AI4Science101

DeepModeling Community

Sep 08, 2022

CONTENTS:

CHAPTER

ONE

FIRST EDITION

1.1 Announcing AI for Science Blog Series

1.1.1 Background

With the rapid development of AI, people have started to apply AI methods to almost every field, from natural language processing to computer vision. Recent breakthroughs have demonstrated the power of AI in solving grand challenges in the scientific community. Particular examples include predicting highly accurate protein structures with AlphaFold2, simulating 100 million particle systems with DPMD, imagining the first-ever picture of a black hole, etc. Nevertheless, many researchers in both AI and scientific fields are not able to approach AI for Science research due to many gaps, from limited domain knowledge to the misunderstanding of AI capability. In addition, the educational materials for AI for science are scattered and poorly organized. We announce this initiative (a blog series) to bring people who are interested in AI for Science into the forefront of AI for Science with knowledge collected at different levels, from motivational overview of the field, and lecture-style tutorials on specific topics to knowledge base over common terminologies.

1.1.2 Aim and Scope

We are a group of students, researchers, and practitioners who are interested in AI for science and devoted to advancing AI for science as a new field and community. We write blogs to promote AI for science research at different levels from motivations for new researchers, resources for interdisciplinary researchers, etc. As we announce this AI for science blog series, we release two main documents with titles *AI for Scientific Discovery* and *Scientific Discovery in the era of AI*, which are different views on AI for science from the AI and scientific communities. In addition, we compile a list of common terminologies in different disciplines as a *knowledge base*. As our first *lecture-style tutorial*, we highlight a study of molecular dynamics, one of the most commonly used tools in computational chemistry.

1.1.3 Acknowledgement

The project is a part of the DeepModeling community, an open-source community that aims to define the future of scientific computing together. This effort is primarily led by Yuanqi Du (Cornell), Yingze Wang (UCB), Yanze Wang (PKU), Yibo Wang (DP) and contributors Jiayue Wang (DP), Jiameng Huang (PKU), Arian Jamasb (Cambridge), Jihao Long (Princeton), Guiyu Cao (PKU), Zhenfeng Deng (PKU), Xi Chen (DP), Siyuan Zhou (BFSU), Yinkai Wang (Tufts). We also like to express our gratitude to Weinan E (Princeton & PKU), Linfeng Zhang (DP), Ping Tuo (DP), Zheng Cheng (AISI), Han Wen (DP), Dongdong Wang (DP), Xinming Tu (UW), Nilay Shah (UCLA), Hannes Stark (MIT), Chaitanya Joshi (Cambridge), Ryan-Rhys Griffiths (Cambridge), Sang Truong (Stanford), Junhan Chang (PKU), Chenbing Wang (PKU), Ziming Liu (MIT), Weiliang Luo (PKU), Zhen Wang (DP), Yucheng Zhang (UTokyo), Ferry Hooft (UvA), Ziyao Li (PKU) for providing expertise, feedback and support.

1.1.4 Feedback/comment or Join us

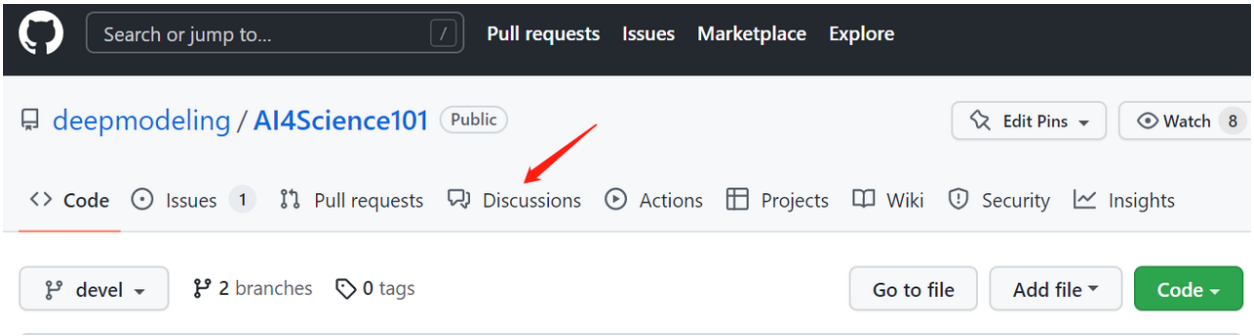
Please reach out to us at ai4science101@deepmodeling.com or join our slack channel if you have any feedback or comments. As this is a community effort, we welcome anyone interested to join us. Any kind of volunteer work is welcomed, including writing tutorials, drawing illustrations, etc. Do not hesitate to let us know!

1.1.5 Contribution Guidelines

We are looking for contributors/experts for specific areas related to AI for Science. The expected contributions include a three-level write-up, a one-paragraph introduction and learning material in section 2 or 3 (depending on the topic in AI or Science), common terminologies and short explanations in section 5, and a specialized chapter similar to section 4. For each specialized chapter, we expect to include (1) target audience and motivations, (2) brief review of literature/history, (3) current advances and future promises, (4) takeaways, and (5) a running sample/demo (optional).

How to get involved

- Github discussion



Welcome everyone to participate in the discussion about AI4Science in the discussion module of our GitHub. The website is here

- Email

Our email address is ai-for-science101@googlegroups.com. If you are interested in sharing your knowledge about any particular aspect of AI for Science (e.g. a common AI tool, practical guidance, an overview of a scientific topic, etc.), we encourage you to send us an email before you start preparing the material.

- Slack group

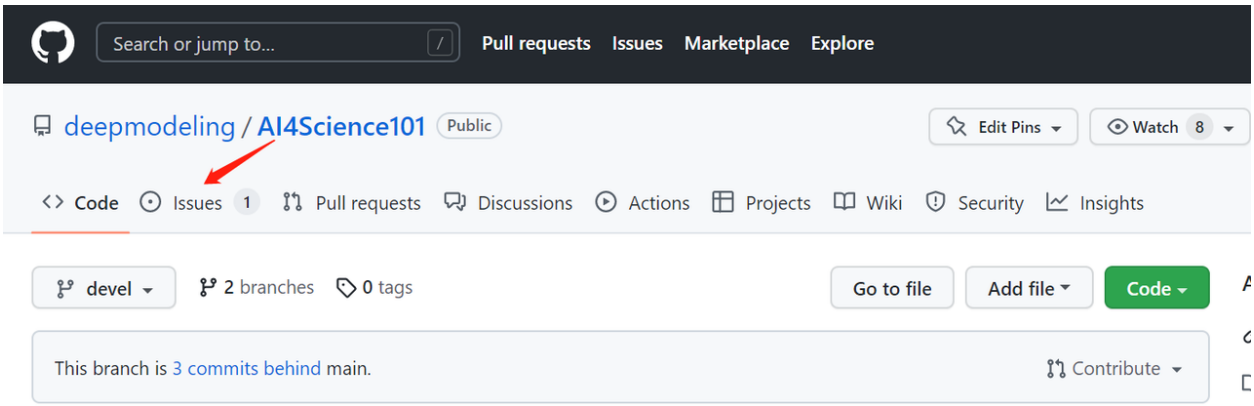
In addition to our reading documents, you can also join the AI4Science101 Slack channel to introduce yourself, drop comments/feedback, discuss related material, network with peers, and contribute new material.

How to make a new request

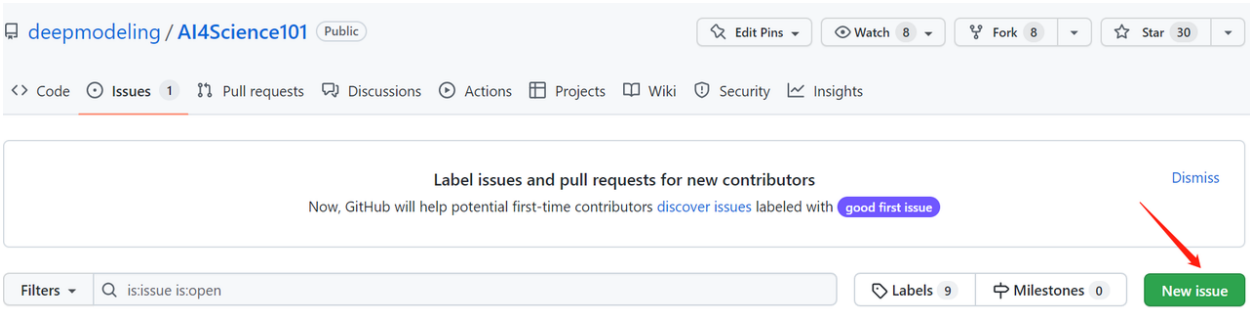
- Make a new issue

If you have any suggestions for any of the documents, have any new requests for material that you are interested in, or you are interested in contributing your knowledge and expertise to this initiative, we encourage you to participate in the AI4Science101 project.

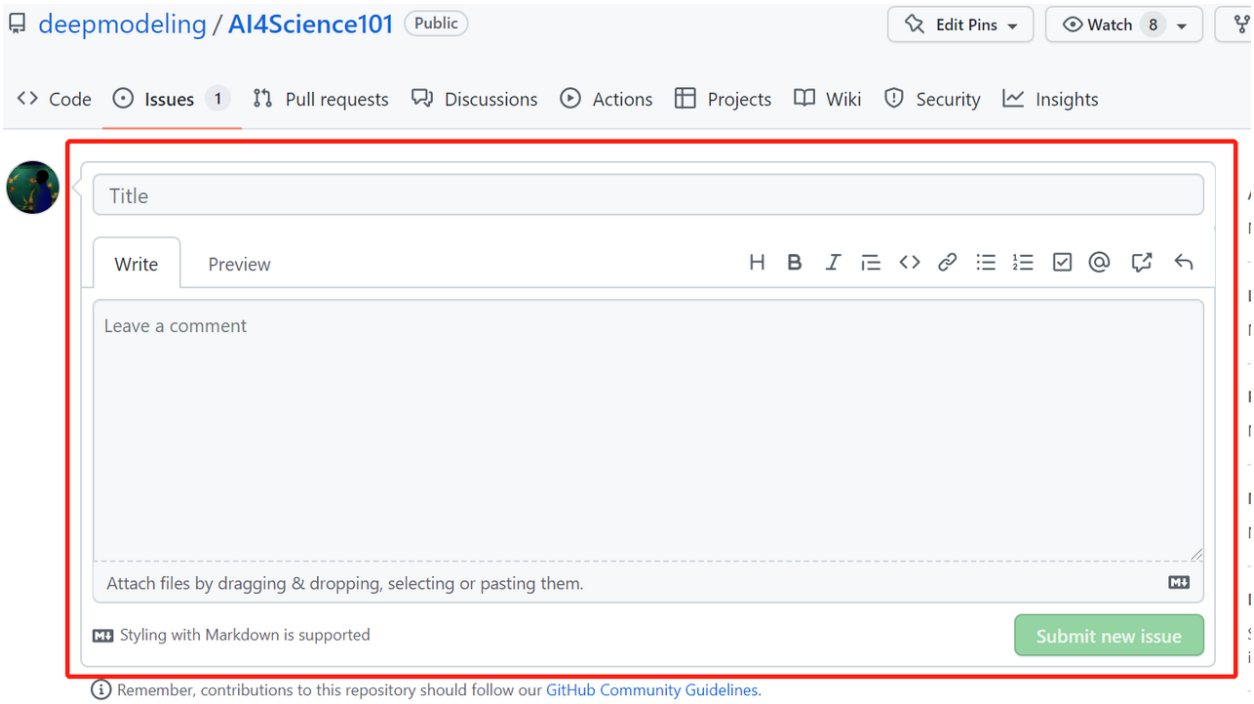
In order to increase the visibility of all requests and comments, and to facilitate the organization of this project, we recommend that you submit a new issue with examples shown below.



