to access the ability of the models' ability to classify active molecules from the inactive ones. Enrichment factor (*EF*), which was defined in equation 37, was also calculated given the top 1%, 5%, 10%, and 20% ligands.

For the unseen ligands, the ligand/LBD interaction fingerprint-based predictions were performed using the models generated above. Multiple randomized models were adopted to ensure a low false positive prediction. The all-pass strategy was used: we generated a pool of models, if a ligand is predicted to be an active ligand by all models in the pool, then the ligand would be trusted as an active molecule. Later, we used SwissAdme (http://www.swissadme.ch) web server to predict the ADME properties of the active molecules and filter these molecules by solubility and lead likeness. Only if a molecule does not have lead likeness violations and has rather a high solubility, would be finally selected. We obtained 335 molecules as active ones, whereas with the SwissAdme server, we successfully predicted ADME properties for 324 molecules. After the filtering process, 21 molecules

are identified as potential lead like active molecules. The ligand/LBD interactions are visualized using LigPlus.

6.2 Results

6.2.1 Decomposed binding interaction fingerprints feature space is not linearly separable

From the PCA analysis of the original high dimensional feature space, the active and decoy molecules could not be clearly separated based on the first two principal components. Though, for the 2nd PC (PC2), active molecules are more aggregated than the decoys. It is not surprising to have this observation. Since we hypothesized that the active molecules adopt similar binding patterns and binding spaces, whereas the decoys were randomly docked into the binding pocket, and their binding within the receptor pocket may not be physical and biologically meaningful. And the large chemical space of the diverse decoys could also result in the diverse instinct of the dataset. Given that PCA applying a linear transformation of the original features, we initially thought that non-linear embedding of the original feature space would be more effective. We used the manifold learning method isometric map embedding (isomap) to project the dataset into a two-dimensional space, however, it turned out that it is still impossible to separate the two classes according to the new lower dimensional dataset (Figure 34).

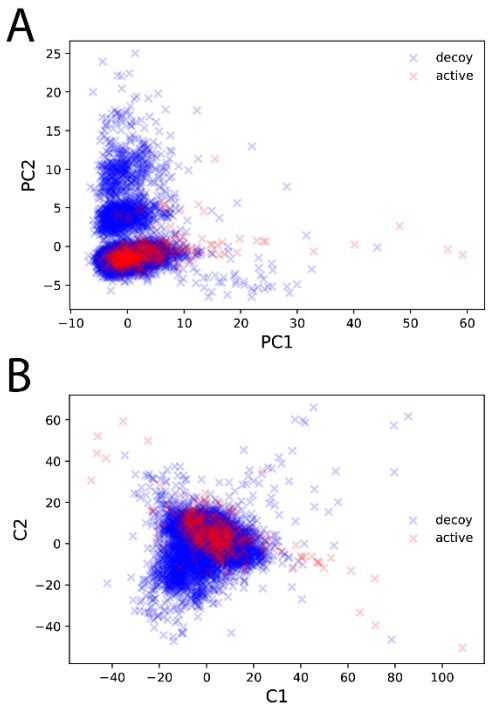


Figure 34. PCA and Isomap transformations of the preprocessed DUD-PR dataset.

Although it is not doable to classify the active ligands from the large quantity of the decoys, it is still valuable to have a low-dimensional, information enriched and low-noise dataset for further machine learning based classifications.

6.2.2 Classification performances of individual models

SVM model with "*rbf*" kernel was performed with a grid search for *C* and *gamma* hyperparameters. The optimal *C* and *gamma* parameters were determined according to the prediction power of the model, the accuracy. The 5-fold CV SVM models have quite a good AUC (>0.94). The original normalized dataset outperforms the engineered dataset, and by using it, higher AUC, MCC, and EF values were obtained (Table 10, 11 and 12). With the engineered dataset, the 5-fold models could achieve an average AUC around 0.944, meanwhile, they could enrich the active molecules from a large number of decoys and achieve an average EF (10%) = 7.912. However, training with the original dataset, the SVM models could have an average AUC around 0.966 and an average EF (10%) = 8.579. The 5-fold CV MLP models did not perform better results than the 5-fold CV SVM models. However, still, the models achieve better scores using original dataset than the engineered dataset. With the original dataset, the MLP models could achieve a 0.960 AUC, 0.672 MCC score, and 8.579 EF (10%) score. The enrichment power of both the MLP models and SVM models is much higher than the GoldScore EF (10%) = 1.081 and Vina Score EF (10%) = 1.865. Not only the EF (10%), the SVM models and MLP models achieve higher scores for all EF at a different level than Vina Score and GoldScore (Table 13).

From the ROC curves, it is quite clear that the SVM and MLP models could quite successfully maintain a high true positive rate, meanwhile, they could retain a low false positive rate. For the testing dataset, all models could achieve rather a high AUC (>94%). Especially, the SVM models training with the original dataset, have well converged ROC curves and quite high AUC values, indicating that the 5-fold SVM models with original dataset are the best models for PR LBD true binder classification. Besides, our models show excellent enrichment ability and extreme low false positive rate.

In virtual screening, the necessity of finding true positive molecules is quite urgent. However, current docking scoring functions bring a large amount of false positive results. In this study, we hope to lower down the false positive rate α. It is worth noting that the sensitivities (1-β) of the models are not very satisfactory, though the AUC and EF are maximized. The sensitivities of the models are around 0.5, thus means that the type II error β is around 0.5. The type II error β represents the false negative rate. It thus indicates that the false negative molecules occupy a very large portion when comparing to true positive molecules. While the specificity, which is defined as the true positive rate, represents the type I error and the false positive rate α = 1 – specificity. Especially, we want to

minimize the α, which is equivalent to maximize the specificity. The average α values obtained are 0.7% and 1.3% for SVM models and MLP models respectively (Table 10 and 11).

Table 10. Prediction power of the 5-fold SVM models.

Dataset	K-Fold	Sensitivity	Specificity	Accuracy	F1 score	MCC	AUC
Engineered Dataset	1	0.492	0.986	0.959	0.564	0.549	0.959
	2	0.465	0.989	0.969	0.526	0.516	0.946
	3	0.438	0.993	0.955	0.569	0.577	0.944
	4	0.448	0.997	0.966	0.600	0.625	0.930
	5	0.466	0.996	0.970	0.607	0.624	0.942
	Aver.	0.462	0.992	0.964	0.573	0.578	0.944
Original Dataset	1	0.619	0.990	0.970	0.690	0.680	0.969
	2	0.628	0.988	0.975	0.643	0.630	0.962
	3	0.600	0.993	0.966	0.706	0.701	0.958
	4	0.612	0.998	0.976	0.745	0.754	0.967
	5	0.586	0.996	0.975	0.701	0.704	0.972
	Aver.	0.609	0.993	0.973	0.697	0.694	0.966

Table 11. Prediction power of the 5-fold MLP models.

Dataset	K-Fold	Sensitivity	Specificity	Accuracy	F1 score	MCC	AUC
Engineered Dataset	1	0.508	0.986	0.960	0.577	0.562	0.950
	2	0.581	0.990	0.975	0.633	0.623	0.954
	3	0.563	0.987	0.958	0.647	0.634	0.955
	4	0.507	0.994	0.966	0.630	0.633	0.934
	5	0.569	0.995	0.974	0.680	0.682	0.938
	Aver.	0.546	0.990	0.967	0.633	0.627	0.946
Original Dataset	1	0.667	0.988	0.971	0.712	0.699	0.954
	2	0.721	0.984	0.975	0.674	0.662	0.962
	3	0.613	0.985	0.960	0.676	0.659	0.959
	4	0.582	0.988	0.965	0.655	0.643	0.969
	5	0.672	0.988	0.973	0.709	0.696	0.957
	Aver.	0.651	0.987	0.969	0.685	0.672	0.960

Table 12. EF of the 5-fold SVM, MLP models and GoldScore

Dataset	K-fold	5-fold SVM				5-fold MLP			
		0.001	0.050	0.100	0.200	0.001	0.050	0.100	0.200
Engineered Dataset	1	17.460	12.063	8.413	4.444	15.873	11.746	8.571	4.524
	2	23.256	12.558	7.907	4.651	23.256	12.558	8.372	4.651
	3	13.750	10.500	7.250	4.375	12.500	11.250	7.875	4.563
	4	16.418	11.343	8.060	4.179	16.418	11.642	7.761	4.328
	5	18.966	13.448	7.931	4.310	18.966	13.793	7.586	4.483

	Aver.	17.970	11.983	7.912	4.392	17.402	12.198	8.033	4.510
Original Dataset	1	17.460	14.286	8.571	4.603	15.873	13.651	8.413	4.603
	2	25.581	13.023	8.372	4.651	25.581	14.884	9.070	4.651
	3	13.750	12.250	8.375	4.625	12.500	12.000	8.125	4.625
	4	16.418	14.030	8.955	4.776	14.925	12.836	8.657	4.776
	5	18.966	14.483	8.621	4.828	18.966	14.138	8.448	4.483
	Aver.	18.435	13.614	8.579	4.697	17.569	13.502	8.542	4.628

Table 13. Enrichment factors of vina scores and GoldScore.

Enrichment	0.01	0.05	0.1	0.2
Vina Score	5.169	2.202	1.865	1.393
GoldScore	0.450	1.081	1.171	1.216

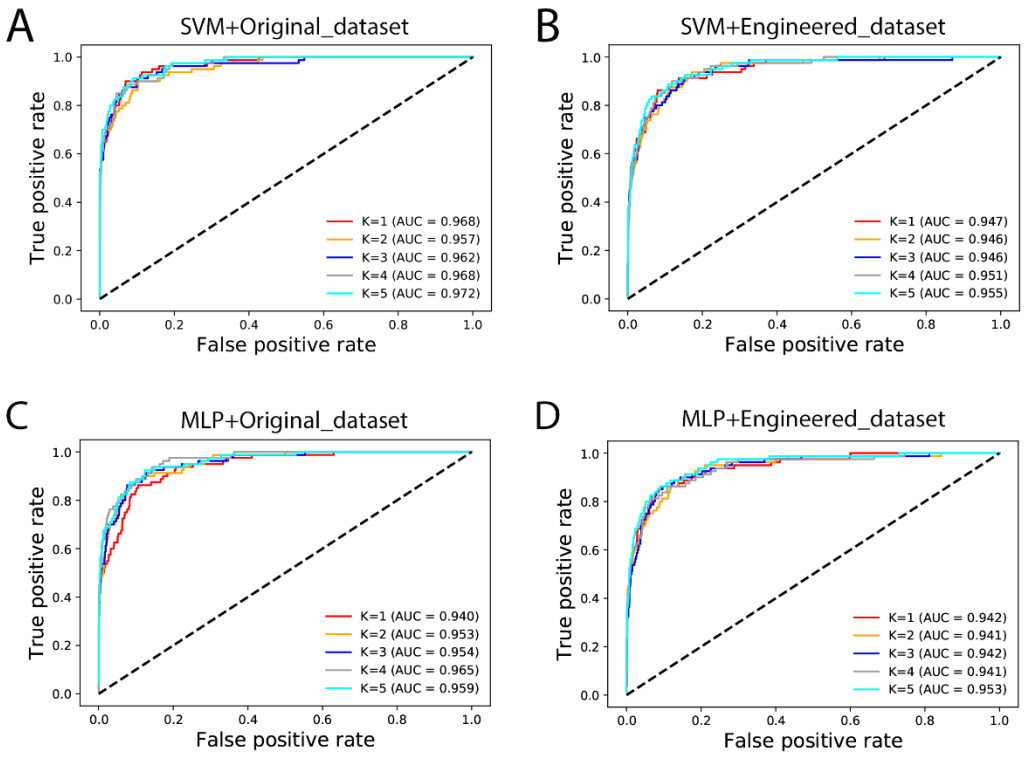


Figure 35. The ROC plots of the SVM and MLP models.

A, the 5-fold CV of the SVM model with the original dataset. B, the 5-fold CV of the SVM model with the engineered dataset. C, the 5-fold CV of the MLP model with the original dataset. D, the 5-fold CV of the MLP model with the engineered dataset.

6.2.3 Predicted PR LBD true binders

We hypothesize that when we incorporate multiple rounds of random and parallel 5-fold SVM or MLP models and assess a molecule as a true binder only if the majority of the models give a positive voting to this molecule. The random split of the training set in 5-fold CV models could enable all cases being treated as training data. The ensemble-models based majority voting, as indicated by other studies [reference], would significantly improve the performances than single models. However, in our study, we found that the inclusion of multiple rounds of SVM or MLP predictions does not significantly improve the AUC, nor the specificity (Table 14). Nevertheless, we applied the 20 repeats of the 5-fold CV SVM models and MLP models to achieve high specificity in on hand and the highest AUC and EF at the other hand. The 20 rounds 5-fold CV SVM models consider 779 molecules as active molecules, while the 20 rounds 5-fold MLP models provide a list of 573 molecules as positive ligands. There are 335 ligands are classified as active molecules with both the SVM and the MLP models. The GoldSocre of the ligands, when docked to PR LBD, are distributed mostly around 50, where the GoldScores for P4 and asoprisnil are 54.25 and 60.08. Around 80% of the ML method predicted active molecules have lower GoldScore than asoprisnil, around 35% of these ligands have lower GoldScore than P4.

Table 14. The prediction power of majority voting of multiple rounds of 5-fold CV SVM models

Round (*N*)	Sensitivit	Specificit	Accurac	F1-Score	MCC	EF	AUC
1	0.6000	0.9957	0.9742	0.7164	0.718	1.375	0.8282
2	0.5875	0.9957	0.9736	0.7068	0.709	7.250	0.8529
5	0.6000	0.9957	0.9742	0.7164	0.718	7.250	0.8521
10	0.5875	0.9964	0.9742	0.7121	0.717	7.250	0.8513
20	0.5875	0.9957	0.9736	0.7068	0.709	7.750	0.8757
50	0.6000	0.9957	0.9742	0.7164	0.718	7.750	0.8750

We rank the intersection molecules predicted through ADMET predictions. The molecules with high drug-likeness, low toxicity, and high solubility were considered. We select 21 top ranking molecules with optimal ADMET scores (Figure 37).