Then you can click the button pointed by the red arrow to open an issue.



After this, you can write your issue in the section inside the red box.



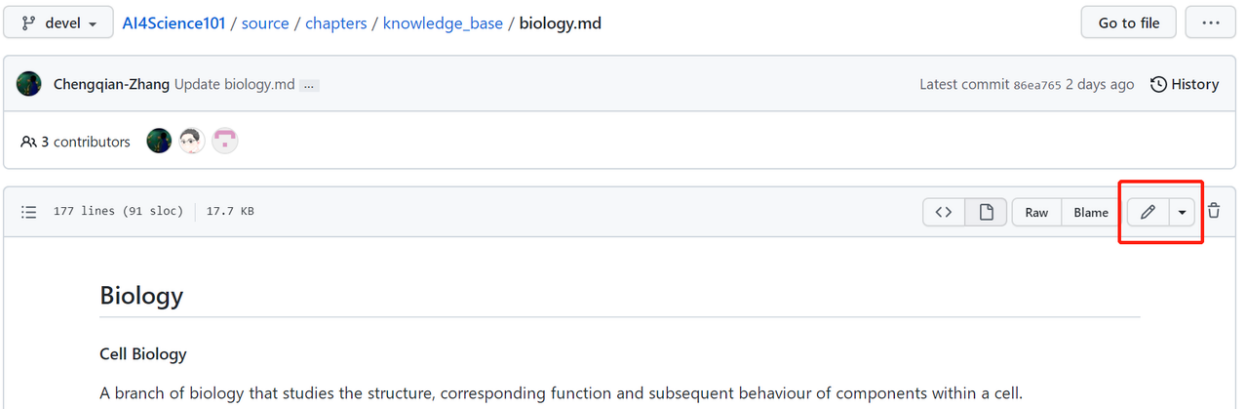
In order to make your request/comments more organized, we hope you help us classify the type of request by stating in the title as one of [error report|question|new material request|new material contribution|others], e.g., [new material request] Protein Structure Prediction Tutorial.

How to make corrections to the docs/pull request

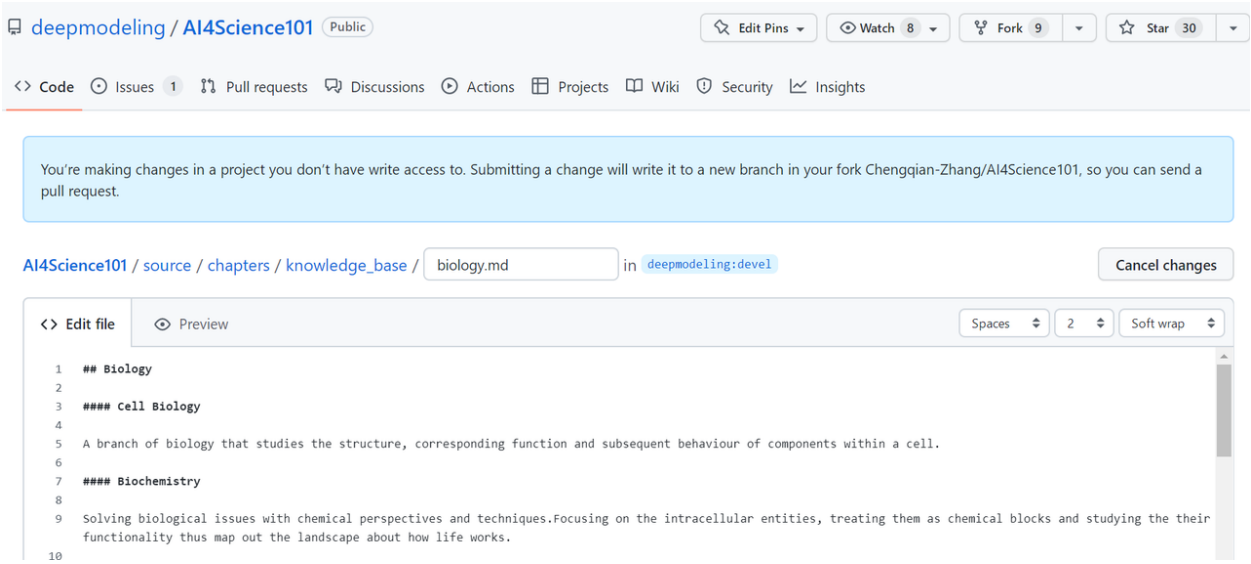
If you find any inaccurate expressions/typos/grammatical errors in our documents or you would like to add more relevant content to our documents, you are welcome to submit a pull request on our GitHub. The community's volunteers will merge your pull request after reviewing your submission.

There are two ways for you to modify the document?

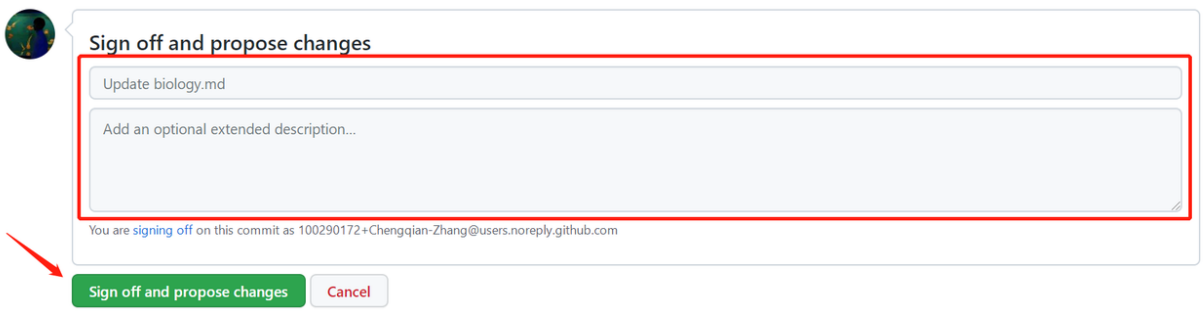
First, if you just want to modify a sentence or a few sentences, you can do it directly in the document,as shown below.



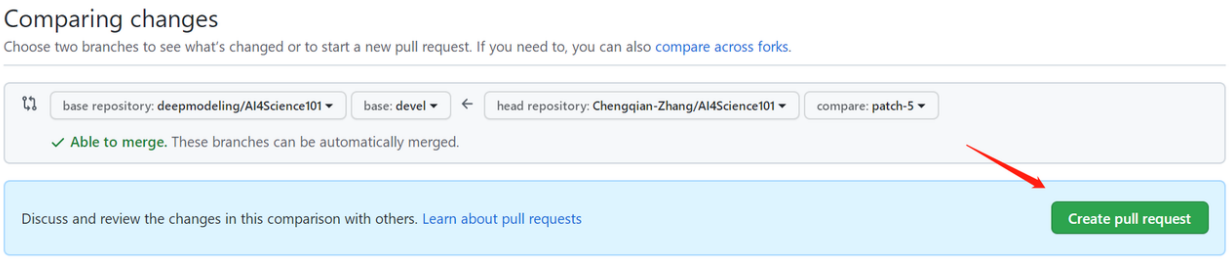
Then find the place you want to modify and modify it.



You can then describe your changes at the bottom of the page and submit them.

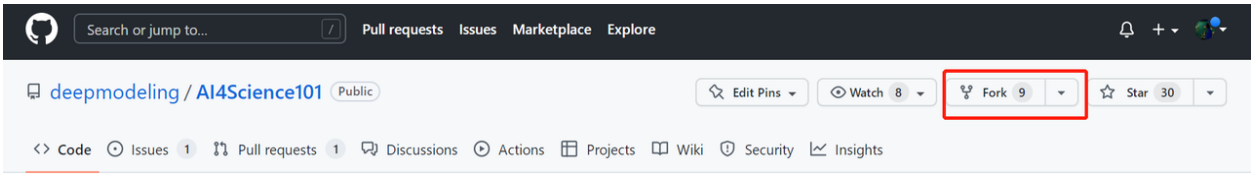


The system will automatically generate a new branch, and you can click the button in the green box to create a new pull request

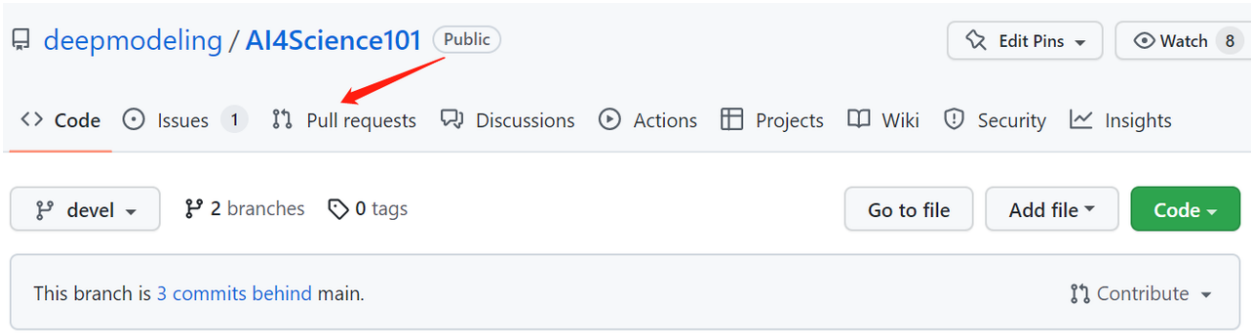


(Note: please commit pr to the **devel branch**)

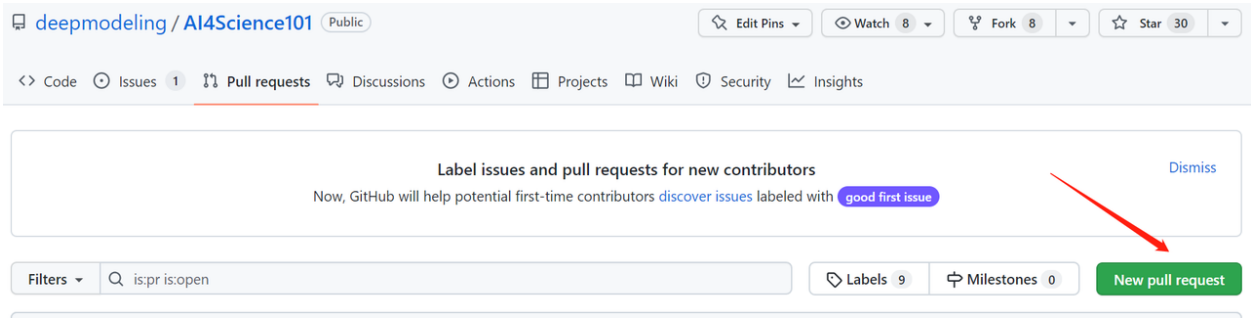
Second, if you want to modify a lot of places, we recommend that you fork to your own repository to modify, and then create a pull request after modifying all the places.