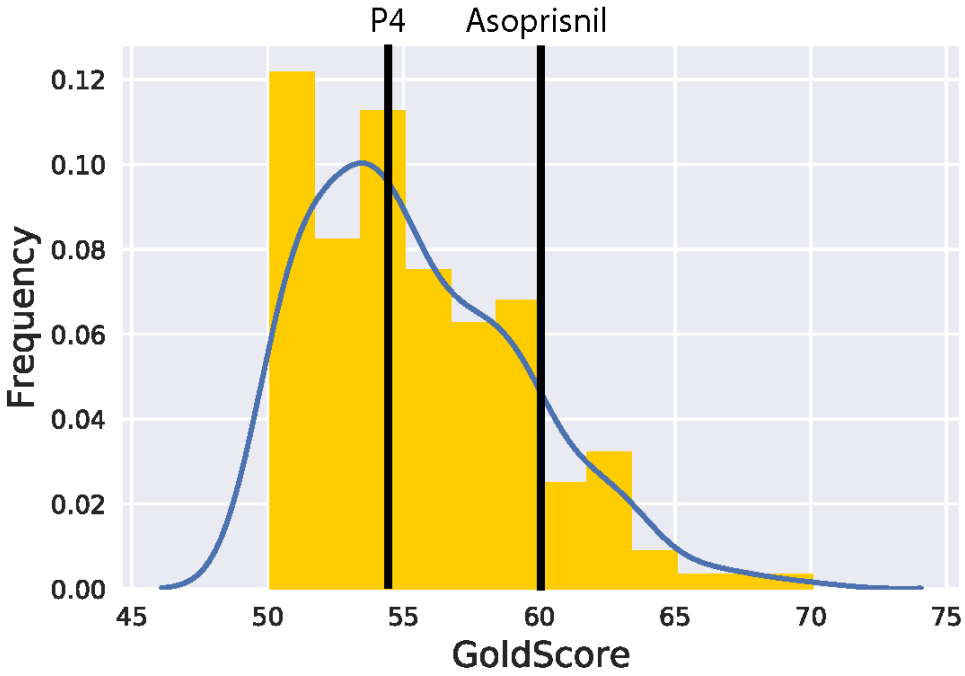


Figure 36. The GoldScore distributions of the ML selected 335 active molecules when they are docked into PR LBD.

The Goldscores of P4 and asoprisnil, when docked to PR LBD, are marked by black vertical lines.

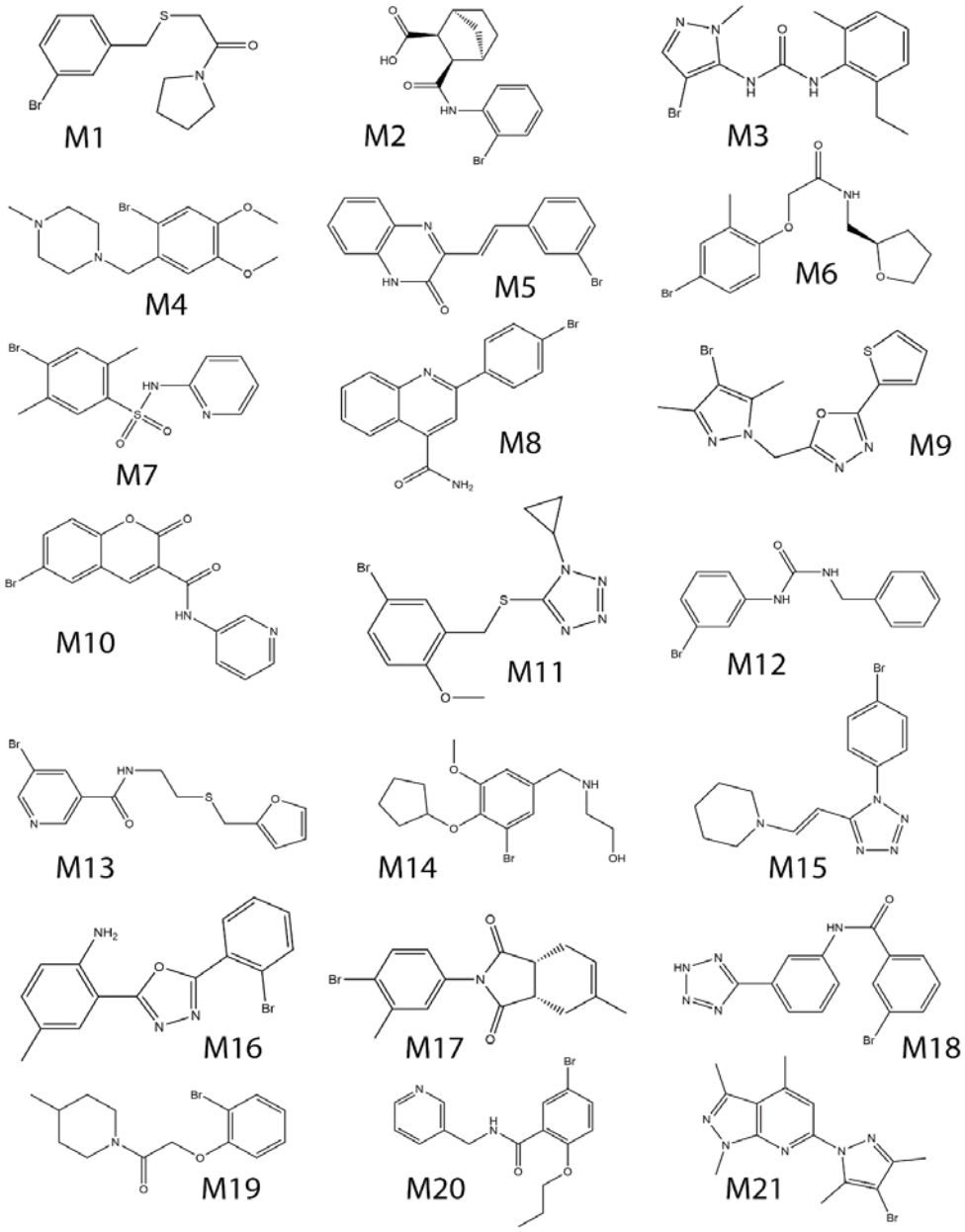


Figure 37. The key active molecules selected by ML models, as well as lead likeness and solubility filtering.

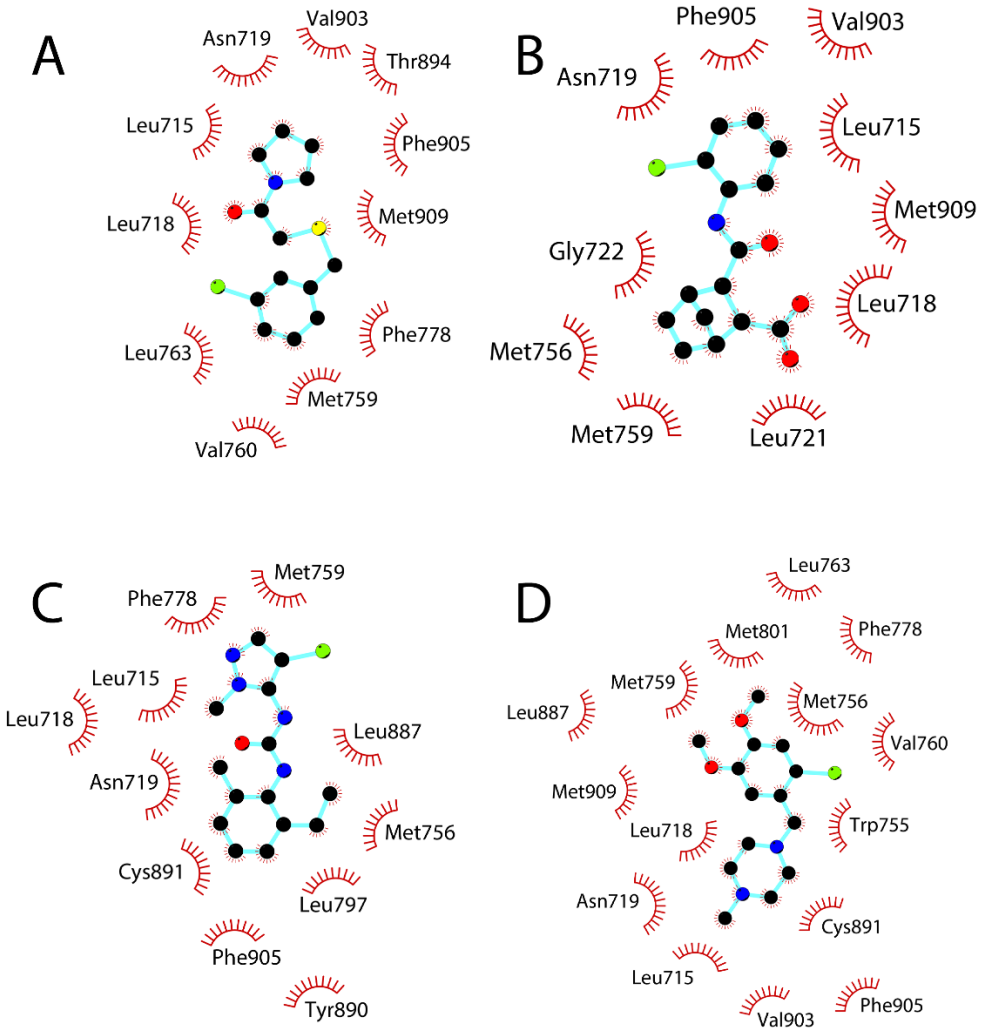


Figure 38. Interactions between the first 4 selected ligands and the LBD.

From A to D, the 4 molecules are M1 to M4.

The interactions between the LBD and the first 4 high potential PR ligands are presented in Figure 38. It is quite clear that the hydrophobic interactions dominate the binding contacts. Leu715, Leu718, and Asn719, which locate in helix 3, are all preserved in the four complexes. Met909 lies in the binding interface for three ligands (M1, M2, and M4). This Met909, which locates in *N*-terminal helix 12, plays a key role in agonistic conformation stability. Met759 also shows up in all the 4 ligand/LBD complex interfaces. This residue also involves hydrophobic interactions in P4/LBD complex as well as asoprisnil/LBD complex. As we observed in crystal P4/LBD and asoprisnil/LBD complexes, Arg766 has actively involved hydrogen bonds with P4 or asoprisnil. However, we noticed that these 4 ligands do not form hydrogen bonds with the LBD. It is not clear where hydrogen bonds would be formed after complex equilibration in water using MD simulations or within a realistic environment. Thus, we believe that although ML methods would provide reliable results, the experimental binding

affinity measurement, as well as cellular agonistic or antagonistic effects, would be required to continually discover the true PR LBD binders.

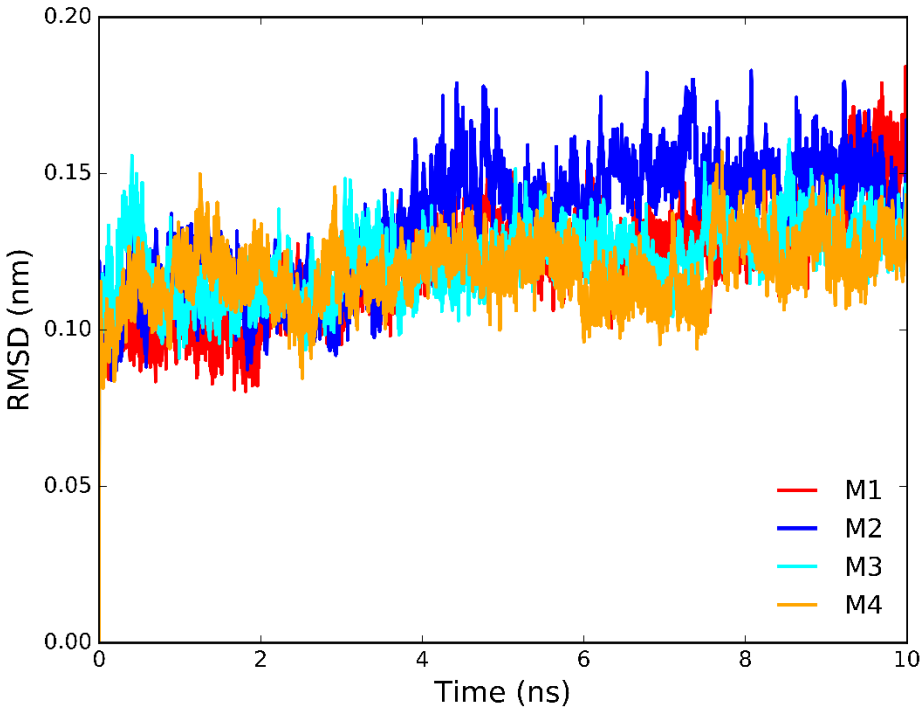


Figure 39. The αC RMSD of PR LBD during the ligand-LBD complex short simulations.

6.3 Discussions

6.3.1 GOLD captures the "right" poses, but "wrong" affinities

The initiative of this ML aided rescoring of the docking poses lies in the fact that GOLD could successfully recover the native docking poses, however, performs badly in binding strength prediction using GoldScore. When the co-crystallized ligands were mixed with FDA approved drugs to generate a test library, GOLD could perfectly generate native-like poses. For most co-crystalized ligands, the docked pose is well aligned with its native conformations (Table 15). For example, the RMSD between P4 docked pose and the native pose is 0.422 Å, and it is 1.506 Å for mifepristone. However, the binding affinities are not well correlated with their GoldScore. For K_i values, the Pearson correlation is around 0.301, while it is around 0.29 for IC50. Meanwhile, some native ligands are badly ranked when they are re-docked into their original receptor conformations. For example, A2K is docked into the PR LBD conformation (4APU, chain B) with a reasonable RMSD (=2.013 Å), however, its GoldScore is quite low and its ranking lies in the last 25% of the GoldScore list. In 6 out of the 21 trials, the ligand's GoldScore is not ranked in the top 10% of the result GoldScore list.

It is quite clear that GoldScore is not strong enough to predict the true binding affinity with acceptable accuracy. So, we come to the conclusion that GOLD could recover the native binding modes, but not the accurate binding strength, at least in PR LBD dockings.

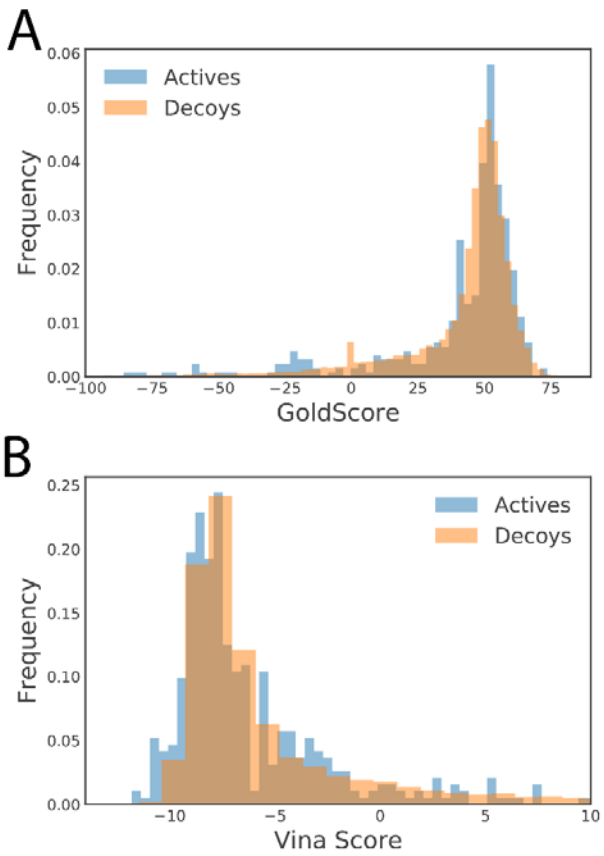


Figure 40. The docking score distributions of active and decoy molecules.

A, the distribution of the GoldScore for both active molecules (blue) and decoys (orange). B, the distribution of the Vina scores for both active molecules (blue) and decoys (orange). The docking scores using GoldScore SF are more positive for strong binders, while those are more negative for AutoDock Vina Score.

Not only GoldScore, but the AutoDock Vina scores also are not sufficient to differentiate the active molecules from large amount decoys. The active molecules have both positive and negative values using Gold docking method and AutoDock Vina. When we perform docking with GOLD, the highest frequent score for active small molecules is 52, while the score is around 50 for the DUD decoys (Figure 40A). Similarly, the most populous docking score for active and decoy molecules is around 7.5 kCal/mol using AutoDock Vina (Figure 40B). The finding here indicates the inefficiency of both GOLD and AutoDock Vina for binding affinity predictions of the ligand-receptor complexes.

Based on this finding, we further consider the possibility of using the "good" poses predicted using GOLD, and apply a more precise scoring function, or a rescoring method, to identify the true binders from the vast number of small molecules.

Table 15. Performance of the redocking co-crystalized ligands into crystal PR LBD conformations

PDB ID/Chain*	Ligands Code	Ligand Name	RMSD (Å)	GoldScore	Ki (nM) **	IC50 (nM)**	Kd (nM) **	Ranking (out of 3280)
1a28_A	STR	P4	0.422	65.87	5.1	0.2-17	NA	71
1e3k_A	R18	R1881	1.062	62.53	0.5	NA	NA	263
1sqn_A	NDR	NA	0.905	58.92	1.9	3.2	0.4	437
1sr7_A	MOF	Mometasone Furoate	0.91	88.43	NA	NA	0.1	1
1zuc_A	T98	Tanaproget	0.783	64.95	NA	0.5	NA	132
2ovh_A	AS0	Asoprisnil	0.755	76.93	NA	1	NA	36
2ovm_A	AS0	Asoprisnil	0.799	74.59	NA	NA	NA	52
2w8y_A	486	Mifepristone	1.506	50.6	0.82	0-10	NA	1441
2w8y_B	NDR	NA	0.879	58.51	1.2	3.2	0.4	489
3d90_A	NOG	Norgestrel	1.058	60.76	NA	NA	NA	323
3g8o_A	30X	NA	3.408	54.96	NA	125	NA	839
3hq5_A	GKK	NA	2.181	79.7	NA	16	NA	7
3kba_A	WOW	NA	5.064	70.03	NA	16	NA	105
3zr7_A	OR8	NA	2.418	69.43	NA	NA	NA	28
3zra_A	ORB	NA	3.319	68.86	NA	NA	NA	56
3zrb_A	OR8	NA	1.94	71.37	NA	NA	NA	19
3zrb_B	ORC	NA	0.958	68.79	NA	NA	NA	49
4a2j_A	AS0	Asoprisnil	0.818	81.89	NA	NA	NA	13
4apu_A	OR8	NA	1.814	66.33	NA	0.6	NA	78
4apu_B	A2K	NA	2.013	38.25	NA	NA	NA	2471
4oar_A	2s0	ulipristal acetate	2.175	59.67	NA	0.2	NA	448

* if there are homodimer chains in a PDB file, only the first chain (with both PR LBD and the ligand) was selected, otherwise both chains would be adopted.

** the data were collected from BindingDB and PDBbind. Only the strongest binding affinities were kept if multiple values exist.

6.3.2 Feature importance and feature engineering

The interaction features are high dimensional data, which could not be easily visualized and understood by humans. In the original dataset, there 1650 features in total. Incorporating all the features would gain more information than engineered features, thus the ML models trained from original dataset would harness more power to draw precise decisions (Table 9, 10 and 11). However, it would be relatively more time consuming and high probability to have over-fitting and noises, if using the original dataset in practice. Thus, we employed a feature selection method which

incorporating decision tree-based feature importance calculation, feature Pearson correlation analysis and a PCA based feature reduction to further decrease the noise in the dataset. Finally, we obtained a clean dataset with 274 features, which is much smaller than the original dataset. We didn't observe the performance increase after the feature cleaning. Instead, the engineered dataset is smaller in size, meanwhile, it loses some part of the information, thus leading to less ability of classification performance (Table 16). In the situation that computational power is not limited, it is unnecessary to process the original dataset to save training time.

Table 16 The performance of clusters of features using planar (n=7) SVM model

SN	Feature Clusters	Sensitivity	Specificity	Accuracy	F1 Score	MCC
1	atomTCount	0.605	0.950	0.853	0.695	0.616
2	columbback	0.000	1.000	0.717	0.000	NA
3	columbside	0.534	0.872	0.778	0.571	0.426
4	countback	0.619	0.919	0.833	0.677	0.573
5	countside	0.623	0.925	0.840	0.685	0.586
6	vdwside	0.602	0.920	0.829	0.666	0.561
7	vdwback	0.590	0.913	0.822	0.650	0.540

Since the binding pattern features could be grouped into feature clusters based on the way how they are generated. From table 16, the feature cluster 1 (atom-type based contact number) shows the best performance among all feature clusters, with the highest MCC score and F1 score, which indicates that these features show higher values in true binder classifications. The backbone and sidechain atom contact counting parameters show relatively stronger prediction power than the vdw and columbic decomposed energies. This may not necessarily reflect the fact that the energy decomposition is not realistic or accurate. The linear combination of the energy terms could be the reason.

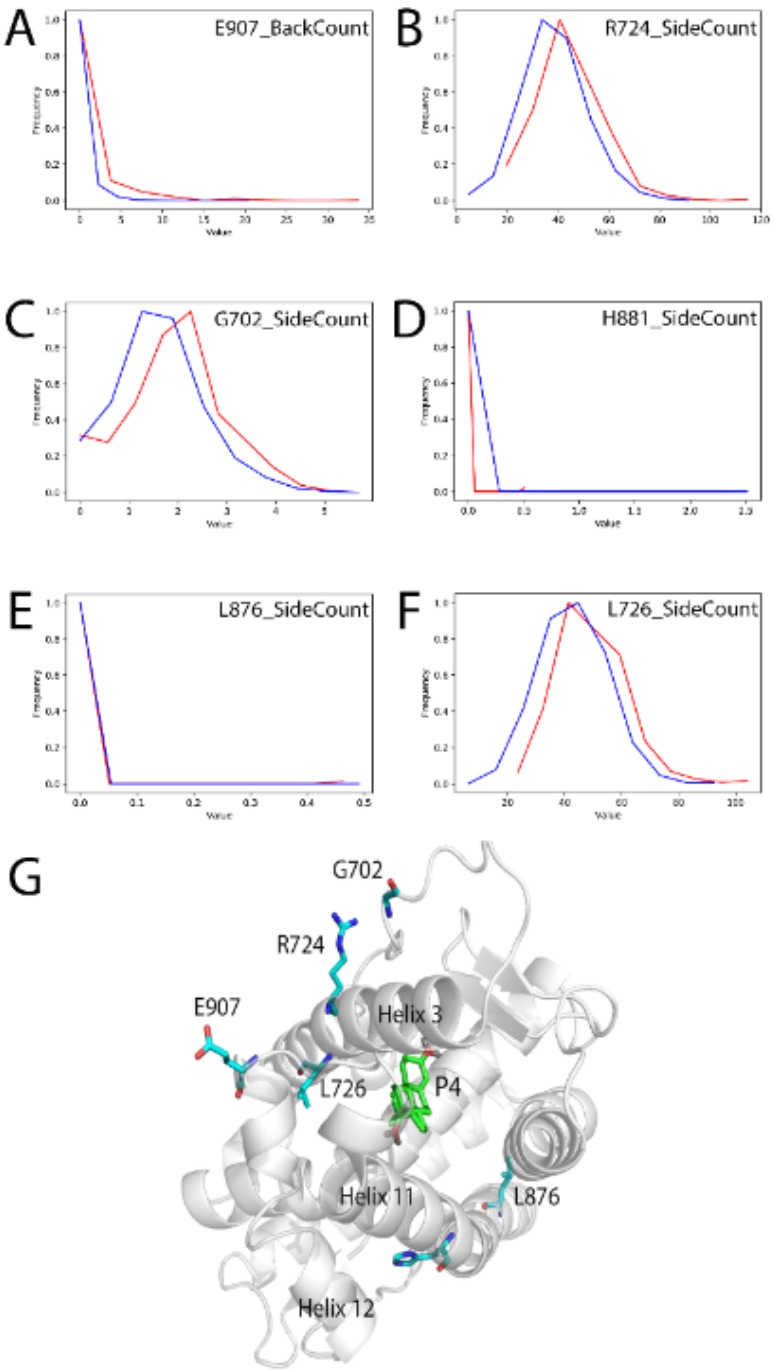


Figure 41. The distribution of the first 6 highest importance score features and the positions of the related residues.

From A-F, the active (red) and decoy (blue) data points share quite different distributions. G, the relative positions of the key features related residues are presented as sticks with carbon colored as cyan, while P4 carbon atoms are in green color and two lone pair electrons in oxygen atoms are shown as grey. The complex structure is from the GOLD docking results of P4-LBD docking.

Using DT based feature scoring model, the most important features have been identified. The distribution of the first 6 highest-importance-score features (Figure 41) were compared between active ligands and decoy molecules. For Glu907, Arg724, Gly702, and Lue726, the contacts in the active-LBD complex are relatively higher, thus indicates the necessity of sufficient contacts are required for high-affinity binding. It seems like that the middle range distance (~10-20 Å) residues related contacts features play a more important role in the classification. The switch function and summation applied in the contact feature calculations take long-range contacts in consideration, thus it explains why Gly702 has some contacts with docked ligands.

6.3.3 One-short learning VS planar learning

The real-world dataset of the protein-ligand complexes is also heavily imbalanced. The strong binders are always masked by majority low importance molecules. It is quite natural since the suitable ligands which could be well accommodated into the specific binding pocket in the receptor is only a very small subset of the large chemoinformatic space. Therefore, given a training set with a relative few positive samples, how to keep the positive sample information and avoid them from flooding of the great portion of the negative samples is quite a common problem in machine learning model construction for drug discovery, as well as other types of dataset. The inequality of the dataset, which could lead to the imbalance of a classification model, could not be ignored. Here, in our case, we observed the bias towards the negative samples.

In order to solve the data inequality problem, it is proposed that by splitting the negative samples into smaller subsets and combine them with the positive dataset individually, the prediction power could be greatly enhanced. We applied the same psychology to bias the dataset less towards the negative samples, through a strategy called “planar learning”.