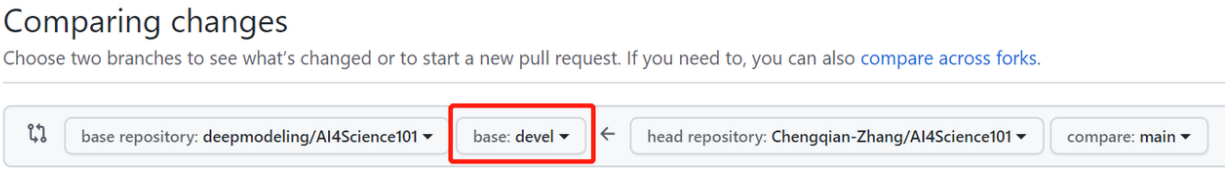
After making changes in your repository, you can create a pull request.



Then you can click the button pointed by the red arrow to open a new pull request.



It is worth noting that you must switch to the devel branch before submitting the pull request. After collecting and sorting out a certain number of changes, we will merge them into the main branch, as shown below. So when submitting a pull request, please change the comparison branch to the devel branch, as shown below.



In addition, due to the problem of markdown format, some formulas of documents on GitHub may appear garbled. Please refer to the content on the project website

1.2 AI for Scientific Discovery

1.2.1 Manifesto

Ever since the time of Isaac Newton, there have been two different paradigms for scientific research: the **Keplerian** paradigm and the **Newtonian** paradigm.

The Keplerian paradigm, often referred to as the "data-driven" approach, expects to extract new physical rules or trends through data analysis and utilize these rules to solve actual problems. The discovery of Kepler's laws of planetary motion was the canonical implementation of this paradigm. Nowadays, many successful examplesbioinformatics and cheminformatics have demonstrated the effectiveness of this paradigm in areas from multi-scale modeling, protein structure prediction, to drug discoveryincluding drug discovery or disease treatment.

The Newtonian paradigm is based on working from first principles, with the aim to figure out fundamental physical rules that govern the world as we know it. Based on these principles, scientists are able to explain most of the experimentally observed phenomenons. One of the most successful theories is quantum mechanics because it almost prepares us with all necessary laws for much of engineering and natural sciences. However, as pointed out by Dirac, "the exact application of these laws leads to equations much too complicated to be solved". The central difficulty is called "the curse of dimensionality", i.e., the problems we are encountered are actually too high-dimensional and cannot be solved efficiently. For a long time, natural scientists have only had limited ability to handle these equations with at most thousands of variables.

BUT things are going to change!

Machine learning, especially deep learning (or generally AI) techniques emerge as effective tools to approximate arbitrary high-dimensional functions as illustrated by its unprecedented success in computer vision (CV) and natural language processing (NLP). In the Newtonian paradigm, AI methods have been applied to incorporate physical laws to solve more much complicated problems or system simulations than toy examples. In the Keplerian paradigm, AI can be directly applied to analyze and learn from data in an end-to-end manner. With the promises of AI in solving real and challenging scientific problems, **"AI for Science" (AI4Science)** has become established as a new term and prevailed in both AI and scientific research communities. In the past few years, successful applications of AI methods have opened up a wide research avenue for both communities, from AlphaFold2 [1] that solves the 50-year-old protein structure prediction puzzle, DeePMD [2] that extending *ab initio* simulation to unprecedentedly large scales, to controlling nuclear reactor with AI agents [3]. The new paradigm of scientific research empowered by AI has been formed, and aforementioned successful examples have paved the way for this new paradigm. However, as scientific discovery has a very broad scope with many different disciplines many grand challenges that are critical to our lives still remain unsolved. Despite the early success, we have to acknowledge that AI for Science is still nascent and requires joint efforts from both AI and scientific communities.

We are living in an era with the opportunity and means to tackle grand challenges in scientific discovery. To facilitate this emergent field and bridge gaps between AI and scientific communities, this blog aims to equip researchers in the AI community with some basic scientific knowledge and an overview of new challenges in scientific discovery, which may appear significantly different from common AI application areas such as computer vision and speech recognition.

1.2.2 Success of AI in Scientific Discovery

Protein Structure Prediction

When it comes to AI for Science, arguably the most famous and successful example of AI-advanced scientific discovery is AlphaFold2 which addresses the problem of accurately predicting protein 3D structures from their sequence, one of the "holy grail" problems in structural biology. The structures of proteins are essential to their biological functions and accurate 3D modeling at the atomistic level is significant for a variety of field from drug discovery to synthetic biology. However, resolving structures experimentally is highly expensive and time-consuming, so computational methods to predict accurate protein structures have long been studied, represented by the series of CASP competitions [4]. Anfinsen's hypothesis, stating that for most proteins, the native structures in standard physiological environments are determined solely by the proteins' amino acid sequences, grounds the study of computational method for the protein structure prediction problem. Traditionally, structural biologists utilize "homology modeling" to make such predictions. In those methods, multiple proteins, whose sequences are similar to the query one and structures have already been resolved experimentally, will be found. These structures will be then be assembled to provide the predicted result. The accuracy of homology modeling depends heavily on sequence identity, and sometimes fails to reach a satisfying level. Deep learning methods, however, have stronger ability to discover correlations between sequences and structures. With carefully-designed attention-based neural networks and multiple-sequence-alignment (MSA) information, in 2020, the AlphaFold2 model achieved an astonishing average RMSD of 0.96 angstroms on the test cases in the prestigious CASP competition. Such accuracy is even comparable to experimental errors and AF2 was considered to make a breakthrough in solving this 50-year structural biological puzzle.

Multi-scale Modeling

As mentioned in the manifesto, the principles of our physical world are almost completely known but the mathematical equations are too complicated to be solved accurately. Therefore, for problems at different time and space scales, scientists have to develop different computational methods with the necessary approximations to reduce the computational complexity which is called multi-scale modeling in computational science. In this area, *ab initio* and classical molecular dynamics (shortened as AIMD and classical MD) are two widely used techniques. However, there has been a longstanding trade-off problem between the accuracy and efficiency of these methods. Specifically, in AIMD, the energy and forces of given systems are calculated by quantum mechanics (usually density functional theory or DFT), thus it is more accurate but more time-consuming. The computational complexity is $O\left(N^3\right)$, where N is the number of electrons in the systems. In classical MD, the potential energy surface is given by fixed mathematical forms with manually curated parameters (named "force fields"), so it is much faster ($O\left(N\right)$) but less accurate. Recently, AI methods have been largely developed to bridge this gap [5,6,7]. Specifically, researchers designed neural networks that preserving necessary physical symmetries and generated a neural representation of atomic configurations which could be used to fit DFT-level potential energy surfaces. However, it still requires large-scale dataset to train such models. Concurrent learning [8] workflows are here to rescue, by heuristically exploring the configuration space to collect as few data as possible. In this manner, construction of neural network potentials becomes more automatic, which further enables a variety of applications in complex systems in condensed matters as well as material science [9].

1.2.3 Why is AI for Science different?

Data is Important

Undoubtedly, AI methods rely on datasets with both high-quality and high-quantity to achieve excellent performance in solving problems, which has been demonstrated by ImageNet [10] , Cifar-10 [11], etc. This is also very true for scientific problems, examplified by RCSB Protein Data Bank (PDB) [12]. This database, containing approximate 200,000 data entries and maintained by researchers all over the world for over 40 years, is one of the best database for experimentally resolved biomolecular structures. Deep learning methods would never reach such a success without efforts from maintainers of PDB. In particular, scientific problems usually involve real and challenging scenarios for many AI-interested topics, e.g., out-of-distribution generalization, low-data regime learning, etc. And it is often the case that in many topics,

no high-quality dataset is available to generate effective deep learning models. Therefore, it is encouraged that AI researchers to pay more attention to data collection and structuralization, during which domain expertise and joint efforts are required.

Problem Formulation

Many scientific discovery problems are much more complex than simple classification or regression tasks. For AI researchers, scientific problems have to be decomposed and well-formulated to an extent that inputs, outputs, and objective functions (often needs to be differentiable) can be clearly defined. For example, "drug design" is a huge pipeline which consists of a sequence of steps and is obviously not a well-formulated problem itself. Instead, we can decompose this problem into different pieces that are AI-solvable: molecular property prediction for virtual screening molecular databases, molecular generation for proposing better drug candidates, etc. Problems failed to be compliant with this standard are often referred to as "dirty" ones, and are unlikely to be addressed solely with AI methods.

1.2.4 Real-world Challenges in AI for Science

Next Steps in Protein Structure Prediction

Represented by AlphaFold2 [1], a variety of AI-based protein structure prediction models [13,14] successfully solve the protein structure prediction problem, but it is only the first step towards understanding protein structures and functions. There are still many remaining challenges to be solved:

Protein Multimers Current predictive models of protein structure can only provide reliable results for monomers (single peptide chain). But in reality, peptide chains can interact with each other and form complexes (multimers). In many scenarios, only by doing so can the proteins perform their biological functions correctly. In structural biology, such behavior is defined as quarternary protein structure.

Protein-ligand Complex Protein-ligand interactions and the induced-fit models are key to understanding drugs' potency. Small organic molecules often interact with a certain area (referred to as a pocket) in target proteins and may cause the protein structure to change significantly. Traditional computational methods, such as molecular docking, model the protein-ligand binding free energies with physical-based scoring functions, which are parametrized in an empirical and error-prune way. AI models will be a breakthrough in this area if accurate prediction of the ligand binding pose, and/or the protein structural changes during the ligand binding process can be made.

Protein Conformation Ensembles Most of the recent successful models are based on multiple sequence alignment (MSA), which can be viewed as an augmented version of "homologous modeling". The scientific logic behind this is that proteins follow rules of evolution, so more or less, any protein found naturally is subject to have some structural similarities with those proteins in other organisms which some have been studied before. However, there are still a variety of proteins that are de novo-designed (manually designed) or lacking MSA, current models fail to provide reliable prediction results. Thus one promising direction of AI-based protein structure prediction may be the development of MSA-free models.

Quantum Mechanics

One of the central goals in quantum mechanics is to find accurate solutions (wave functions and energies) to Schrödinger equations on real systems, which is hindered by many-body problems because the dimensionality of the equation is $3N$, where N is the number of electrons (and a real system can easily have hundreds of electrons, where the many-body Schrödinger equations can not be solved exactly. To compromise, researchers have come up with many approximation methods, such as DFT (density functional theory), to make the computational cost acceptable by sacrificing some accuracy. This work have reached great achievements in many areas such as material science, but in cases where the DFT results are not accurate enough, researchers have to rely on more accurate but more time-consuming methods (CCSD(T), with a computation complexity of $O\left(N^7\right)$). Recent work, such as DeePKS [15] and DM21 [16] have been proposed to tackle this issue with AI models, but are still far from perfect. One particular challenge is how to represent an anti-symmetric function under permutation in a neural-network manner (wave functions of electrons are of this property).