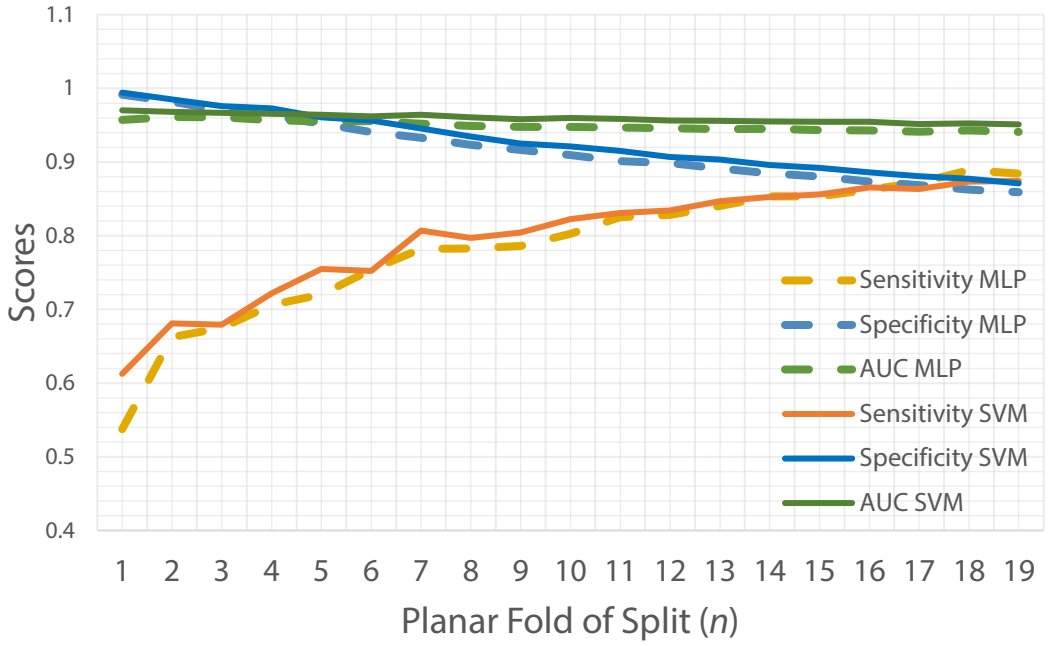


Figure 42. The prediction power of planar learning with SVM and MLP dataset.

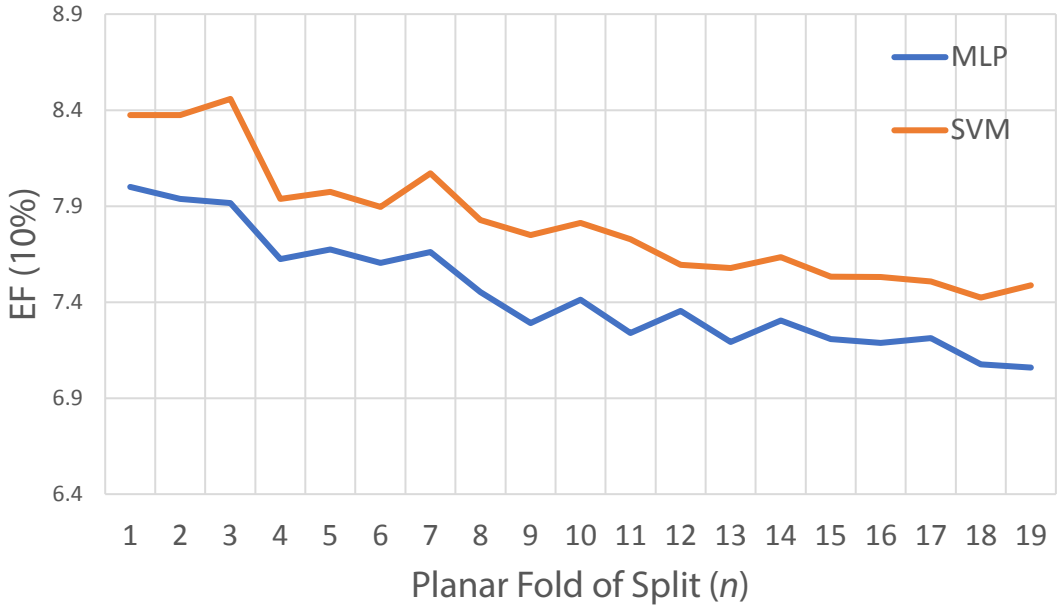


Figure 43. The EF (at 10%) of the planar learning with SVM and MLP.

In a one-shot learning process, we just throw the training dataset into a classification model and obtain the learning performance. Though it may be straightforward to using this strategy, the performance would not be optimal for an imbalanced dataset. Using either the original normalized dataset or engineered dataset, the specificity of the models is very satisfied, indicating that the models could

correctly classify a true negative molecule as the negative group. However, these models failed to make the right predictions about the active (positive) molecules. Only about 50% to 62% of the true active molecules could be correctly classified as positive molecules.

When we split the negative samples into *n* (planar fold) different equally random distributed subsets, we constructed *n* * 5 models and using the averaging scores as the indications of the prediction power. We noticed the decreasing of the specificity, as the sensitivity increases (Figure 42). However, as *n* rises, the overall enrichment ability drops quickly, whereas the AUC remains almost stationary (Figure 43).

In the case that we are trying to lower down the false positive rate, or maximize the specificity, we found that the planar learning doesn't improve our prediction power, meanwhile, it doesn't increase the specificity. The advantage of the planar learning here, in our case, is not an optimal choice.

6.4 Summary

The high false positive ratio is one of the major weakness of docking simulations. To handle this issue, the improvements in both the searching algorithms and the scoring functions would be valuable. Current docking packages are effective to discover native-like ligand poses, as proved in this study, but are hardly accurate to assess the binding strength between a ligand and its target. Here we proposed that by including ML-aided rescoring strategy, the false positive rate would be greatly decreased. Basically, we prepared a rescoring working flow after the completion of the docking simulations, to provide more accurate binding strength estimations. In detail, after the docking both the active and decoy molecules to PR LBD, the ligand-protein complexes were preprocessed to extract the per-residue based decomposed binding energies and atomic contacts. Thus, we had a positive dataset describing the active ligands and PR LBD interactions, and a negative dataset accounting for the 3D interactions between the decoy molecules and PR LBD. Then the SVM and ANN models were constructed to classify the active compounds from the decoy ones. The models achieved false positive rate less than 5%, rather high AUC (0.97) and strong enrichment power. Further, using the models, we screened unseen compound libraries and identified 21 most possible PR LBD binders. The molecules were proved to stably bind to PR LBD through MD simulations. The ML-aided rescoring model is also very practical to discover new lead-like molecules for other targets.

Chapter 7. Conclusions and recommendations

There is always a void in the structural, dynamics and lead discovery study of PR LBD, especially the ligand-free form. Current biochemistry and molecular biology techniques are not economic and practical to solve the problem. We for the first time, applied computational methods to tackle the structure-dynamics issue of PR LBD, and we identified the possible reason of the flexibility of PR LBD, and identified key residues governing the dynamics and ligand-induced adaptations of PRL BD. The ensemble docking, and ML-aided rescoring strategies were also introduced for PR LBD lead discovery for the first time.

The *apo*-form LBD is quite flexible and has several different metastable conformations. It is clear the agonistic conformation is a stable state of *apo*-form PR LBD, whereas the antagonistic conformation is not. We also observed the ligand binding-induced conformational adaptation. Both an agonist or antagonist binding would shift the antagonistic PR LBD towards the agonistic state, where the H11-H12 could not be well positioned though. During the antagonistic state to antagonistic state, the hydrophobic core is firstly formed and is followed by the formation of stable electrostatic interaction networks around the ligand. Among the interacting residues, Glu723, Trp755, Val912, Arg899, Met909 and Ile913 are the central nodes of the interaction networks. Meanwhile, the co-existence of the antagonist and the co-repressor peptide is the necessity to maintain the antagonistic LBD conformation.

The drug discovery is never an easy task. Many efforts have been endowed in this active area, and no such a method could provide high enrichment power in meanwhile is resources friendly. Here, we adopted the traditional docking methodology followed by ML-based models to re-score the docking results. Based on the conformations we sampled in MD simulations, we performed large scale VS against LBD. Several potential ligands have been identified. Further MD simulations of these ligand-LBD complexes were performed, and their binding energies were assessed. We further identified 4 ligands for future examinations. To lower down the possible high false positive rate of the VS, we employed MLP and SVM based ML models, using the interaction fingerprint dataset as features and the DUD ligand database as training samples. The rescoring model constructed by ML model aggregations demonstrates quite high specificities than GOLD and AutoDock Vina. We incorporated the rescoring model to predict the PR LBD lead like molecules and identified 21 highly possible ligands, based on their probabilities and drug-like parameters. We believe that the docking-rescoring strategy is a more reliable method for more precise drug discovery in the future. The dynamics of *apo*-form, as well as the conformational adaptations of PR LBD, have been explored. However, further analysis of the free energy surface of ligand bound PR LBD is still interesting. In the future, we could adopt ligand/receptor interaction-based funnel bias, or CVs, to simulate the dissociation or binding process of a ligand, using metadynamics or other enhanced samplings methods.

Meanwhile, how ligand affects the energy landscape or dynamics of PR LBD, is still very difficult to capture. But the information would be quite valuable for understanding the relationship of the LBD and a ligand. In order to obtain the energy landscape, the HMM or MSM methods could be applied. Firstly, we generate an artificial transition pathway, from "open" state to "closed" state with ligand in the binding pocket, then starting from the middle states of the pathway, multiple short cMD simulations could be performed. Based the short trajectories, the seed structures could be selected by partitioning the PR LBD configurational space. Now, starting from the seed structures, adaptive sampling strategy could be applied to thoroughly sample the configurational space, until a convergence condition is met. Then, using all trajectories, an MSM model of the metastable states could be constructed, and the transition possibilities, as well as transition time-scale, between states, could be estimated. The real transition pathways and the realistic state conformations thus could be obtained. Using MSM or HMM methods, we could compare the transition pathway of PR LBD with or without ligand, or with different ligands. The knowledge would explain where the partial agonism rises for antagonists.

Secondly, it has been proved that the cooperation between the PR domains is vital for the normal biological functions, such as the ligand-induced dimerization and subsequent co-activators binding, and downstream gene expressions. Thus, we need to answer these questions, 1) what are the dynamics of the NTD and DBD of PR? 2) and how these three domains interact with each other? 3) how collaborations between the domains occur? 4) how ligand binding changes the overall dynamics of PR and the collective interaction networks between domains? To elucidate the unknown mechanism of the PR, we may firstly model the NTD structure, then utilize the coarse-graining methods to model the whole PR dynamics. However, when considering ligand-induced conformational changes of the whole PR, it would be desirable to use the all-atom cMD or enhanced sampling methods to have the detailed atomic-level resolution, though enormous computational resources would be required to solve the problem.

Thirdly, more accurate rescoring functions could be constructed to screening lead like molecules. We have generated two binding fingerprint features based machine learning models. However, the models could not be verified yet. Next step, we would purchase the predicted high potential good binders, and perform experimental binding affinity and cellular agonism effect tests. The real power of the machine learning models could thus be verified with the experimental results. Further, we would fine tune the machine learning models, and apply more stochastic features, incorporating hydrogen bonding, salt bridging, as well as π-π stacking information. More detailed fingerprint features would add up to the accuracy of the machine learning models. Besides the SVM and MLP models, we would like to introduce deep learning models to classify the good and bad molecules, and eventually predict the binding affinities of the molecule/LBD complexes.

In one way, we would like to more thoroughly explore the dynamic properties of PR. In another hand, we hope to use the more accurate scoring function, or rescoring method to screen lead like molecules against PR. Finally, we would incorporate the dynamics information to guide more precisely drug discovery towards PR.

References

1. Mangelsdorf DJ, Thummel C, Beato M, Herrlich P, Schutz G, Umesono K, et al. The nuclear receptor superfamily: the second decade. Cell. 1995;83(6):835-9. PubMed PMID: 8521507.

2. Bain DL, Heneghan AF, Connaghan-Jones KD, Miura MT. Nuclear receptor structure: implications for function. Annu Rev Physiol. 2007;69:201-20. doi: 10.1146/annurev.physiol.69.031905.160308. PubMed PMID: 17137423.

3. Hill KK, Roemer SC, Churchill ME, Edwards DP. Structural and functional analysis of domains of the progesterone receptor. Mol Cell Endocrinol. 2012;348(2):418-29. doi: 10.1016/j.mce.2011.07.017. PubMed PMID: 21803119.

4. Kastner P, Krust A, Turcotte B, Stropp U, Tora L, Gronemeyer H, et al. Two distinct estrogen-regulated promoters generate transcripts encoding the two functionally different human progesterone receptor forms A and B. EMBO J. 1990;9(5):1603-14. PubMed PMID: 2328727; PubMed Central PMCID: PMC551856.

5. Takimoto GS, Tung L, Abdel-Hafiz H, Abel MG, Sartorius CA, Richer JK, et al. Functional properties of the N-terminal region of progesterone receptors and their mechanistic relationship to structure. J Steroid Biochem Mol Biol. 2003;85(2-5):209-19. PubMed PMID: 12943706.

6. Fernandez-Valdivia R, Mukherjee A, Mulac-Jericevic B, Conneely OM, DeMayo FJ, Amato P, et al. Revealing progesterone's role in uterine and mammary gland biology: insights from the mouse. Semin Reprod Med. 2005;23(1):22-37. doi: 10.1055/s-2005-864031. PubMed PMID: 15714387.

7. Meyer ME, Pornon A, Ji JW, Bocquel MT, Chambon P, Gronemeyer H. Agonistic and antagonistic activities of RU486 on the functions of the human progesterone receptor. EMBO J. 1990;9(12):3923-32. PubMed PMID: 2249658; PubMed Central PMCID: PMC552163.

8. Richer JK, Jacobsen BM, Manning NG, Abel MG, Wolf DM, Horwitz KB. Differential gene regulation by the two progesterone receptor isoforms in human breast cancer cells. J Biol Chem. 2002;277(7):5209-18. doi: 10.1074/jbc.M110090200. PubMed PMID: 11717311.

9. Bellance C, Khan JA, Meduri G, Guiochon-Mantel A, Lombès M, Loosfelt H. Progesterone receptor isoforms PRA and PRB differentially contribute to breast cancer cell migration through interaction with focal adhesion kinase complexes. Molecular biology of the cell. 2013;24(9):1363-74.

10. Recouvreux MS, Sampayo R, Simian M. Progesterone receptor isoform ratio regulates the stem cell population in the mouse mammary gland. Cancer Research. 2015;75(15 Supplement):2240-.

11. Guiochon-Mantel A, Loosfelt H, Lescop P, Sar S, Atger M, Perrot-Applanat M, et al. Mechanisms of nuclear localization of the progesterone receptor: evidence for interaction between monomers. Cell. 1989;57(7):1147-54. PubMed PMID: 2736623.

12. Jacobsen BM, Horwitz KB. Progesterone receptors, their isoforms and progesterone regulated transcription. Molecular and cellular endocrinology. 2012;357(1-2):18-29.

13. Zhang C, McFarlane C, Lokireddy S, Bonala S, Ge X, Masuda S, et al. Erratum to: 'Inhibition of myostatin protects against diet-induced obesity by enhancing fatty acid oxidation and promoting a brown adipose phenotype in mice' and 'Myostatin-deficient mice exhibit reduced insulin resistance through activating the AMP-activated protein kinase signalling pathway'. Diabetologia. 2015;58(3):643. doi: 10.1007/s00125-014-3450-2. PubMed PMID: 25500699.

14. Stanczyk FZ. All progestins are not created equal. Steroids. 2003;68(10-13):879-90. PubMed PMID: 14667980.

15. Sitruk-Ware R. Pharmacological profile of progestins. Maturitas. 2008;61(1-2):151-7. PubMed PMID: 19434887.

16. Obr AE, Edwards DP. The biology of progesterone receptor in the normal mammary gland and in breast cancer. Mol Cell Endocrinol. 2012;357(1-2):4-17. doi: 10.1016/j.mce.2011.10.030. PubMed PMID: 22193050; PubMed Central PMCID: PMC3318965.

17. Fournier A, Berrino F, Clavel-Chapelon F. Unequal risks for breast cancer associated with different hormone replacement therapies: results from the E3N cohort study. Breast Cancer Res Treat. 2008;107(1):103-11. doi: 10.1007/s10549-007-9523-x. PubMed PMID: 17333341; PubMed Central PMCID: PMC2211383.

18. Bentel JM, Birrell SN, Pickering MA, Holds DJ, Horsfall DJ, Tilley WD. Androgen receptor agonist activity of the synthetic progestin, medroxyprogesterone acetate, in human breast cancer cells. Mol Cell Endocrinol. 1999;154(1-2):11-20. PubMed PMID: 10509795.

19. Zheng ZY, Bay BH, Aw SE, Lin VC. A novel antiestrogenic mechanism in progesterone receptor-transfected breast cancer cells. J Biol Chem. 2005;280(17):17480-7. doi: 10.1074/jbc.M501261200. PubMed PMID: 15728178.

20. Chen CC, Hardy DB, Mendelson CR. Progesterone receptor inhibits proliferation of human breast cancer cells via induction of MAPK phosphatase 1 (MKP-1/DUSP1). J Biol Chem. 2011;286(50):43091-102. doi: 10.1074/jbc.M111.295865. PubMed PMID: 22020934; PubMed Central PMCID: PMC3234857.

21. Bardou VJ, Arpino G, Elledge RM, Osborne CK, Clark GM. Progesterone receptor status significantly improves outcome prediction over estrogen receptor status alone for adjuvant endocrine therapy in two large breast cancer databases. J Clin Oncol. 2003;21(10):1973-9. doi: 10.1200/JCO.2003.09.099. PubMed PMID: 12743151.

22. Savouret JF, Chauchereau A, Misrahi M, Lescop P, Mantel A, Bailly A, et al. The progesterone receptor. Biological effects of progestins and antiprogestins. Hum Reprod. 1994;9 Suppl 1:7-11. PubMed PMID: 7962473.

23. Mohammed H, Russell IA, Stark R, Rueda OM, Hickey TE, Tarulli GA, et al. Progesterone receptor modulates ERalpha action in breast cancer. Nature. 2015;523(7560):313-7. doi: 10.1038/nature14583. PubMed PMID: 26153859.

24. Seton-Rogers S. Breast cancer: Untangling the role of progesterone receptors. Nat Rev Cancer. 2015;15(8):456. doi: 10.1038/nrc3991. PubMed PMID: 26205336.

25. Madauss KP, Grygielko ET, Deng SJ, Sulpizio AC, Stanley TB, Wu C, et al. A structural and in vitro characterization of asoprisnil: a selective progesterone receptor modulator. Mol Endocrinol. 2007;21(5):1066-81. doi: 10.1210/me.2006-0524. PubMed PMID: 17356170.

26. Klijn JG, Setyono-Han B, Foekens JA. Progesterone antagonists and progesterone receptor modulators in the treatment of breast cancer. Steroids. 2000;65(10-11):825-30. PubMed PMID: 11108894.

27. Spitz IM, Chwalisz K. Progesterone receptor modulators and progesterone antagonists in women's health. Steroids. 2000;65(10-11):807-15. PubMed PMID: 11108892.

28. Spitz IM. Progesterone antagonists and progesterone receptor modulators: an overview. Steroids. 2003;68(10-13):981-93. PubMed PMID: 14667991.

29. Gagne D, Pons M, Philibert D. RU 38486: a potent antiglucocorticoid in vitro and in vivo. J Steroid Biochem. 1985;23(3):247-51. PubMed PMID: 2864478.

30. De Leo V, Morgante G, La Marca A, Musacchio MC, Sorace M, Cavicchioli C, et al. A benefit-risk assessment of medical treatment for uterine leiomyomas. Drug Saf. 2002;25(11):759-79. PubMed PMID: 12222988.

31. DeManno D, Elger W, Garg R, Lee R, Schneider B, Hess-Stumpp H, et al. Asoprisnil (J867): a selective progesterone receptor modulator for gynecological therapy. Steroids. 2003;68(10-13):1019-32.

32. Spitz IM. Progesterone antagonists and progesterone receptor modulators. Expert Opin Investig Drugs. 2003;12(10):1693-707. doi: 10.1517/13543784.12.10.1693. PubMed PMID: 14519088.

33. Roemer SC, Donham DC, Sherman L, Pon VH, Edwards DP, Churchill ME. Structure of the progesterone receptor-deoxyribonucleic acid complex: novel interactions required for binding to half-site response elements. Molecular endocrinology. 2006;20(12):3042-52.

34. Vente-Spreeuwenberg MA, Verdonk JM, Verstegen MW, Beynen AC. Villus height and gut development in weaned piglets receiving diets containing either glucose, lactose or starch. Br J Nutr. 2003;90(5):907-13. PubMed PMID: 14667184.

35. Bourguet W, Ruff M, Chambon P, Gronemeyer H, Moras D. Crystal structure of the ligand-binding domain of the human nuclear receptor RXR-alpha. Nature. 1995;375(6530):377-82. doi: 10.1038/375377a0. PubMed PMID: 7760929.

36. Batista MR, Martinez L. Dynamics of nuclear receptor Helix-12 switch of transcription activation by modeling time-resolved fluorescence anisotropy decays. Biophys J. 2013;105(7):1670-80. doi: 10.1016/j.bpj.2013.07.032. PubMed PMID: 24094408; PubMed Central PMCID: PMC3791304.

37. Tanenbaum DM, Wang Y, Williams SP, Sigler PB. Crystallographic comparison of the estrogen and progesterone receptor's ligand binding domains. Proceedings of the National Academy of Sciences. 1998;95(11):5998-6003.

38. Xu J, Li Q. Review of the in vivo functions of the p160 steroid receptor coactivator family. Mol Endocrinol. 2003;17(9):1681-92. doi: 10.1210/me.2003-0116. PubMed PMID: 12805412.

39. Nocker M, Cozzini P. Induced fit simulations on nuclear receptors. Curr Top Med Chem. 2011;11(2):133-47. PubMed PMID: 20939793.

40. Williams SP, Sigler PB. Atomic structure of progesterone complexed with its receptor. Nature. 1998;393(6683):392-6. doi: 10.1038/30775. PubMed PMID: 9620806.

41. Petit-Topin I, Turque N, Fagart J, Fay M, Ulmann A, Gainer E, et al. Met909 plays a key role in the activation of the progesterone receptor and also in the high potency of 13-ethyl progestins. Mol Pharmacol. 2009;75(6):1317-24. doi: 10.1124/mol.108.054312. PubMed PMID: 19289570.

42. Raaijmakers HC, Versteegh JE, Uitdehaag JC. The X-ray structure of RU486 bound to the progesterone receptor in a destabilized agonistic conformation. J Biol Chem. 2009;284(29):19572-9. doi: 10.1074/jbc.M109.007872. PubMed PMID: 19372222; PubMed Central PMCID: PMC2740583.

43. Lusher SJ, Raaijmakers HC, Vu-Pham D, Kazemier B, Bosch R, McGuire R, et al. X-ray structures of progesterone receptor ligand binding domain in its agonist state reveal differing mechanisms for mixed profiles of 11beta-substituted steroids. J Biol Chem. 2012;287(24):20333-43. doi: 10.1074/jbc.M111.308403. PubMed PMID: 22535964; PubMed Central PMCID: PMC3370215.

44. Palsson B. The challenges of in silico biology. Nature biotechnology. 2000;18(11):1147-50.

45. Bloom M. Biology in silico: the bioinformatics revolution. The American Biology Teacher. 2001;63(6):400-7.

46. Nussinov R, Bonhoeffer S, Papin JA, Sporns O. From “What Is?” to “What Isn't?” Computational Biology. PLoS Comput Biol. 2015;11(7):e1004318.

47. Setubal JC, Meidanis J, Setubal-Meidanis. Introduction to computational molecular biology: PWS Pub.; 1997.

48. Frenkel D, Smit B. Understanding molecular simulation: from algorithms to applications: Academic press; 2001.

49. McCammon JA, Gelin BR, Karplus M. Dynamics of folded proteins. Nature. 1977;267(5612):585-90.

50. Ogata K, Yuki T, Hatakeyama M, Uchida W, Nakamura S. All-atom molecular dynamics simulation of photosystem II embedded in thylakoid membrane. Journal of the American Chemical Society. 2013;135(42):15670-3.

51. Allen MP, Tildesley DJ. Computer simulation of liquids: Oxford university press; 1989.

52. Halgren TA. Merck molecular force field. I. Basis, form, scope, parameterization, and performance of MMFF94. J Comput Chem. 1996;17(5‐6):490-519.

53. Huang K. Statistical Mechanics, 18.3. Wiley; 1987.

54. Kubelka J, Hofrichter J, Eaton WA. The protein folding ‘speed limit’. Current opinion in structural biology. 2004;14(1):76-88.

55. Tolman RC. The principles of statistical mechanics: Courier Corporation; 1938.

56. Sugita Y, Okamoto Y. Replica-exchange molecular dynamics method for protein folding. Chemical physics letters. 1999;314(1):141-51.

57. Liu P, Kim B, Friesner RA, Berne B. Replica exchange with solute tempering: A method for sampling biological systems in explicit water. Proceedings of the National Academy of Sciences of the United States of America. 2005;102(39):13749-54.

58. Marinari E, Parisi G. Simulated tempering: a new Monte Carlo scheme. EPL (Europhysics Letters). 1992;19(6):451.

59. Zhang T, Nguyen PH, Nasica-Labouze J, Mu Y, Derreumaux P. Folding Atomistic Proteins in Explicit Solvent Using Simulated Tempering. The Journal of Physical Chemistry B. 2015.

60. Torrie GM, Valleau JP. Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling. Journal of Computational Physics. 1977;23(2):187-99.

61. Laio A, Parrinello M. Escaping free-energy minima. Proceedings of the National Academy of Sciences. 2002;99(20):12562-6.

62. Sutto L, Marsili S, Gervasio FL. New advances in metadynamics. Wiley Interdisciplinary Reviews: Computational Molecular Science. 2012;2(5):771-9.

63. Mu Y, Yang Y, Xu W. Hybrid Hamiltonian replica exchange molecular dynamics simulation method employing the Poisson–Boltzmann model. The Journal of chemical physics. 2007;127(8):084119.

64. Bussi G, Gervasio FL, Laio A, Parrinello M. Free-energy landscape for β hairpin folding from combined parallel tempering and metadynamics. Journal of the American Chemical Society. 2006;128(41):13435-41.

65. Torrie GM, Valleau JP. Monte Carlo free energy estimates using non-Boltzmann sampling: application to the sub-critical Lennard-Jones fluid. Chemical Physics Letters. 1974;28(4):578-81.

66. Kästner J. Umbrella sampling. Wiley Interdisciplinary Reviews: Computational Molecular Science. 2011;1(6):932-42.

67. Rosenbergl JM. The weighted histogram analysis method for free-energy calculations on biomolecules. I. The method. Journal of computational chemistry. 1992;13(8):1011-21.

68. Banavali NK, Roux B. Free energy landscape of A-DNA to B-DNA conversion in aqueous solution. J Am Chem Soc. 2005;127(18):6866-76. doi: 10.1021/ja050482k. PubMed PMID: 15869310.

69. Barducci A, Bussi G, Parrinello M. Well-tempered metadynamics: a smoothly converging and tunable free-energy method. Phys Rev Lett. 2008;100(2):020603. PubMed PMID: 18232845.

70. Piana S, Laio A. A bias-exchange approach to protein folding. J Phys Chem B. 2007;111(17):4553-9. doi: 10.1021/jp067873l. PubMed PMID: 17419610.

71. Bussi G, Gervasio FL, Laio A, Parrinello M. Free-energy landscape for beta hairpin folding from combined parallel tempering and metadynamics. J Am Chem Soc. 2006;128(41):13435-41. doi: 10.1021/ja062463w. PubMed PMID: 17031956.