Molecular Dynamics

Molecular dynamics (MD) is a computer simulation method for analyzing the physical movements of atoms and molecules. The atoms and molecules are allowed to interact for a fixed period of time, providing a view of the dynamic "evolution" of the system. The trajectory can be considered as a sample under the Boltzmann distribution of a given system and temperature. Thus, many thermodynamic properties such as density and free energy can be calculated by MD.

General neural-network-based force field Although deep learning methods have already shown their capabilities in accelerating AIMD, a neural network potential able to be generalized to different systems and simulation settings is of high practical value. This could be achieved by pre-training treatment and thus repeated work can be avoided, as users will no longer need to establish a model from scratch, but fine-tune the pre-trained models against specific systems instead. For example, a model describing arbitrary organic molecules at a very accurate quantum mechanics level will be useful in drug design, and a model describing any components of alloy/materials is valued in material science. Besides, the requirement of higher transferability also challenge current methods with more generalizable representation of atomic configuration, which further brings demand to architecture enhancement.

Coarse-grained models Simulation of extremely large and complicated systems, such as a whole virus, needs coarse-grained force fields that treat several atoms as one "bead". Then the interactions between these beads are expected to reflect certain properties of interest, such as free energy or conformation distribution. However, it is nontrivial to find optimal forms and parameters to describe such interactions, and currently there are no general protocols like empirical atomistic force fields. AI models may be an effective tool just as they are between DFT/AIMD and classical MD, but more research need to be conducted to answer questions including what targets to fit, how to generate training data efficiently.

Enhanced sampling Enhanced sampling assists to overcome free energies barriers in a molecular dynamic simulation. If the free energy landscape of a given system is not smooth, the simulation will be stuck in one local minimal and ergodicity in molecular dynamics simulation will not be satisfied. This phenomenon is manifested by inadequate sampling over the whole landscape, especially over transition states or other local minima, and occurs frequently in simulation of biological systems. Computational chemists have employed bias-potential-based techniques (such as meta-dynamics [17], and umbrella sampling [18]) to enhance sample efficiency. But these methods require well-defined collective variables (CVs) and fail to handle situations where the number of CVs is large. The key challenge is how to learn an accurate representation of the free energy surface (FES) with high-dimensional CVs. AI models have recently been introduced, e.g., NN-VES [18], Reinforced Dynamics [19], and NN-based CV selections [20,21]. The main challenges lie in better models with generalizability and more effective workflows to take training data generation into consideration [22].

Partial differential equations

High dimensional partial differential equations (PDEs) arise in many scientific problems. Notable examples include high dimensional nonlinear Black-Scholes equations in finance, many electronic Schrödinger equations in quantum mechanics, and high dimensional Hamilton-Jacobi-Bellman equations in control theory. However, traditional numerical algorithms like finite difference or finite element methods suffer from the curse of dimensionality and are unable to deal with PDEs beyond 10 dimensions. The practical success of deep-learning-based PDE solvers such as physics-informed neural networks and deep BSDE method shows the ability of the deep neural networks to efficiently approximate the solutions of high dimensional PDEs. Hence, once we can reformulate the PDE by a variational problem, deep learning techniques can be easily applied to the variational problem and the original PDE can be solved. Successful examples in this direction include the deep Ritz method[23] the deep BSDE method[1] , and Physics-informed neural networks[24]

- **Variational problem:** Find the maxima or minima of a functional, which maps functions to scalars, over a given domain.
- **Finite difference method:** A class of numerical algorithms to solve the differential equations. It approximates the derivative or partial derivative by finite differences and solves the resulting linear or nonlinear systems.
- **Finite element method:** A class of numerical algorithms to solve the differential equations. It converts the differential equations to a variational problem, uses a finite-dimensional linear space to approximate the domain of the variational problem, and solves the variational problem over the finite-dimensional linear space.

Control theory

Control algorithms are widely used in engineering and industry, which aim to govern the application of system inputs to drive the dynamic system to satisfying specific conditions. Since the time of Bellman [25] , a long-lasting problem in control theory is to solve the high dimensional closed-loop control problems, which aims to find the policy function: the input as a function of the state. Indeed, the terminology "curse of dimensionality" was originally coined by Bellman in order to highlight these difficulties. The practical success of deep learning shows that deep neural networks can approximate high dimensional functions and hence raise the hope to solve high dimensional closed-loop control problems. Although this field is still immature and faces many challenges such as stability and robustness of policy function, pioneering works [24,26] show the potential of this field. Another related field is reinforcement learning. Roughly speaking, control algorithms and reinforcement learning problems solve the same problems. However, in contrast, to control algorithms, which make heavy use of the underlying models, reinforcement learning algorithms make minimum use of the model. Comparison and combination of control algorithms and reinforcement learning algorithms are interesting topics and helpful if one wants to deal with complex practical problems.

- **Reinforcement learning:** Reinforcement learning concerns how an agent takes actions to maximize the long-term reward when faced with an unknown environment. One feature of the reinforcement learning algorithm is that it does not require the exact form of the underlying model.

Fluid Mechanics

Fluid mechanics studies the systems with fluid (liquids, gases, and plasmas) at rest and in motion [27,28,29]. Many scientific and engineering disciplines get involved with fluid mechanics (as shown in Figure above), including astrophysics, oceanography, meteorology, aerospace engineering, chip industry, and physics-based animation. Overall, the fluid mechanics can be roughly divided into inviscid flows vs. viscous flows, laminar flows vs. turbulence, incompressible flows vs. compressible flows, continuum flows vs. rarefied flows, single-phase flows vs. multiphase flows, Newtonian flows vs. non-Newtonian flows, etc.

Mathematical analysis, experimental studies, and numerical simulations are three major approaches to exploring fluid mechanics. Fundamentally, a fluid system is assumed to be governed by mathematical equations in the conservation of mass, momentum, and energy. In different physical modeling scales, the governing equations of fluid are in different forms [30], the newton dynamics, Boltzmann equation, Euler or Navier-Stokes equations (NSE), and coarse-grained turbulence models. In the hierarchy of governing equations, the hyperbolic Euler equations for inviscid flows are usually utilized to validate the performance of the numerical scheme of its accuracy, efficiency, and robustness. Additionally, the NSE is widely used in continuum viscous fluid mechanics, while the Boltzmann equation works well in rarefied gas dynamics. With the rapid growth of high-performance computing, numerical simulation called computational fluid dynamics (CFD) not only gradually becomes the indispensable tool to validate the key mathematical conclusions and experimental observations in fluid dynamics, but also provides more abundant and practical fluid information (macroscopic velocities, pressure and temperature distribution, drag and lift force, heat load, noise level) for engineering applications. With the aid of AI methods, research on numerical and experimental fluid mechanics may be improved.

- **Design data-driven turbulence models**, such as modeling high-Reynolds number wall-bounded turbulent flows and complex separated turbulent flows [31] (i.e., the simulation and design in advanced aircraft).
- **Conduct data assimilation in flow fields**, which combines the sparse measured data and numerical solutions together to provide more complete and accurate flow fields (i.e., the prediction of ocean circulation, weather forecast, and city environment simulations).
- **Refresh multiphase and multiscale fluid models**, which modify the ad-hoc models of turbulence combustion, multiphase flows, and rarefied gas dynamics (i.e., efficient moment closure models for simulating rarefied flows).

1.2.5 See the opportunities - Why AI for Science?

- **Challenging scenarios.** For AI algorithms, scientific applications are usually much more challenging, compared to common applications in images, texts, or audio, where "rules" are defined by humans. Science is about finding and understanding the nature, so it is usually more challenging.
- **Real-world impacts.** Scientific discovery works for the good of human beings. For example, boosting drug design will reduce the price of drugs and save more people's lives.
- **New discovery.** Curiosity is the nature of human beings that motivates the development of science. As Kepler derived the laws of planetary motions hundreds of years ago, we are in an era in which AI may help us to discover new science systematically.

1.2.6 Mind your steps

- **Be careful with data.** Datasets in scientific problems have many problems: it may be highly-screwed, with 99% positive cases and only 1% negative cases, because researchers will not publish their bad results; it may be very small, because much data is hard to generate and collect; it may be very dirty, for example, some experimental results are noisy and not reliable.
- **Understand the problems.** Scientific concepts are not as easy to understand as classifying cats and dogs in computer vision. Take a humble and respectful manner toward scientific problems and learn more scientific backgrounds (physics, chemistry, biology, etc.) about the problem of interest. Understand the reason for solving the problem and the practical application of the research are the key to success.
- **Be patient.** "Rome was not built in a day." Scientific problems are often challenging and taking years to solve. But don't be afraid if you miss any of the deadlines for NeurIPS/ICML/ICLR, good work will be recognized and published and become impactful eventually.
- **Enjoy interdisciplinary collaborations.** Good collaborations between AI and Science communities are key to make impactful work both in terms of real-world challenges and methodologies. Be open-minded while talking to people from the other community.

1.2.7 A Roadmap of Basic Scientific Knowledge

Classical Mechanics

- Kibble, Tom, and Frank H. Berkshire. Classical mechanics. world scientific publishing company, 2004.Classical Mechanics

Statistic Mechanics:

- Tuckerman, Mark. Statistical mMechanics: tTheory and mMolecular sSimulation. Oxford university press, 2010. Mark E. Tuckerman
- Pathria, Raj Kumar.Statistical Statistical mMechanics. Elsevier (R. K. Pathria, Paul D. Beale, 2016.1)

Quantum Mechanics:

- A. Szabo, A., and N. S. Ostlund. ", Modern Quantum Chemistry (Dover." New York (, 1996).
- L. Piela, Lucjan. Ideas of qQuantum cChemistry, 2nd Ed. (Elsevier, 2006.14)
- Sholl, David S., and Janice A. Steckel. Density functional theory: aA practical introduction. John Wiley & Sons (David Sholl, Janice A Steckel, 2011.09)

Solid State Physics:

- Kittel, Charles. "Introduction to Solid State Physics Solution Manual." (2021). (Charles Kittel, 2004)

Multi-scale Modeling:

- Weinan, E. Principles of mMultiscale mModeling. Cambridge University Press (Weinan, E, 2011.)

Control Theory:

- Evans, Lawrence C. "An introduction to mathematical optimal control theory version 0.2." Lecture notes available at http://math. berkeley. edu/~ evans/control. course. pdf ((L.C. Evans, 1983).

Partial Differential Equations:

- Evans, Lawrence C. Partial dDifferential eEquations. Vol. 19. American Mathematical Soc. (L. C. Evans, 2010.)

Fluid Dynamics:

- Kundu, Pijush K., Ira M. Cohen, and David R. Dowling. Fluid mechanics. Academic pressFluid mechanics (P. K. Kundu, I. M. Cohen & D. R. Dowling, 2015.5)

1.2.8 References

[1] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A A Kohl, Andrew J Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. Highly accurate protein structure prediction with AlphaFold. Nature, 596(7873):583–589, August 2021.