72. Bonomi M, Branduardi D, Bussi G, Camilloni C, Provasi D, Raiteri P, et al. PLUMED: A portable plugin for free-energy calculations with molecular dynamics. Computer Physics Communications. 2009;180(10):1961-72.

73. Biarnés X, Pietrucci F, Marinelli F, Laio A. METAGUI. A VMD interface for analyzing metadynamics and molecular dynamics simulations. Computer Physics Communications. 2012;183(1):203-11.

74. Sugita Y, Okamoto Y. Replica-exchange molecular dynamics method for protein folding. Chem Phys Lett. 1999;314(1-2):141-51. doi: Doi 10.1016/S0009-2614(99)01123-9. PubMed PMID: WOS:000083955300022.

75. La Penna G, Morante S, Perico A, Rossi GC. Designing generalized statistical ensembles for numerical simulations of biopolymers. Journal of Chemical Physics. 2004;121(21):10725-41. doi: 10.1063/1.1795694. PubMed PMID: WOS:000225136300051.

76. Mitsutake A, Sugita Y, Okamoto Y. Generalized‐ensemble algorithms for molecular simulations of biopolymers. Peptide Science. 2001;60(2):96-123.

77. Tai K. Conformational sampling for the impatient. Biophysical chemistry. 2004;107(3):213-20.

78. García AE, Sanbonmatsu KY. α-Helical stabilization by side chain shielding of backbone hydrogen bonds. Proceedings of the National Academy of Sciences. 2002;99(5):2782-7.

79. Sindhikara D, Meng YL, Roitberg AE. Exchange frequency in replica exchange molecular dynamics. Journal of Chemical Physics. 2008;128(2). doi: Artn 024103

10.1063/1.2816560. PubMed PMID: WOS:000252450100006.

80. Cecchini M, Rao F, Seeber M, Caflisch A. Replica exchange molecular dynamics simulations of amyloid peptide aggregation. Journal of Chemical Physics. 2004;121(21):10748-56. doi: 10.1063/1.1809588. PubMed PMID: WOS:000225136300053.

81. Zhang W, Wu C, Duan Y. Convergence of replica exchange molecular dynamics. Journal of Chemical Physics. 2005;123(15). doi: Artn 154105

10.1063/1.2056540. PubMed PMID: WOS:000232697900007.

82. Swendsen RH, Wang J-S. Nonuniversal critical dynamics in Monte Carlo simulations. Physical review letters. 1987;58(2):86.

83. Kone A, Kofke DA. Selection of temperature intervals for parallel-tempering simulations. The Journal of chemical physics. 2005;122(20):206101.

84. Lin E, Shell MS. Convergence and Heterogeneity in Peptide Folding with Replica Exchange Molecular Dynamics. J Chem Theory Comput. 2009;5(8):2062-73. Epub 2009/08/11. doi: 10.1021/ct900119n. PubMed PMID: 26613148.

85. Denschlag R, Lingenheil M, Tavan P. Efficiency reduction and pseudo-convergence in replica exchange sampling of peptide folding-unfolding equilibria. Chemical Physics Letters. 2008;458(1-3):244-8. doi: 10.1016/j.cplett.2008.04.114. PubMed PMID: WOS:000256284900052.

86. Okur A, Roe DR, Cui G, Hornak V, Simmerling C. Improving Convergence of Replica-Exchange Simulations through Coupling to a High-Temperature Structure Reservoir. J Chem Theory Comput. 2007;3(2):557-68. Epub 2007/03/01. doi: 10.1021/ct600263e. PubMed PMID: 26637035.

87. Zhang W, Wu C, Duan Y. Convergence of replica exchange molecular dynamics. J Chem Phys. 2005;123(15):154105. Epub 2005/10/29. doi: 10.1063/1.2056540. PubMed PMID: 16252940.

88. Mu Y. Dissociation aided and side chain sampling enhanced Hamiltonian replica exchange. J Chem Phys. 2009;130(16):164107. doi: 10.1063/1.3120483. PubMed PMID: 19405561.

89. Zhang J-H, Chung TD, Oldenburg KR. A simple statistical parameter for use in evaluation and validation of high throughput screening assays. Journal of biomolecular screening. 1999;4(2):67-73.

90. Bleicher KH, Böhm H-J, Müller K, Alanine AI. A guide to drug discovery: Hit and lead generation: beyond high-throughput screening. Nature reviews Drug discovery. 2003;2(5):369.

91. Koeppen H, Kriegl J, Lessel U, Tautermann CS, Wellenzohn B. Ligand‐Based Virtual Screening. Virtual Screening: Principles, Challenges, and Practical Guidelines. 2011:61-85.

92. Geppert H, Vogt M, Bajorath Jr. Current trends in ligand-based virtual screening: molecular representations, data mining methods, new application areas, and performance evaluation. Journal of chemical information and modeling. 2010;50(2):205-16.

93. Ghosh S, Nie A, An J, Huang Z. Structure-based virtual screening of chemical libraries for drug discovery. Current opinion in chemical biology. 2006;10(3):194-202.

94. Villoutreix BO, Eudes R, Miteva MA. Structure-based virtual ligand screening: recent success stories. Combinatorial chemistry & high throughput screening. 2009;12(10):1000-16.

95. Rueda M, Bottegoni G, Abagyan R. Consistent improvement of cross-docking results using binding site ensembles generated with elastic network normal modes. Journal of chemical information and modeling. 2009;49(3):716-25.

96. Awuni Y, Mu Y. Reduction of false positives in structure-based virtual screening when receptor plasticity is considered. Molecules. 2015;20(3):5152-64. Epub 2015/03/27. doi: 10.3390/molecules20035152. PubMed PMID: 25808156.

97. Okamoto M, Takayama K, Shimizu T, Ishida K, Takahashi O, Furuya T. Identification of death-associated protein kinases inhibitors using structure-based virtual screening. Journal of medicinal chemistry. 2009;52(22):7323-7.

98. Awuni Y, Mu Y. Reduction of False Positives in Structure-Based Virtual Screening When Receptor Plasticity Is Considered. Molecules. 2015;20(3):5152-64.

99. Case DA. Normal mode analysis of protein dynamics. Current Opinion in Structural Biology. 1994;4(2):285-90.

100. Amadei A, Linssen A, Berendsen HJ. Essential dynamics of proteins. Proteins: Structure, Function, and Bioinformatics. 1993;17(4):412-25.

101. Jones G, Willett P, Glen RC, Leach AR, Taylor R. Development and validation of a genetic algorithm for flexible docking. Journal of molecular biology. 1997;267(3):727-48.

102. Ewing TJ, Makino S, Skillman AG, Kuntz ID. DOCK 4.0: search strategies for automated molecular docking of flexible molecule databases. Journal of computer-aided molecular design. 2001;15(5):411-28.

103. Morris GM, Goodsell DS, Halliday RS, Huey R, Hart WE, Belew RK, et al. Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. Journal of computational chemistry. 1998;19(14):1639-62.

104. Muegge I. PMF scoring revisited. Journal of medicinal chemistry. 2006;49(20):5895-902.

105. Gohlke H, Hendlich M, Klebe G. Knowledge-based scoring function to predict protein-ligand interactions1. Journal of molecular biology. 2000;295(2):337-56.

106. Mooij W, Verdonk ML. General and targeted statistical potentials for protein–ligand interactions. Proteins: Structure, Function, and Bioinformatics. 2005;61(2):272-87.

107. Wang R, Lai L, Wang S. Further development and validation of empirical scoring functions for structure-based binding affinity prediction. Journal of computer-aided molecular design. 2002;16(1):11-26.

108. Eldridge MD, Murray CW, Auton TR, Paolini GV, Mee RP. Empirical scoring functions: I. The development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes. Journal of computer-aided molecular design. 1997;11(5):425-45.

109. Böhm H-J. Prediction of binding constants of protein ligands: a fast method for the prioritization of hits obtained from de novo design or 3D database search programs. Journal of computer-aided molecular design. 1998;12(4):309-.

110. Ain QU, Aleksandrova A, Roessler FD, Ballester PJ. Machine‐learning scoring functions to improve structure‐based binding affinity prediction and virtual screening. Wiley Interdisciplinary Reviews: Computational Molecular Science. 2015;5(6):405-24.

111. Chen H, Lyne PD, Giordanetto F, Lovell T, Li J. On evaluating molecular-docking methods for pose prediction and enrichment factors. J Chem Inf Model. 2006;46(1):401-15. Epub 2006/01/24. doi: 10.1021/ci0503255. PubMed PMID: 16426074.

112. Friesner RA, Banks JL, Murphy RB, Halgren TA, Klicic JJ, Mainz DT, et al. Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. Journal of medicinal chemistry. 2004;47(7):1739-49.

113. Trott O, Olson AJ. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. Journal of computational chemistry. 2010;31(2):455-61.

114. Jain AN. Surflex: fully automatic flexible molecular docking using a molecular similarity-based search engine. Journal of medicinal chemistry. 2003;46(4):499-511.

115. Venkatachalam CM, Jiang X, Oldfield T, Waldman M. LigandFit: a novel method for the shape-directed rapid docking of ligands to protein active sites. Journal of Molecular Graphics and Modelling. 2003;21(4):289-307.

116. Schölkopf B, Smola AJ. Learning with kernels: support vector machines, regularization, optimization, and beyond: MIT press; 2002.

117. Weaver DC. Applying data mining techniques to library design, lead generation and lead optimization. Curr Opin Chem Biol. 2004;8(3):264-70. Epub 2004/06/09. doi: 10.1016/j.cbpa.2004.04.005. PubMed PMID: 15183324.

118. Lavecchia A. Machine-learning approaches in drug discovery: methods and applications. Drug discovery today. 2015;20(3):318-31.

119. Walters WP, Namchuk M. A guide to drug discovery: designing screens: how to make your hits a hit. Nature reviews Drug discovery. 2003;2(4):259.

120. Gertrudes J, Maltarollo V, Silva R, Oliveira P, Honorio K, Da Silva A. Machine learning techniques and drug design. Current medicinal chemistry. 2012;19(25):4289-97.

121. Mitchell JB. Machine learning methods in chemoinformatics. Wiley Interdisciplinary Reviews: Computational Molecular Science. 2014;4(5):468-81.

122. Geppert H, Vogt M, Bajorath J. Current trends in ligand-based virtual screening: molecular representations, data mining methods, new application areas, and performance evaluation. J Chem Inf Model. 2010;50(2):205-16. Epub 2010/01/22. doi: 10.1021/ci900419k. PubMed PMID: 20088575.

123. Huang S-Y, Grinter SZ, Zou X. Scoring functions and their evaluation methods for protein–ligand docking: recent advances and future directions. Physical Chemistry Chemical Physics. 2010;12(40):12899-908.

124. Cheng T, Li Q, Zhou Z, Wang Y, Bryant SH. Structure-based virtual screening for drug discovery: a problem-centric review. The AAPS journal. 2012;14(1):133-41.

125. Deng W, Breneman C, Embrechts MJ. Predicting protein− ligand binding affinities using novel geometrical descriptors and machine-learning methods. Journal of chemical information and computer sciences. 2004;44(2):699-703.

126. Zhang S, Golbraikh A, Tropsha A. Development of Quantitative Structure− Binding Affinity Relationship Models Based on Novel Geometrical Chemical Descriptors of the Protein− Ligand Interfaces. Journal of medicinal chemistry. 2006;49(9):2713-24.

127. Artemenko N. Distance dependent scoring function for describing protein− ligand intermolecular interactions. Journal of chemical information and modeling. 2008;48(3):569-74.

128. Ballester PJ, Mitchell JB. A machine learning approach to predicting protein–ligand binding affinity with applications to molecular docking. Bioinformatics. 2010;26(9):1169-75.

129. Das S, Krein MP, Breneman CM. Binding affinity prediction with property-encoded shape distribution signatures. Journal of chemical information and modeling. 2010;50(2):298-308.

130. Durrant JD, McCammon JA. NNScore: a neural-network-based scoring function for the characterization of protein− ligand complexes. Journal of chemical information and modeling. 2010;50(10):1865-71.

131. Liu Q, Kwoh CK, Li J. Binding affinity prediction for protein–ligand complexes based on β contacts and B factor. Journal of chemical information and modeling. 2013;53(11):3076-85.

132. Zilian D, Sotriffer CA. SFCscore RF: a random forest-based scoring function for improved affinity prediction of protein–ligand complexes. Journal of chemical information and modeling. 2013;53(8):1923-33.

133. Ding B, Wang J, Li N, Wang W. Characterization of small molecule binding. I. Accurate identification of strong inhibitors in virtual screening. Journal of chemical information and modeling. 2013;53(1):114-22.

134. Melville JL, Burke EK, Hirst JD. Machine learning in virtual screening. Combinatorial chemistry & high throughput screening. 2009;12(4):332-43.

135. Xue L, Godden JW, Stahura FL, Bajorath J. Design and evaluation of a molecular fingerprint involving the transformation of property descriptor values into a binary classification scheme. Journal of chemical information and computer sciences. 2003;43(4):1151-7.

136. Kier LB, Hall LH, Hall K. Molecular structure description: the electrotopological state: Academic press New York; 1999.

137. Maw HH, Hall LH. E-state modeling of dopamine transporter binding. Validation of the model for a small data set. Journal of chemical information and computer sciences. 2000;40(5):1270-5.

138. González MP, Helguera AM. TOPS-MODE versus DRAGON descriptors to predict permeability coefficients through low-density polyethylene. Journal of computer-aided molecular design. 2003;17(10):665-72.

139. Karthikeyan M, Vyas R. Chemoinformatics approach for the design and screening of focused virtual libraries. Practical Chemoinformatics: Springer; 2014. p. 93-131.