[2] Linfeng Zhang, Jiequn Han, Han Wang, Roberto Car, and Weinan. Deep potential molecular dynamics: A scalable model with the accuracy of quantum mechanics. Phys. Rev. Lett., 120(14), April 2018.

[3] Jonas Degrave, Federico Felici, Jonas Buchli, Michael Neunert, Brendan Tracey, Francesco Carpanese, Timo Ewalds, Roland Hafner, Abbas Abdolmaleki, Diego de las Casas, Craig Donner, Leslie Fritz, Cristian Galperti, Andrea Huber, James Keeling, Maria Tsimpoukelli, Jackie Kay, Antoine Merle, Jean-Marc Moret, Seb Noury, Federico Pesamosca, David Pfau, Olivier Sauter, Cristian Sommariva, Stefano Coda, Basil Duval, Ambrogio Fasoli, Pushmeet Kohli, Koray Kavukcuoglu, Demis Hassabis, and Martin Riedmiller. Magnetic control of tokamak plasmas through deep reinforcement learning. Nature, 602(7897):414–419, February 2022.

[4] CASP competitions 2022. https://predictioncenter.org/

[5] Justin S Smith, Olexandr Isayev, and Adrian E Roitberg. Ani-1: an extensible neural network potential with dft accuracy at force field computational cost. Chemical science, 8(4):3192–3203, 2017.

[6] Linfeng Zhang, Jiequn Han, Han Wang, Wissam Saidi, Roberto Car, and Weinan E. End-to-end symmetry preserving inter-atomic potential energy model for finite and extended systems. Advances in Neural Information Processing Systems, 31, 2018.

[7] Oliver T Unke and Markus Meuwly. Physnet: A neural network for predicting energies, forces, dipole moments, and partial charges. Journal of chemical theory and computation, 15(6):3678–3693, 2019.

[8] Yuzhi Zhang, Haidi Wang, Weijie Chen, Jinzhe Zeng, Linfeng Zhang, Han Wang, and Weinan. DP-GEN: A concurrent learning platform for the generation of reliable deep learning based potential energy models. Comput. Phys. Commun., 253(107206):107206, August 2020.

[9] Tongqi Wen, Linfeng Zhang, Han Wang, E Weinan, and David J Srolovitz. Deep potentials for materials science. Materials Futures, 2022.

[10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In 2009 IEEE conference on computer vision and pattern recognition, pages 248–255. Ieee, 2009.

[11] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[12] Helen M Berman, John Westbrook, Zukang Feng, Gary Gilliland, Talapady N Bhat, Helge Weissig, Ilya N Shindyalov, and Philip E Bourne. The protein data bank. Nucleic acids research, 28(1):235–242, 2000.

[13] Minkyung Baek, Frank DiMaio, Ivan Anishchenko, Justas Dauparas, Sergey Ovchinnikov, Gyu Rie Lee, Jue Wang, Qian Cong, Lisa N Kinch, R Dustin Schaeffer, et al. Accurate prediction of protein structures and interactions using a three-track neural network. Science, 373(6557):871–876, 2021. 52

[14] Jian Peng and Jinbo Xu. Raptorx: exploiting structure information for protein alignment by statistical inference. Proteins: Structure, Function, and Bioinformatics, 79(S10):161–171, 2011.

[15] Yixiao Chen, Linfeng Zhang, Han Wang, and Weinan E. Deepks: A comprehensive data-driven approach toward chemically accurate density functional theory. Journal of Chemical Theory and Computation, 17(1):170–181, 2021. PMID: 33296197.

[16] James Kirkpatrick, Brendan McMorrow, David H. P. Turban, Alexander L. Gaunt, James S. Spencer, Alexander G. D. G. Matthews, Annette Obika, Louis Thiry, Meire Fortunato, David Pfau, Lara Román Castellanos, Stig Petersen, Alexander W. R. Nelson, Pushmeet Kohli, Paula Mori-Sánchez, Demis Hassabis, and Aron J. Cohen. Pushing the frontiers of density functionals by solving the fractional electron problem. Science, 374(6573):1385–1389, 2021.

[17] Alessandro Laio and Michele Parrinello. Escaping free-energy minima. Proceedings of the National Academy of Sciences, 99(20):12562–12566, 2002.

[18] Glenn M Torrie and John P Valleau. Nonphysical sampling distributions in monte carlo free-energy estimation: Umbrella sampling. Journal of Computational Physics, 23(2):187–199, 1977.

[19] Luigi Bonati, Yue-Yu Zhang, and Michele Parrinello. Neural networks-based variationally enhanced sampling. Proceedings of the National Academy of Sciences, 116:201907975, 08 2019.

[20] Linfeng Zhang, Han Wang, and Weinan E. Reinforced dynamics for enhanced sampling in large atomic and molecular systems. The Journal of chemical physics, 148(12):124113, 2018.

[21] Dongdong Wang, Yanze Wang, Junhan Chang, Linfeng Zhang, Han Wang, et al. Efficient sampling of high-dimensional free energy landscapes using adaptive reinforced dynamics. Nature Computational Science, 2(1):20–29, 2022.

[22] Luigi Bonati, Valerio Rizzi, and Michele Parrinello. Data-driven collective variables for enhanced sampling. The Journal of Physical Chemistry Letters, 11:2998–3004, 04 2020.

[23] Andrew W Senior, Richard Evans, John Jumper, James Kirkpatrick, Laurent Sifre, Tim Green, Chongli Qin, Augustin Žídek, Alexander W R Nelson, Alex Bridgland, Hugo Penedones, Stig Petersen, Karen Simonyan, Steve Crossan, Pushmeet Kohli, David T Jones, David Silver, Koray Kavukcuoglu, and Demis Hassabis. Improved protein structure prediction using potentials from deep learning. Nature, 577(7792):706–710, January 2020.

[24] Richard Evans, Michael O'Neill, Alexander Pritzel, Natasha Antropova, Andrew Senior, Tim Green, Augustin Žídek, Russ Bates, Sam Blackwell, Jason Yim, Olaf Ronneberger, Sebastian Bodenstein, Michal Zielinski, Alex Bridgland, Anna Potapenko, Andrew Cowie, Kathryn Tunyasuvunakool, Rishub Jain, Ellen Clancy, Pushmeet Kohli, John Jumper, and Demis Hassabis. Protein complex prediction with AlphaFold-Multimer. October 2021.

[25] E Weinan. The dawning of a new era in applied mathematics. Notices of the, volume 68. American Mathematical Society, 2021.

[26] Linfeng Zhang, Jiequn Han, Han Wang, Wissam A Saidi, Roberto Car, and Weinan E. End-to-end symmetry preserving inter-atomic potential energy model for finite and extended systems. May 2018.

[27] Pijush K Kundu, Ira M Cohen, and Howard H Hu. Fluid Mechanics. Academic Press Inc. (London), London, England, 3 edition, December 2004. 53

[28] Frank M White. Viscous fluid flow (int'l ed). McGraw-Hill Professional, New York, NY, 3 edition, April 2005.

[29] David J Acheson. Elementary fluid dynamics. Clarendon Press, 2009.

[30] Kun Xu. Direct modeling for computational fluid dynamics: Construction and application of unified gas-kinetic schemes. Advances In Computational Fluid Dynamics. World Scientific Publishing, Singapore, Singapore, March 2015.

[31] Karthik Duraisamy, Gianluca Iaccarino, and Heng Xiao. Turbulence modeling in the age of data. March 2018.

1.3 Scientific Discovery in the era of AI

1.3.1 Manifesto

In recent years, AI has almost been everywhere, from our daily life to life-critical systems. AI for science is a new terminology representing a growing community that attracts more and more people from both AI and science communities to work on scientific discovery with AI. You may wonder what AI really is? How AI for science works? How is it related to my daily work? In this blog, we introduce AI to people who are interested in AI for science (especially from the view of the scientific community) and answer the above questions, including what AI brings to the scientific community, successful examples of AI in scientific applications, AI mindsets in tackling different types of scientific problems.

1.3.2 News you may hear about AI

- DeepBlue [1] is the first computer chess player to win a game, and the first to win a match, against a reigning world champion under regular time controls. Deep Blue's victory was considered a milestone in the history of artificial intelligence.
- AlphaGo [2] is a computer Go player that defeats a professional human Go player, and defeats a Go world champion for the first time.
- AlphaFold2 [3] provides a solution to a 50-year-old grand challenge in biology, determining protein structure given its sequence.
- DALL·E 2 (2022) [4] is one of the largest AI systems that can create realistic images and art from a description in human-readable language.

1.3.3 What is Artificial Intelligence?

Many words describe areas closely related to AI, sometimes they could be called AI generally, but we illustrate their relationships below. (examples shown in each level are disjoint examples from the overlapping subjects):

- **Artificial Intelligence** is a generic word that represents intelligence demonstrated by machines which include a broad set of methods from traditional reasoning and planning methods to modern machine learning approaches.
- **Machine Learning** is the study of computer algorithms that can improve automatically through experience and by the use of data. It is seen as a part of artificial intelligence.
- **Deep Learning** is one type of machine learning methods that leverages artificial neural networks with back propagation for representation learning.
- **Statistics** is the discipline that concerns the collection, organization, analysis, interpretation, and presentation of data, which shares great overlaps with AI in terms of methodologies.
- **Data Mining** is a process of extracting and discovering patterns in large data sets involving methods at the intersection of AI, statistics, and database systems (access and storage of data).

1.3.4 What does AI bring to the scientific Community?

Traditional Discovery Process in Science (a day of a scientist)

- Goal - Drug design (AI can help in different phases of drug design)
- Hypothesis - Hand-crafted design (AI can help search the vast chemical space and propose drug candidates)
- Simulation - Computer simulation (AI can accelerate and improve the accuracy of computer simulation)
- Experiment - lab experiment (AI can guide lab experiment design)
- Result analysis - Statistical analysis (AI can analyze high-dimensional data)
- Repeat
- Conclusion & Finding

Goals	Traditional Methods	AI-Based Methods
Drug Design	- Error-prone and intuition-based workflows	- AI aided rational drug design
Virtual Screening	- Docking (based on physical scoring functions) - 2D/3D/based on pharmacophore similarity search	- Neural network-based scoring functions - Generative models
Lead Optimization	- MD-based free energy calculation with empirical force field - Wet-lab experiment	- AI-predicted binding- Simulation with more accurate force fields developed with neural networks affinity
Drug Synthesize	- Error-prone experiments to find optimal synthesis route and reaction conditions	- Retro-synthesis analysis by AI models - AI predicted reaction outcomes- Automated wet-Lab experiment
ADMET	- Experiments	- AI-based prediction models

Even decades ago, AI was widely used in the scientific community

- Example: principal component analysis/PCA, linear regression, Kalman Filter, clustering algorithms, etc.

Data analysis empowered by modern AI

- **Why has AI become so popular recently? (What changed the game?)**
  - **Accumulated Big Data**
  - **Advanced Algorithms (especially Rising of Deep Learning)**
  - **Improved Computing Power and Storage**
- **What is data?** Data are individual facts, statistics, or items of information, often numeric. In a more technical sense, data are a set of values of qualitative or quantitative variables about one or more persons or objects.
- **What is learning?** Learning refers to machine learning, which is the study of computer algorithms that improve automatically through experience.
- **Why do we learn from data?** Learning from accumulated data enables us to analyze data and execute certain tasks.
- **What are common tasks that could be solved by AI?**
  - **Predictive tasks** refer to predicting the value or status of something of interest.

    * Example: predict whether an image is a cat or dog.
  - **Generative tasks** refer to generating new data by learning from existing data.
    * Example: generate new drug-like molecules.
  - **Decision-making tasks** refer to making decisions based on the information provided.
    * Example: decide on trading strategies for the stock market.

- **How do we learn from data?** Two essential components of learning from data are (1) data and (2) learning.
  - **Data** includes two main components, data points, and labels, data points refer to the single instances of facts, statistics, or items of information, while labels are meaningful and informative tags for the data points. (Note the labels are often expensive to obtain, thus most of the data are unlabelled)
    * Data Points:(X) cab be in any format, text, image, graph, etc.

      - Example: images of cats and dogs
    * Labels: (Y) are informative tags

      - Example: cat or dog
  - **Learning** essentially involves three main broad categories, **Supervised Learning**, **Unsupervised Learning, and Reinforcement Learning**. We include an additional learning diagram, **Active Learning**, which is commonly used in scientific discovery.
    * **Supervised Learning** learns with labeled data, usually, for predictive tasks, a special case is Semi-supervised Learning where only partial data are labeled, since labels allow us to directly assess the predictive performance of a machine learning model in certain circumstances.

      - Example: predicting molecular properties from molecular structures
    * **Unsupervised Learning** learns with unlabeled data and discovers patterns from data, usually for clustering, dimensionality reduction, and visualization tasks. Another rising topic, Self-supervised Learning, also requires no labeled data by training models to predict "missing" or masked parts of the input, and it focuses more on predictive tasks similar to supervised learning.

      - Example 1: clustering galaxy images with similar patterns

      - Example 2: clustering molecule conformations with similar patterns which reduces the workloads for MD analysis
    * **Reinforcement Learning** concerns how intelligent agents ought to take actions in an environment in order to maximize the notion of cumulative reward.

      - Example: AI agent for nuclear fusion reactor control
    * **Active Learning** is a learning algorithm that can request labels (or propose experiments) that provide information that be most useful for it to improve predictive performance. It is also referred to as optimal experimental design.

      - Example: uncertainty estimation to guide the experiment/data collection

- **How do we collect or represent data?**
  - **Representation Learning** is a set of techniques that allows a system to automatically discover the representations needed for feature detection or classification from raw data. This replaces manual feature engineering and allows a machine to both learn the features and use them to perform a specific task.
    * **Tabular Representation (columns are features)**

      - Example: weather data stored in tables

    * **Grid Representation (stored in grids)**

      - Example: cell images

    * **Sequence Representation (stored in sequences)**

      - Example: genes

    * **Graph Representation (stored in nodes and edges)**

      - Example: molecular graphs

    * **Geometry Representation (stored in 3D geometries)**

      - Example: protein structures

- **Could we generate new data?**
  - **Generative Modeling** learns the distributions of observed samples and generates unseen samples.
    * Example: goal-oriented molecule generation, protein conformations sampling

1.3.5 Mindsets for AI

Key abilities of AI

- **Automated feature learning.** Instead of traditionally manual design of features for various tasks, AI takes data in raw formats and automatically learns the features while optimizing the task objectives.
- **Learning from big data**. AI can learn from the accumulated "big" data in many domains that traditional methods are not capable of.
- **Inductive bias** (e.g. symmetry preservation). AI models are flexible and can be designed to respect natural laws such as symmetries.
- **Generalizability** (to unseen data). After training the AI model, it is expected to generalize to new scenarios and unseen data. In some scenarios, it is also expected to generalize to new dataset or similar task after training with one general dataset or task.
- **Fit high-dimensional function**. AI models can fit complex functions, such as free energy surface, Schrodinger equation, etc.
- **Differentiable programming**. AI brings a new wave of differentiable programming and pushes forward the development of many tools for automatic differentiation, such as PyTorch, Tensorflow, Theano, etc.

Limitations of AI

- **Overfitting**. AI models sometimes overfit into the given data set which hinders their abilities to generalize to other datasets.
- **Data requirement.** No free lunch, AI models usually rely on large-scale datasets.
- **Computational cost.** AI models usually consume plenty of computational resources, especially with the growth of the model and data sizes.
- **Explainability.** AI models usually have poor explainability and are thus considered "black-boxes", though it is an active area of research.

1.3.6 How does AI work?

Typical pipeline

- **Problem Formulation** is an essential step to formulate a problem in a "machine learning language"
  - What is the input? What is the output? What is the task?
    - Input: images of cats and dogs;
    - Output: whether the images are cats or dogs;
    - Task: Predictive or Classification task.
  - What is the input representation? What is the output representation? What is the objective function?
    - Input representation: Images.
    - Output representation: Scalars.
    - Objective function: measurement between the predicted labels and ground truth labels (cross-entropy loss is very common for classification tasks)
- **Data Preparation/Processing** is conducted after problem formulation. The data could be accumulated, or specifically curated for the formulated problem. It is utilized to collect and manipulate the data to produce meaningful information under the formulation of the problem.
- **Data Representation** is another important step to represent data in a machine-readable (or numeric) format. The type of representation is also critical to model choice and which type of specific information it aims to capture.
- **Model Choice** is another important step in the pipeline which determines the key model used to learn to fulfill the task. It mainly includes traditional machine learning models and deep learning models. In addition to the model choice itself, for a model to learn from data efficiently, we often need to design a measurement or objective function such that the model can be aware of how good or bad it is performing, then an optimizer is needed to adjust the model accordingly.
  - **Traditional ML Models** (Modeling data w/ limited structured inductive bias (i.e. data is not always assumed to be in certain structure like graph))
    - **Random Forests, Support Vector Machine, Gradient Boosting,** etc.
  - **Deep Learning Models** (Modeling data w/ structured inductive bias)
    - **Multi-layer Perceptron (MLP)** models all types of data (without structure inductive bias)
    - **Convolutional Neural Network (CNN)** models grid data
    - **Recurrent Neural Network (RNN)** models sequence data
    - **Graph Neural Network (GNN)** models graph data
    - **Transformer** models sequence data originally, but later adapted to model all types of data
  - **Objective/Loss Function**
    - Mean-squared error loss for regression task
    - Cross-entropy loss for classification task
    - More to read: Common Loss functions in machine learning [5]
  - **Optimizer** is the algorithm used to minimize the objective/loss function and update the parameters of the machine learning model. More to read [6]
    - Common optimizers include SGD, Adam, RMSProp, etc.

Evaluation/Result Analysis is conducted to evaluate the performance of the model and provide feedback to improve the whole pipeline.

  - Training/Validation/Testing Set Evaluation (Common procedure: tuning parameters on the training set, select the parameters that have the best performance on the validation set and report the result on the testing set to mimic the real-world scenario when unseen/new data come as the testing set)
  - Evaluation Metrics measure the performance of the model.

Real-world Example (Protein Structure Prediction - Alphafold2)

- Problem Formulation
  - Input: protein sequence (a sequence of N amino acids)
  - Output: protein structure (coordinates of amino acids in 3D space -> (N \times 3))
  - Task: predictive or regression task
- Data Preparation
  - Accumulated protein structures from Protein Data Bank (PDB) (sequence, structure pairs)
  - Searching Multiple Sequence Alignment (MSA) for each protein sequence (demonstrated to help with learning coevolutionary information)
  - Accumulated protein templates (existing templates for some proteins)
- Model Choice - Deep Learning Models
  - Transformers in modeling MSA embeddings and producing pairwise and single-sequence features
  - Transformers in modeling pairwise and single-sequence features and output structures
- Objective Function
  - Cross-entropy loss, mean-squared-error loss, etc.
- Optimizer - Adam Optimizer
- Evaluation/Result Analysis
  - TMScore, lDDT (measurement between two structures)
  - plDDT, pTMScore (predicted lDDT/TMScore for uncertainty estimation)

1.3.7 AI Systematic Learning Roadmap

- AI for Everyone
- Python (Programming)
- Calculus
- Linear Algebra
- Discrete Math
- Probabilistic Statistics
- Machine Learning
- Deep Learning
- Specialized/Advances Topics

  - Machine Learning with Graphs
  - Computer Vision
  - Reinforcement Learning
  - More…

1.3.8 References

[1] Wikipedia Contributors. Deep learning, 05 2019.

[2] DeepMind. Alphago: the story so far, 2016.

[3] The AlphaFold Team. Alphafold: a solution to a 50-year-old grand challenge in biology, 11 2020.

[4] OpenAI. Dall·e 2, Apr 2022

[5] Ravindra Parmar. Common loss functions in machine learning, 09 2018.

[6] Sebastian Ruder. An overview of gradient descent optimization algorithms, 01 2016.

1.4 Molecular Dynamics

This tutorial aims to equip you with the knowledge about molecular dynamics and answer the following questions: (1) what molecular dynamics is, (2) why we run molecular dynamics, (3) how molecular dynamics works and (4) how to run a real molecular dynamics process.

1.4.1 Before you start

- This tutorial assumes that you have already learned:
  - Basic physics and chemistry knowledge
  - Calculus
- If you want to deeply understand the principles and roles of molecular dynamics, you need to master the following:
  - Mechanics, theoretical mechanics (analytical mechanics)
  - Statistical Mechanics
  - Physical Chemistry
  - Numerical Methods / Computational Physics / Computational Methods
  - Further topics:
    - Quantum Chemistry
    - Biochemistry or Solid State Physics, depending on the specific application

1.4.2 What is Molecular Dynamics?

Molecular Dynamics (MD) studies how atomic coordinates evolve under given conditions. It relies on the framework of **classical mechanics**(also known as **Newtonian mechanics**) and simulates the motion of molecular systems numerically. For instance, you can imagine the motion of two rigid balls connected by a spring.

Experiments on a computer

MD is a computational method, a chemical experiment performed on computers. Let's imagine a chemical experiment in the real world first. One may have to prepare experimental instruments and drugs, set instrument parameters, conduct experiments, wait for chemical reactions to proceed, obtain experimental results, and then analyze experimental results. MD is quite similar; we use a table to compare MD with the traditional chemical experiments:

Chemical experiment	Molecular dynamics simulation
Prepare experimental instruments	Prepare computing hardware (CPU, GPU, etc.)Prepare calculation software (Gromacs, Lammps, etc.)
Prepare experimental drugs	Prepare a file describing the molecular structure as an initial conformation for the simulation process
Set instrument parameters	Set simulation parameters (simulation temperature, simulation time, etc.)Set force field parameters (parameters that can be analogous to spring coefficients)
Conduct chemical experiments	Run MD simulations
Get experimental results	Get the simulated trajectory
Analyze experimental results	Analysis of physico-chemical properties from trajectories obtained from simulations (calculation of statistics)

Why do we run MD?