140. Ma XH, Jia J, Zhu F, Xue Y, Li ZR, Chen YZ. Comparative analysis of machine learning methods in ligand-based virtual screening of large compound libraries. Comb Chem High Throughput Screen. 2009;12(4):344-57. Epub 2009/05/16. PubMed PMID: 19442064.

141. Sutter JM, Kalivas JH. Comparison of forward selection, backward elimination, and generalized simulated annealing for variable selection. Microchemical journal. 1993;47(1-2):60-6.

142. Guyon I, Weston J, Barnhill S, Vapnik V. Gene selection for cancer classification using support vector machines. Machine learning. 2002;46(1-3):389-422.

143. Lucasius CB, Kateman G. Understanding and using genetic algorithms Part 1. Concepts, properties and context. Chemometrics and intelligent laboratory systems. 1993;19(1):1-33.

144. Ballester PJ, Schreyer A, Blundell TL. Does a more precise chemical description of protein–ligand complexes lead to more accurate prediction of binding affinity? Journal of chemical information and modeling. 2014;54(3):944-55.

145. Xie X-QS. Exploiting PubChem for virtual screening. Expert opinion on drug discovery. 2010;5(12):1205-20.

146. Woutersen S, Pfister R, Hamm P, Mu Y, Kosov DS, Stock G. Peptide conformational heterogeneity revealed from nonlinear vibrational spectroscopy and molecular-dynamics simulations. The Journal of chemical physics. 2002;117(14):6833-40.

147. Snow CD, Zagrovic B, Pande VS. The Trp cage: folding kinetics and unfolded state topology via molecular dynamics simulations. J Am Chem Soc. 2002;124(49):14548-9.

148. Marinelli F, Pietrucci F, Laio A, Piana S. A kinetic model of trp-cage folding from multiple biased molecular dynamics simulations. Plos Comput Biol. 2009;5(8):e1000452.

149. Gnanakaran S, Nymeyer H, Portman J, Sanbonmatsu KY, Garcıa AE. Peptide folding simulations. Current opinion in structural biology. 2003;13(2):168-74.

150. Blanco FJ, Rivas G, Serrano L. A short linear peptide that folds into a native stable β-hairpin in aqueous solution. Nature Structural & Molecular Biology. 1994;1(9):584-90.

151. Ma B, Nussinov R. Stabilities and conformations of Alzheimer's β-amyloid peptide oligomers (Aβ16–22, Aβ16–35, and Aβ10–35): sequence effects. Proceedings of the National Academy of Sciences. 2002;99(22):14126-31.

152. Arkin MR, Tang Y, Wells JA. Small-molecule inhibitors of protein-protein interactions: progressing toward the reality. Chemistry & biology. 2014;21(9):1102-14.

153. Baaden M, Marrink SJ. Coarse-grain modelling of protein–protein interactions. Current opinion in structural biology. 2013;23(6):878-86.

154. Rajamani D, Thiel S, Vajda S, Camacho CJ. Anchor residues in protein–protein interactions. Proceedings of the National Academy of Sciences of the United States of America. 2004;101(31):11287-92.

155. Shrivastava IH, Sansom MS. Simulations of ion permeation through a potassium channel: molecular dynamics of KcsA in a phospholipid bilayer. Biophys J. 2000;78(2):557-70.

156. Im W, Roux Bt. Ions and counterions in a biological channel: a molecular dynamics simulation of OmpF porin from Escherichia coli in an explicit membrane with 1M KCl aqueous salt solution. Journal of molecular biology. 2002;319(5):1177-97.

157. Khalili-Araghi F, Gumbart J, Wen P-C, Sotomayor M, Tajkhorshid E, Schulten K. Molecular dynamics simulations of membrane channels and transporters. Current opinion in structural biology. 2009;19(2):128-37.

158. Gumbart J, Wang Y, Aksimentiev A, Tajkhorshid E, Schulten K. Molecular dynamics simulations of proteins in lipid bilayers. Current opinion in structural biology. 2005;15(4):423-31.

159. Freddolino PL, Arkhipov AS, Larson SB, McPherson A, Schulten K. Molecular dynamics simulations of the complete satellite tobacco mosaic virus. Structure. 2006;14(3):437-49.

160. Miao Y, Johnson JE, Ortoleva PJ. All-atom multiscale simulation of cowpea chlorotic mottle virus capsid swelling. The Journal of Physical Chemistry B. 2010;114(34):11181-95.

161. Beauchamp KA, McGibbon R, Lin YS, Pande VS. Simple few-state models reveal hidden complexity in protein folding. Proceedings of the National Academy of Sciences of the United States of America. 2012;109(44):17807-13. Epub 2012/07/11. doi: 10.1073/pnas.1201810109. PubMed PMID: 22778442; PubMed Central PMCID: PMCPMC3497769.

162. McGibbon RT, Pande VS. Learning kinetic distance metrics for Markov state models of protein conformational dynamics. Journal of chemical theory and computation. 2013;9(7):2900-6.

163. Suárez E, Adelman JL, Zuckerman DM. Accurate estimation of protein folding and unfolding times: beyond Markov state models. Journal of chemical theory and computation. 2016;12(8):3473-81.

164. van der Kamp MW, Mulholland AJ. Combined quantum mechanics/molecular mechanics (QM/MM) methods in computational enzymology. Biochemistry. 2013;52(16):2708-28. Epub 2013/04/06. doi: 10.1021/bi400215w. PubMed PMID: 23557014.

165. Hu L, Soderhjelm P, Ryde U. On the Convergence of QM/MM Energies. J Chem Theory Comput. 2011;7(3):761-77. Epub 2011/03/08. doi: 10.1021/ct100530r. PubMed PMID: 26596307.

166. Riccardi D, Yang S, Cui Q. Proton transfer function of carbonic anhydrase: Insights from QM/MM simulations. Biochim Biophys Acta. 2010;1804(2):342-51. Epub 2009/08/15. doi: 10.1016/j.bbapap.2009.07.026. PubMed PMID: 19679196.

167. Senn HM, Thiel W. QM/MM studies of enzymes. Curr Opin Chem Biol. 2007;11(2):182-7. Epub 2007/02/20. doi: 10.1016/j.cbpa.2007.01.684. PubMed PMID: 17307018.

168. Zhang Y, Liu H, Yang W. Free energy calculation on enzyme reactions with an efficient iterative procedure to determine minimum energy paths on a combined ab initio QM/MM potential energy surface. The Journal of Chemical Physics. 2000;112(8):3483-92. doi: 10.1063/1.480503.

169. Noé F, Schütte C, Vanden-Eijnden E, Reich L, Weikl TR. Constructing the equilibrium ensemble of folding pathways from short off-equilibrium simulations. Proceedings of the National Academy of Sciences. 2009;106(45):19011-6.

170. McGibbon R, Ramsundar B, Sultan M, Kiss G, Pande V, editors. Understanding protein dynamics with L1-regularized reversible hidden Markov models. International Conference on Machine Learning; 2014.

171. Thayer KM, Lakhani B, Beveridge DL. A Molecular Dynamics–Markov State Model of Protein Ligand Binding and Allostery in CRIB-PDZ: Conformational Selection and Induced Fit. The Journal of Physical Chemistry B. 2017.

172. Banerjee S, Prakash H, Mazumdar S. Evidence of molecular fragmentation inside the charged droplets produced by electrospray process. Journal of the American Society for Mass Spectrometry. 2011;22(10):1707-17. doi: 10.1007/s13361-011-0188-7. PubMed PMID: 21952884.

173. Pronk S, Páll S, Schulz R, Larsson P, Bjelkmar P, Apostolov R, et al. GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit. Bioinformatics. 2013:btt055.

174. Piana S, Lindorff-Larsen K, Shaw DE. How robust are protein folding simulations with respect to force field parameterization? Biophysical journal. 2011;100(9):L47-L9.

175. Frisch M, Trucks G, Schlegel HB, Scuseria G, Robb M, Cheeseman J, et al. Gaussian 09, Revision A. 02, Gaussian. Inc, Wallingford, CT. 2009;200.

176. Case DA, Darden T, Cheatham T, Simmerling CL, Wang J, Duke RE, et al. Amber 11. University of California, 2010.

177. Wang J, Wolf RM, Caldwell JW, Kollman PA, Case DA. Development and testing of a general amber force field. Journal of computational chemistry. 2004;25(9):1157-74.

178. Jorgensen WL, Chandrasekhar J, Madura JD, Impey RW, Klein ML. Comparison of simple potential functions for simulating liquid water. The Journal of chemical physics. 1983;79(2):926-35.

179. Bussi G, Donadio D, Parrinello M. Canonical sampling through velocity rescaling. The Journal of chemical physics. 2007;126(1):014101.

180. Darden T, York D, Pedersen L. Particle mesh Ewald: An N· log (N) method for Ewald sums in large systems. The Journal of chemical physics. 1993;98(12):10089-92.

181. Ryckaert J-P, Ciccotti G, Berendsen HJ. Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. Journal of Computational Physics. 1977;23(3):327-41.

182. Hess B, Bekker H, Berendsen HJ, Fraaije JG. LINCS: a linear constraint solver for molecular simulations. Journal of computational chemistry. 1997;18(12):1463-72.

183. Mu Y, Nguyen PH, Stock G. Energy landscape of a small peptide revealed by dihedral angle principal component analysis. Proteins: Structure, Function, and Bioinformatics. 2005;58(1):45-52.

184. !!! INVALID CITATION !!! .

185. Mecozzi S, West AP, Dougherty DA. Cation-pi interactions in aromatics of biological and medicinal interest: electrostatic potential surfaces as a useful qualitative guide. Proceedings of the National Academy of Sciences. 1996;93(20):10566-71.

186. McCammon JA, Harvey SC. Dynamics of proteins and nucleic acids: Cambridge University Press; 1988.

187. Seeber M, Cecchini M, Rao F, Settanni G, Caflisch A. Wordom: a program for efficient analysis of molecular dynamics simulations. Bioinformatics. 2007;23(19):2625-7. doi: 10.1093/bioinformatics/btm378. PubMed PMID: 17717034.

188. Seeber M, Felline A, Raimondi F, Muff S, Friedman R, Rao F, et al. Wordom: a user-friendly program for the analysis of molecular structures, trajectories, and free energy surfaces. J Comput Chem. 2011;32(6):1183-94. doi: 10.1002/jcc.21688. PubMed PMID: 21387345; PubMed Central PMCID: PMC3151548.

189. Hsin J, Arkhipov A, Yin Y, Stone JE, Schulten K. Using VMD: an introductory tutorial. Curr Protoc Bioinformatics. 2008;Chapter 5:Unit 5 7. doi: 10.1002/0471250953.bi0507s24. PubMed PMID: 19085979; PubMed Central PMCID: PMC2972669.

190. Humphrey W, Dalke A, Schulten K. VMD: visual molecular dynamics. J Mol Graph. 1996;14(1):33-8, 27-8. PubMed PMID: 8744570.

191. Le Guilloux V, Schmidtke P, Tuffery P. Fpocket: an open source platform for ligand pocket detection. BMC bioinformatics. 2009;10(1):168.

192. Lopéz-Blanco JR, Garzón JI, Chacón P. iMod: multipurpose normal mode analysis in internal coordinates. Bioinformatics. 2011;27(20):2843-50.

193. Branduardi D, Gervasio FL, Parrinello M. From A to B in free energy space. J Chem Phys. 2007;126(5):054103. doi: 10.1063/1.2432340. PubMed PMID: 17302470.

194. Laio A, Gervasio FL. Metadynamics: a method to simulate rare events and reconstruct the free energy in biophysics, chemistry and material science. Reports on Progress in Physics. 2008;71(12):126601.

195. Tiwary P, Parrinello M. A time-independent free energy estimator for metadynamics. J Phys Chem B. 2015;119(3):736-42. doi: 10.1021/jp504920s. PubMed PMID: 25046020.

196. Palmer D, Frater J, Phillips R, McLean AR, McVean G. Integrating genealogical and dynamical modelling to infer escape and reversion rates in HIV epitopes. Proc Biol Sci. 2013;280(1762):20130696. doi: 10.1098/rspb.2013.0696. PubMed PMID: 23677344; PubMed Central PMCID: PMC3673055.

197. Pietrucci F, Laio A. A collective variable for the efficient exploration of protein beta-sheet structures: Application to SH3 and GB1. Journal of Chemical Theory and Computation. 2009;5(9):2197-201.

198. Pfaendtner J, Branduardi D, Parrinello M, Pollard TD, Voth GA. Nucleotide-dependent conformational states of actin. Proc Natl Acad Sci U S A. 2009;106(31):12723-8. doi: 10.1073/pnas.0902092106. PubMed PMID: 19620726; PubMed Central PMCID: PMC2722336.

199. Limongelli V, Marinelli L, Cosconati S, La Motta C, Sartini S, Mugnaini L, et al. Sampling protein motion and solvent effect during ligand binding. Proc Natl Acad Sci U S A. 2012;109(5):1467-72. doi: 10.1073/pnas.1112181108. PubMed PMID: 22238423; PubMed Central PMCID: PMC3277130.

200. Favia AD, Masetti M, Recanatini M, Cavalli A. Substrate binding process and mechanistic functioning of type 1 11beta-hydroxysteroid dehydrogenase from enhanced sampling methods. PLoS One. 2011;6(9):e25375. doi: 10.1371/journal.pone.0025375. PubMed PMID: 21966510; PubMed Central PMCID: PMC3179505.

201. Tanenbaum DM, Wang Y, Williams SP, Sigler PB. Crystallographic comparison of the estrogen and progesterone receptor's ligand binding domains. Proc Natl Acad Sci U S A. 1998;95(11):5998-6003. PubMed PMID: 9600906; PubMed Central PMCID: PMC27574.