Molecular dynamics simulations can allow us to simulate chemical and physical processes on computers, obtain kinetic information at the microscopic scale, provide theoretical support for experiments and guide chemical experiments. Moreover, computational simulations can help reduce the cost of manually conducting experiments. MD can also be performed under special (usually severe) conditions (ultra-high pressure, ultra-high temperature, strong electric field, magnetic field, etc.)

However, its basis in classical mechanics means that MD can only be effective for physical processes at the molecular scale (e.g. nm). For simulations involving electronic structures MD doesn't work.

- Systems that can be simulated with classical molecular dynamics:
  - Protein systems (protein folding problems, etc.)
  - RNA systems
  - Atomic-scale material systems
  - Free energy calculation
  - Catalytic reactions that do not involve electron transfer
- Problems that classical MD *can not* solve (or cannot be solved only by classical MD):
  - Magnetic and electrical properties of materials (requires quantum chemistry tools)
  - Calculate the energy of molecular conformation (requires quantum chemistry method, current software include Gaussian, Vasp)

- Protease-catalyzed reactions involving electron transfer (requires QM/MM method)
- Chemical reactions involving electron transfer (requires QM/MM method)

The inputs and outputs of MD

Inputs to molecular dynamics simulations are an initial conformation of molecules. They should contain the coordinates of the atoms, the information of chemical bonds among atoms (also called topology information), etc.

The outputs of molecular dynamics are molecular trajectories (trajectory refers to a continuous molecular motion in the coordinate space or Cartesian space) simulated from the initial conformations.

1.4.3 Starting with a simple example

Here we consider a simple example of a molecular dynamics simulation for a hydrogen molecule in vacuum. $H - H$ (a single hydrogen molecule)

Basic knowledge

Let's go over some basic knowledge before simulations:

- Molecules have an associated energy, which can be divided to potential energy (denoted by V) and kinetic energy (denoted by T). Potential energy is related to the coordinates of the atoms in molecules, while kinetic energy is related to the velocity (or momentum) of the atoms.

where q represents the vector coordinates of atoms and p represents the momentum of atoms.

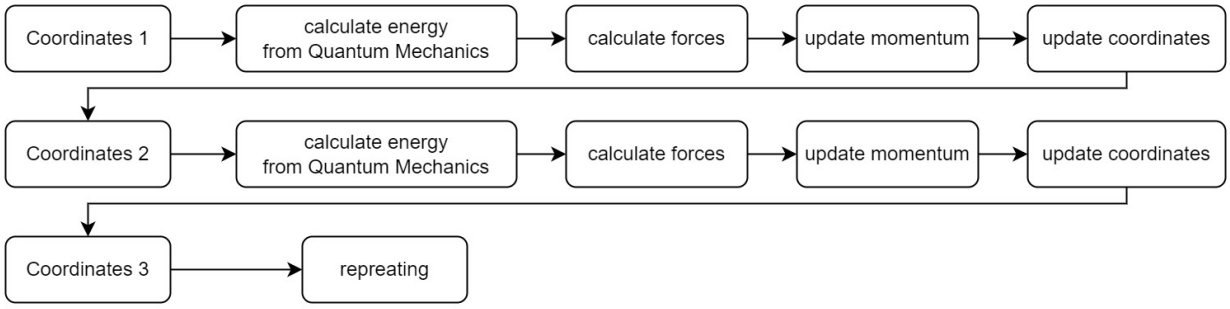
- Force is the negative derivative of the energy over coordinates. Since kinetic energy has nothing to do with positions (coordinates), force is also equal to the negative derivative of potential energy over coordinates.

The construction of theoretical models - generating force fields

Why do we need force fields?

To simulate the motion of hydrogen molecules, we need to know the forces exerted on hydrogen atoms at different positions. These can be used to calculate changes of velocity and coordinates. This is equivalent to knowing the potential energy surface of the H_2 molecule.

How do we get the potential energy surface? The most straightforward way is to calculate the energy by solving Schrödinger equations, and then calculate forces from derivatives. Then our simulation process becomes:



Every time we get a new position in the coordinate space, we need to solve a new set of Schrödinger equations. It is very time-consuming and laborious for complex systems. However, to calculate the motion of hydrogen atoms, we don't need

to know the exact information of the electrons in hydrogen atom, but only the position information of the nucleus. Thus we need **the Born-Oppenheimer approximation**.

The Born-Oppenheimer approximation is a quantum chemical approximation for electrons. Since nuclei are much heavier than electrons, they move much slower. We can assume that the motion of the nuclei is only affected by the mean-field of electrons.

Simply speaking, we only need to know the positions and forces about the hydrogen nucleus. The influence of electrons can be represented by some force field parameters under approximations. For example, we can approximate the hydrogen molecule as a spring model, and regard the interaction from electrons (or chemical bonds) as a spring. The force generated by the chemical bond is described by the spring coefficient. In this way, we transform the problem of solving Schrödinger equations to calculating the force of a spring:

where k represents the spring coefficient, and x_0 represents the offset from the equilibrium position. As a result, calculations are greatly simplified. Here, the **k (spring coefficient, or force constant)** and **x0 (equilibrium positions)** are simplified **force field parameters**, also known as the **force field**.

Note that such simplification can significantly reduce the accuracy of MD simulations, so MD can only reflect the information near equilibrium states. But for macro statistics (e.g. chemical shift, protein contact), it is usually a sufficient description.

The actual force field parameters.

In real MD simulations, almost all inter-atomic and inter-molecular forces are approximated by an analytical expression of classical mechanics. These **empirical parameters are obtained by fitting experimental data or data from quantum calculations}(for example, gradient descent optimization**. Usually, the format of force fields looks like a combination of terms like:

These interactions include:

- Short-range interactions:
  - bonds
  - angles
  - torsion (dihedral angles)
- Long-range interactions
  - Electrostatic interactions (Coulomb interactions, elec)
  - Van der Waals

Preparation for simulation

Now, we have modeled a hydrogen molecule. For simplicity, let the force constant of the hydrogen molecule be $0.1kJ/\mathring{A}^2$. Then let the equilibrium bond length be $0.5\mathring{A}$, and the mass of the hydrogen atom is assumed to be (1) (just for demonstration).

- To start the simulation, we also need the initial conditions (initial conformation and initial velocity). Also for simplicity, we assume that the coordinates of the two atoms are [0.7, 0, 0], [0, 0, 0] (Unit: angstrom), and the velocities are [-0.1, 0, 0], [0.1, 0 , 0] (Unit: angstrom/fs)

```
# ATOM   x,      y,     z,    vx,    vy,    vz
   H1    0.7,   0.0,   0.0, -0.1,   0.0,   0.0
   H2    0.0,   0.0,   0.0,  0.1,   0.0,   0.0
```

- Next, let's set the simulation parameters: since we calculate the differential equation numerically, we need to give the minimum simulation time interval, which is generally set to be 2 fs.(which is actually the grid size for time dimension in finite difference method).
- Set the simulation to 5 steps, so the total simulation time is 10 fs.
- Set the simulated temperature to 300K (this affects the velocity initialization and force field parameters, which should be set before velocity initialization).

Ready? Let's simulate!

Simulation

(Please be aware that units are omitted for simplicity)

1. Calculate the force on the hydrogen atoms from the initial positions:
2. Calculate the accelerations:
3. Update velocities:
4. Update the coordinates and velocities through the calculation above to obtain new coordinates:
5. (Repeat) Calculate the new force
6. (Repeat) Calculate the new acceleration

```
# ATOM   x,      y,     z,    vx,    vy,    vz
   H1    0.5,   0.0,   0.0, -0.15,   0.0,   0.0
   H2    0.2,   0.0,   0.0,  0.15,   0.0,   0.0
```

Results

Finally, we will obtain positions of H atoms at different times, and these "continuous" motions compose the MD trajectory.

Through long-term MD simulations, we can calculate **physical and chemical** properties from the trajectories. (such as RMSD changes, conformation transitions, ensemble averaging of physical quantities, etc.)

Notes: Here, the equations from Newtonian mechanics are used for the purposes of demonstration. MD simulation is **usually more complicated**. Depending on the required **ensemble**, numerical evolution of partial differential equations will be carried out under specific Hamiltonian mechanics.

We will not introduce the concept of the ensemble in detail here. If you are interested in it, please read the related content about **statistical mechanics**.

1.4.4 An advanced example

In this part, we will use an example of a **protein** to illustrate how to conduct MD simulations in real-world research.

Building Model

Again, we use the BO approximation to construct models for proteins. These models include information on bond types, atom types, and force field parameters for various interactions. We usually need the following files:

- topology (`topol.top` for example): Contains molecular bonding information, molecular type information, and atomic type information
- force field (`forcefield.itp` for example): contains chemical bond equilibrium positions, chemical bond force constants, non-bond force constants, etc.
- Commonly used force fields are as follows:
  - Amber
  - Charmm
  - Gromos
  - OPLS
- So far there is no unified database or maintenance of methods for force fields, and each force field is maintained by each company or organization. Developing a force field is a difficult and nuanced process.

Solvent

Unlike systems in vacuum, proteins generally exist in solvents. We therefore need to account for and model the solvent. We can do this in two ways:

- **Explicit solvent model.** It is often used in all-atom simulations, that is, directly introducing solvent molecules into the system. The force field parameters for solvent molecules are usually included in the force field file. Commonly used models for water are SCP, TIP3P, TIP4P, etc.
  - Advantages: Similar to the actual physical process, the results are more accurate. It can describe solvent-involved processes (e.g. protein-ligand binding). It can also explicitly describe solvent effects such as hydrogen bonding.
  - Disadvantages: high computational complexity. A system usually contains hundreds or thousands of solvent molecules.
- **Implicit solvent model.** The effect of the solvent on the solute is described by a continuous electric field model. The Generalized Born model is an example of commonly used model.
  - Advantages: low computational complexity, no need to introduce additional solvent molecules.
  - Disadvantages: Imprecise, solvent-involved reactions cannot be described.

1. **Periodic Boundary Conditions**:

   Due to limited computation resources it is impossible for us to simulate an infinitely large system, nor to simulate infinite steps. For a protein system, tens of thousands of atoms already require a lot of calculations (the required calculation time is measured in days), but this is still far less than Avogadro's constant(10^{23}).

   However, simulating a small system will be seriously affected by the interface and cannot reflect the properties of the bulk phase.

To solve this problem, we often use **periodic boundary conditions**. We confine the system of interest in a box and assume the properties of the actual system can be approximated by an virtual infinite system of repeating side-by-side lattices. If the molecule passes through the box boundary, it will re-enter the box from the opposite boundary, forming a periodic space.

2. **Preparation for simulation**

   Next, you need to prepare the conformation files of the protein and water molecules. Typically, protein structure files are generated by PDBs, then converted into a format that can be read by molecular dynamics software.