202. Blondel A, Renaud JP, Fischer S, Moras D, Karplus M. Retinoic acid receptor: a simulation analysis of retinoic acid binding and the resulting conformational changes. Journal of molecular biology. 1999;291(1):101-15. doi: 10.1006/jmbi.1999.2879. PubMed PMID: 10438609.

203. Baker ME, Chandsawangbhuwana C, Ollikainen N. Structural analysis of the evolution of steroid specificity in the mineralocorticoid and glucocorticoid receptors. BMC Evol Biol. 2007;7:24. doi: 10.1186/1471-2148-7-24. PubMed PMID: 17306029; PubMed Central PMCID: PMC1805736.

204. Celik L, Lund JD, Schiott B. Conformational dynamics of the estrogen receptor alpha: molecular dynamics simulations of the influence of binding site structure on protein dynamics. Biochemistry. 2007;46(7):1743-58. doi: 10.1021/bi061656t. PubMed PMID: 17249692.

205. Gallivan JP, Dougherty DA. Cation-pi interactions in structural biology. Proc Natl Acad Sci U S A. 1999;96(17):9459-64. PubMed PMID: 10449714; PubMed Central PMCID: PMC22230.

206. Fujii S, Yamada A, Nakano E, Takeuchi Y, Mori S, Masuno H, et al. Design and synthesis of nonsteroidal progesterone receptor antagonists based on C, C′-diphenylcarborane scaffold as a hydrophobic pharmacophore. European journal of medicinal chemistry. 2014;84:264-77.

207. Khan JA, Tikad A, Fay M, Hamze A, Fagart J, Chabbert-Buffet N, et al. A new strategy for selective targeting of progesterone receptor with passive antagonists. Molecular endocrinology. 2013;27(6):909-24. doi: 10.1210/me.2012-1328. PubMed PMID: 23579486.

208. Perryman AL, Lin JH, McCammon JA. HIV‐1 protease molecular dynamics of a wild‐type and of the V82F/I84V mutant: Possible contributions to drug resistance and a potential new target site for drugs. Protein Science. 2004;13(4):1108-23.

209. Miao Y, Nichols SE, McCammon JA. Mapping of allosteric druggable sites in activation-associated conformers of the M2 muscarinic receptor. Chemical biology & drug design. 2014;83(2):237-46. doi: 10.1111/cbdd.12233. PubMed PMID: 24112716; PubMed Central PMCID: PMC4012891.

210. Schnecke V, Kuhn LA. Database screening for HIV protease ligands: the influence of binding-site conformation and representation on ligand selectivity. Proceedings / International Conference on Intelligent Systems for Molecular Biology ; ISMB International Conference on Intelligent Systems for Molecular Biology. 1999:242-51. PubMed PMID: 10786307.

211. Schapira M, Abagyan R, Totrov M. Nuclear hormone receptor targeted virtual screening. Journal of medicinal chemistry. 2003;46(14):3045-59. doi: 10.1021/jm0300173. PubMed PMID: 12825943.

212. Schwede T, Kopp J, Guex N, Peitsch MC. SWISS-MODEL: an automated protein homology-modeling server. Nucleic acids research. 2003;31(13):3381-5.

213. Verdonk ML, Cole JC, Hartshorn MJ, Murray CW, Taylor RD. Improved protein–ligand docking using GOLD. Proteins: Structure, Function, and Bioinformatics. 2003;52(4):609-23.

214. Lindorff‐Larsen K, Piana S, Palmo K, Maragakis P, Klepeis JL, Dror RO, et al. Improved side‐chain torsion potentials for the Amber ff99SB protein force field. Proteins: Structure, Function, and Bioinformatics. 2010;78(8):1950-8.

215. Mark P, Nilsson L. Structure and dynamics of the TIP3P, SPC, and SPC/E water models at 298 K. The Journal of Physical Chemistry A. 2001;105(43):9954-60.

216. Abraham MJ, Murtola T, Schulz R, Páll S, Smith JC, Hess B, et al. GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. SoftwareX. 2015;1:19-25.

217. Berendsen HJ, Postma Jv, van Gunsteren WF, DiNola A, Haak J. Molecular dynamics with coupling to an external bath. The Journal of chemical physics. 1984;81(8):3684-90.

218. Zheng LZ, Lin VC, Mu YG. Exploring Flexibility of Progesterone Receptor Ligand Binding Domain Using Molecular Dynamics. Plos One. 2016;11(11). doi: ARTN e0165824

10.1371/journal.pone.0165824. PubMed PMID: WOS:000387615200034.

219. Walt Svd, Colbert SC, Varoquaux G. The NumPy array: a structure for efficient numerical computation. Computing in Science & Engineering. 2011;13(2):22-30.

220. Levy R, Srinivasan A, Olson W, McCammon J. Quasi‐harmonic method for studying very low frequency modes in proteins. Biopolymers. 1984;23(6):1099-112.

221. Notredame C, Higgins DG, Heringa J. T-coffee: a novel method for fast and accurate multiple sequence alignment1. Journal of molecular biology. 2000;302(1):205-17.

222. Sinha SK, Chakraborty S, Bandyopadhyay S. Thickness of the hydration layer of a protein from molecular dynamics simulation. The Journal of Physical Chemistry B. 2008;112(27):8203-9.

223. Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, et al. UCSF Chimera—a visualization system for exploratory research and analysis. Journal of computational chemistry. 2004;25(13):1605-12.

224. Lusher SJ, Raaijmakers HC, Vu-Pham D, Dechering K, Lam TW, Brown AR, et al. Structural basis for agonism and antagonism for a set of chemically related progesterone receptor modulators. The Journal of biological chemistry. 2011;286(40):35079-86. doi: 10.1074/jbc.M111.273029. PubMed PMID: 21849509; PubMed Central PMCID: PMC3186393.

225. Pace CN, Shirley BA, McNutt M, Gajiwala K. Forces contributing to the conformational stability of proteins. The FASEB journal. 1996;10(1):75-83.

226. Callaway DJ. Solvent‐induced organization: A physical model of folding myoglobin. Proteins: Structure, Function, and Bioinformatics. 1994;20(2):124-38.

227. Dill KA. Dominant forces in protein folding. Biochemistry. 1990;29(31):7133-55.

228. Spolar RS, Ha J-H, Record MT. Hydrophobic effect in protein folding and other noncovalent processes involving proteins. Proceedings of the National Academy of Sciences. 1989;86(21):8382-5.

229. Levy Y, Onuchic JN. Water mediation in protein folding and molecular recognition. Annu Rev Biophys Biomol Struct. 2006;35:389-415.

230. Kauzmann W. Some Factors in the Interpretation of Protein Denaturation1. Advances in protein chemistry. 14: Elsevier; 1959. p. 1-63.

231. Harano Y, Kinoshita M. Large gain in translational entropy of water is a major driving force in protein folding. Chemical physics letters. 2004;399(4-6):342-8.

232. Kinoshita M. Importance of translational entropy of water in biological self-assembly processes like protein folding. International journal of molecular sciences. 2009;10(3):1064-80.

233. Hamelberg D, McCammon JA. Standard free energy of releasing a localized water molecule from the binding pockets of proteins: double-decoupling method. Journal of the American Chemical Society. 2004;126(24):7683-9.

234. Levenson AS, Schafer JIM, Bentrem DJ, Pease KM, Jordan VC. Control of the estrogen-like actions of the tamoxifen–estrogen receptor complex by the surface amino acid at position 351. The Journal of steroid biochemistry and molecular biology. 2001;76(1-5):61-70.

235. Mangelsdorf DJ, Thummel C, Beato M, Herrlich P, Schütz G, Umesono K, et al. The nuclear receptor superfamily: the second decade. Cell. 1995;83(6):835-9.

236. Huang P, Chandra V, Rastinejad F. Structural overview of the nuclear receptor superfamily: insights into physiology and therapeutics. Annual review of physiology. 2010;72:247-72.

237. Stephens PJ, Tarpey PS, Davies H, Van Loo P, Greenman C, Wedge DC, et al. The landscape of cancer genes and mutational processes in breast cancer. Nature. 2012;486(7403):400.

238. Webb P, Nguyen P, Valentine C, Weatherman RV, Scanlan TS, Kushner PJ. An antiestrogen-responsive estrogen receptor-α mutant (D351Y) shows weak AF-2 activity in the presence of tamoxifen. Journal of Biological Chemistry. 2000;275(48):37552-8.

239. Herynk MH, Fuqua SA. Estrogen receptor mutations in human disease. Endocrine reviews. 2004;25(6):869-98.

240. Huang X, Margulis CJ, Berne BJ. Dewetting-induced collapse of hydrophobic particles. Proc Natl Acad Sci U S A. 2003;100(21):11953-8. Epub 2003/09/26. doi: 10.1073/pnas.1934837100. PubMed PMID: 14507993; PubMed Central PMCID: PMCPMC218694.

241. Madauss KP, Deng SJ, Austin RJH, Lambert MH, McLay I, Pritchard J, et al. Progesterone receptor ligand binding pocket flexibility: Crystal structures of the norethindrone and mometasone furoate complexes. J Med Chem. 2004;47(13):3381-7. doi: 10.1021/jm030640n. PubMed PMID: WOS:000221963400006.

242. Thompson SK, Washburn DG, Frazee JS, Madauss KP, Hoang TH, Lapinski L, et al. Rational design of orally-active, pyrrolidine-based progesterone receptor partial agonists. Bioorganic & medicinal chemistry letters. 2009;19(16):4777-80. Epub 2009/07/15. doi: 10.1016/j.bmcl.2009.06.055. PubMed PMID: 19595590.

243. Leonhardt SA, Edwards DP. Mechanism of action of progesterone antagonists. Experimental biology and medicine. 2002;227(11):969-80.

244. Fuhrmann U, Hess-Stumpp H, Cleve A, Neef G, Schwede W, Hoffmann J, et al. Synthesis and biological activity of a novel, highly potent progesterone receptor antagonist. J Med Chem. 2000;43(26):5010-6.

245. Chwalisz K, Larsen L, Mattia-Goldberg C, McCrary K, Edmonds A. Asoprisnil, a novel selective progesterone receptor modulator (SPRM), controls abnormal uterine bleeding in subjects with leiomyomata. Human Reproduction. 2005:556-. PubMed PMID: WOS:000232272500119.

246. Germain P, Iyer J, Zechel C, Gronemeyer H. Co-regulator recruitment and the mechanism of retinoic acid receptor synergy. Nature. 2002;415(6868):187-92. Epub 2002/01/24. doi: 10.1038/415187a. PubMed PMID: 11805839.

247. Chen JD, Evans RM. A transcriptional co-repressor that interacts with nuclear hormone receptors. Nature. 1995;377(6548):454-7. Epub 1995/10/05. doi: 10.1038/377454a0. PubMed PMID: 7566127.

248. Glass CK, Rosenfeld MG. The coregulator exchange in transcriptional functions of nuclear receptors. Genes Dev. 2000;14(2):121-41. Epub 2000/02/01. PubMed PMID: 10652267.

249. Case D, Darden T, Cheatham III T, Simmerling C, Wang J, Duke R, et al. AmberTools 16, University of California, San Francisco, 2016. There is no corresponding record for this reference.

250. da Silva AWS, Vranken WF. ACPYPE-Antechamber python parser interface. BMC research notes. 2012;5(1):367.

251. Kumari R, Kumar R, Lynn A. g_mmpbsa A GROMACS Tool for High-Throughput MM-PBSA Calculations. Journal of chemical information and modeling. 2014;54(7):1951-62.

252. Genheden S, Ryde U. The MM/PBSA and MM/GBSA methods to estimate ligand-binding affinities. Expert opinion on drug discovery. 2015;10(5):449-61.

253. Miller III BR, McGee Jr TD, Swails JM, Homeyer N, Gohlke H, Roitberg AE. MMPBSA. py: an efficient program for end-state free energy calculations. Journal of chemical theory and computation. 2012;8(9):3314-21.

254. Cang Z, Wei GW. Integration of element specific persistent homology and machine learning for protein‐ligand binding affinity prediction. International journal for numerical methods in biomedical engineering. 2018;34(2):e2914.

255. Wang B, Buchman CD, Li L, Hurley TD, Meroueh SO. Enrichment of chemical libraries docked to protein conformational ensembles and application to aldehyde dehydrogenase 2. J Chem Inf Model. 2014;54(7):2105-16. Epub 2014/05/27. doi: 10.1021/ci5002026. PubMed PMID: 24856086; PubMed Central PMCID: PMCPMC4114474.

256. Ding B, Wang J, Li N, Wang W. Characterization of small molecule binding. I. Accurate identification of strong inhibitors in virtual screening. J Chem Inf Model. 2013;53(1):114-22. Epub 2012/12/25. doi: 10.1021/ci300508m. PubMed PMID: 23259763; PubMed Central PMCID: PMCPMC3584174.

257. Ain QU, Aleksandrova A, Roessler FD, Ballester PJ. Machine-learning scoring functions to improve structure-based binding affinity prediction and virtual screening. Wiley Interdiscip Rev Comput Mol Sci. 2015;5(6):405-24. Epub 2016/04/26. doi: 10.1002/wcms.1225. PubMed PMID: 27110292; PubMed Central PMCID: PMCPMC4832270.

258. Vyas R, Bapat S, Jain E, Tambe SS, Karthikeyan M, Kulkarni BD. A Study of Applications of Machine Learning Based Classification Methods for Virtual Screening of Lead Molecules. Comb Chem High Throughput Screen. 2015;18(7):658-72. Epub 2015/07/04. PubMed PMID: 26138573.

259. O'Boyle NM, Banck M, James CA, Morley C, Vandermeersch T, Hutchison GR. Open Babel: An open chemical toolbox. Journal of cheminformatics. 2011;3(1):33.