   We also need to prepare simulation parameter files, in which you need to set:

   - temperature
   - time interval and duration for simulations
   - the integrator/numerical algorithm, the ensemble for simulations
   - the temperature/pressure controller
   - the output-frequency and content of output files

3. **Simulation** Commonly used MD simulation software in current research are as follows:

   - Amber [1]
     - Commercial software. There is a free version and a high-performance optimized commercial version. The code is closed source and highly engineered. It supports most MD functions and plays an important role in the simulation of protein systems. Amber is often used in simulations of biological systems.
   - Gromacs [2]
     - Open-source MD simulation software. The latest release is the Gromacs 2022 release version. With efficient GPU optimization, there is a good developer community for Gromacs. This is mostly used for biological system simulation.
   - OpenMM [3]
     - Molecular dynamics simulation software with Python interfaces. Modules are implemented by calling functions from the Python command line, which can directly involve deep learning frameworks such as PyTorch. However, its compatibility has not been fully developed. Mostly used for biological system simulation.
   - Lammps [4]
     - MD software for material simulations.

   Most MD software is fully optimized on GPU devices, which can provide greater efficiency than CPU devices. Another program, Charmm, can be used for preparing structures for MD but it can not be used for simulations.

   In addition, most software is compatible with other formats of force fields; for instance, Gromacs is compatible with Amber or Charmm force fields.

1.4.5 Rare Events and Enhanced Sampling

- Molecular dynamics algorithms have time reversibility and ergodic hypotheses. We assume that all states of a molecule have a probability to be explored (or traversed/sampled) after a sufficiently long simulation. These states include ground states, meta-stable states, and some high-energy states (unstable states). When the simulation reaches equilibrium, the distribution of molecular conformations in the system satisfies **the Boltzmann distribution** under its **ensemble**.
- Different states have different probabilities to be sampled. In most cases, the molecules are stuck in a local minimum on the energy surface, and it is difficult to jump over the energy barriers. Therefore, under finite-time simulations, the probability of sampling some high-energy state or another meta-state separated by an energy barrier is very low. These are **rare events**.
  - Here, it can be compared with Monte Carlo (MC) simulations of a high-dimensional function, starting from a random value of the high-dimensional function to explore the global minima of the function. This can take a lot of time and computational resources.
- To increase the probability of rare events occurring in MD simulations, the simulation process can be interfered with using various methods:
  - Enhanced sampling based on temperature:
    - Raise the temperature, lower the energy barriers, and increase the probability of rare events.
    - Replica-Exchange Molecular Dynamics (REMD)[5], selective integrated tempering sampling (SITS)[6], etc.
  - Enhanced sampling based on bias potential:
    - Collective variables (CV), are functions of system coordinates. The free energy are defined on collective variables. (Please refer to the difference between free energy and potential energy)
    - Add bias potential to the given CVs during the simulations, which can push the trajectory out of the local minima on the energy surface to explore other states.
    - Metadynamics[7], VES[8], RiD[9], etc.
    - Traditional boosted sampling methods based on bias potential suffer from **the curse of dimensionality.**

1.4.6 AI in MD

The current difficulties restricting MD simulation are as follows:

- Simulation accuracy
  - Introduced by the classical force field approximations to quantum mechanics.
- Sampling efficiency
  - This is the most essential problem of MD. It is partly solved by enhanced sampling techniques but traditional enhanced sampling methods cannot solve high-dimensional problems. Excessive heating or aggressively increasing bias potential can lead to denaturation of systems.

Here, some people draw inspiration and seek solutions from deep learning to try solve these problems.

- DeePMD[10]
  - By fitting quantum chemical data with neural networks, a potential energy surface with quantitative accuracy is obtained. The computational complexity of neural networks is far less than solving Schrödinger equations, so DeepMD can achieve molecular dynamics simulations with high-precision.
- Neural-networks-based enhanced sampling

  - These methods obtain the representation of the high-dimensional free energy surface by fitting the mean-force data using neural networks. The bias potential is applied to the system to boost simulations, where neural networks alleviate the high-dimensional problem of multiple CVs. These works include NN-VES, Reinforced Dynamics, etc.

- CV Discovering
  - Reduce the dimensionalities of the data through machine learning methods (SVM, VAE, GAN, etc.) to find the most essential collective variables. TorchCV[11] is a good example for this.

Difficulties:

- How to generate MD data for training ML models?
  - Possible solutions: active learning, concurrent learning
- It is temporarily hard to do end-to-end simulation with similar efficiency to well-developed MD software.

1.4.7 References

[1] Romelia salomon ferrer, David Case, and Ross Walker. An overview of the amber biomolecular simulation package. Wiley Interdisciplinary Reviews: Computational Molecular Science, 3, 03 2013.

[2] Henk Bekker, Herman Berendsen, E.J. Dijkstra, S. Achterop, Rudi Drunen, David van der Spoel, A. Sijbers, H. Keegstra, B. Reitsma, and M.K.R. Renardus. Gromacs: A parallel computer for molecular dynamics simulations. Physics Computing, 92:252–256, 01 1993.

[3] Peter Eastman, Jason Swails, John Chodera, Robert Mcgibbon, Yutong Zhao, Kyle Beauchamp, Lee-Ping Wang, Andrew Simmonett, Matthew Harrigan, Chaya Stern, Rafal Wiewiora, Bernard Brooks, and Vijay Pande. Openmm 7: Rapid development of high performance algorithms for molecular dynamics. PLoS Comput Biol, 06 2017.

[4] Aidan Thompson, H. Metin Aktulga, Richard Berger, Dan Bolintineanu, W. Brown, Paul Crozier, Pieter in 't Veld, Axel Kohlmeyer, Stan Moore, Trung Nguyen, Ray Shan, Mark Stevens, J. Tranchida, Christian Trott, and Steven Plimpton. Lammps - a flexible simulation tool for particle-based materials modeling at the atomic, meso, and continuum scales. Computer Physics communications, 271:108171, 09 2021.

[5] Zhou Ruhong. Replica exchange molecular dynamics method for protein folding simulation. Methods in molecular biology (Clifton, N.J.), 350:205–23, 02 2007.

[6] Lijiang Yang and Yi Gao. A selective integrated tempering method. The Journal of chemical physics, 131:214109, 12 2009.

[7] Alessandro Barducci and Massimiliano Bonomi. Metadynamics. WIREs Comput. Mol. Sci., 1:826, 09 2011.

[8] Luigi Bonati, Yue-Yu Zhang, and Michele Parrinello. Neural networks-based variationally enhanced sampling. Proceedings of the National Academy of Sciences, 116:201907975, 08 2019.

[9] Dongdong Wang, Yanze Wang, Junhan Chang, Linfeng Zhang, Han Wang, et al. Efficient sampling of high-dimensional free energy landscapes using adaptive reinforced dynamics. Nature Computational Science, 2(1):20–29, 2022.

[10] Linfeng Zhang, Jiequn Han, Han Wang, Roberto Car, and Weinan E. Deep potential molecular dynamics: a scalable model with the accuracy of quantum mechanics. July 2017.

[11] Luigi Bonati, Valerio Rizzi, and Michele Parrinello. Data-driven collective variables for enhanced sampling. The Journal of Physical Chemistry Letters, 11:2998–3004, 04 2020.

1.5 Knowledge Base

1.5.1 Physics

Molecular Dynamics

Molecular Dynamic (MD) is a type of molecular simulation method, which aims to study the dynamic evolution of physical systems through computer simulations of atoms and molecules. Based on MD simulations and statistical mechanics, many macroscopic thermodynamic properties, for instance, free energy or density, can be evaluated. Typically, trajectories of atoms in a simulation are generated by solving Newton's laws of motion, where the potential energy function V comes either from force fields or Quantum Mechanic (QM) ab-initio calculations:

Depending on the smallest indivisible unit during a simulation, MD simulations can be roughly divided into two major categories: All-atom Molecular Dynamics and Coarse-grained Molecular Dynamics (CGMD):

- **All-atom Molecular Dynamics**: each individual atom is treated as the smallest indivisible unit for motion and force calculations
- **Coarse-grained Molecular Dynamics**: a set of adjacent atoms (such as an amino acid residue, a water molecule) is treated as a coarse grained unit, usually referred as a "bead". Only interactions between beads are considered, while all intra-bead interactions are neglected during a CGMD. This treatment makes CGMD capable of performing simulations on a larger time scale and for larger physical systems with reduced cost of computation and increased loss of accuracy.

Depending on the accuracy of potential energy functions used during a simulation, MD simulations can be divided into three categories: Classical Molecular Dynamics (Classical MD, cMD), Ab-initio Molecular Dynamics (AIMD) and Machine Learning Molecular Dynamics (MLMD):

- **Classical Molecular Dynamics**: potential energy functions of the physical system come from a force field;
- **Ab-initio Molecular Dynamics**: potential energy functions of the physical system come from ab-initio calculations;
- **Machine Learning Molecular Dynamics**: potential energy functions of the physical system come from a machine learning force field.

Potential Energy Function

Potential Energy Function, usually shortened as "Potential", refers to the function that is used to describe the energy of interaction within a physical system. In an all-atom MD simulation the potential is a function of the atom types and atomic coordinates within the given physical system, and it could be given by quantum mechanics (QM), molecular mechanics (MM) force fields, or machine learning (ML) force fields.

Force Field

Force Field, conventionally called Molecular Mechanics (MM) Force Field, refers to a collection of empirical functions with fixed mathematical formats to describe the potential energy of the physical system. Parameters for these empirical functions are determined by fitting against experimental data or QM-derived data. Compared to ab-initio methods, MM force fields are less accurate but much faster (usually several magnitudes).

Hamiltonian

Under the context of classical mechanics, the concept of the Hamiltonian refers to the total energy of a physical system, which is the sum of the potential energy and the kinetic energy of all particles within the given system.

In quantum mechanics, the Hamiltonian should be considered as an Hamiltonian operator.

Statistical Mechanics

In physics, statistical mechanics is a sub-discipline which applies statistical methods and probability theory to describe large assemblies of microscopic particles so that macroscopic behavior of the physical system (for instance, temperature, pressure) can be related to the behavior of microscopic particles.

State Function

State Function is a physical property to describe the macroscopic property of a physical system. State functions have fixed values for a physical system under certain thermodynamic equilibria and depend only on the current equilibrium state of the system, rather than the path on which the system reaches equilibrium. Examples of State Functions include internal energy, enthalpy, entropy, free energy, etc.

Ensemble

Ensemble is a concept in statistical mechanics, which refers to a collection of a large number of independent systems with identical properties and structures in various motion states under certain macroscopic conditions.

Free Energy

The thermodynamic free energy refers to the energy of a thermodynamic system that can be used to do external work. It can be used as a criterion for whether a thermodynamic process can proceed spontaneously. Under given constraints, the system always tends to transition to a state with low free energy. For example, the process of protein folding is the spontaneous transition from an unfolded state with higher free energy to a folded state with lower free energy. According to the different qualifications, it can be divided into Helmholtz free energy (common notation F) and Gibbs free energy (common notation G). Note: free energy is different from potential energy although many people may confuse them.

Boltzmann Distribution

In statistical mechanics, the Boltzmann distribution describes the In statistical mechanics, the Boltzmann distribution describes the probability distribution of particles in a system in possible microscopic quantum states, and has the following form:

where E is the quantum state energy, k is the Boltzmann constant T is the temperature, p_i is the probability that the particle is in the i quantum state, and ε_i is the energy of the i quantum state.

Collective Variables (Reaction Coordinates)

The representative parameters that can quantitatively describe the change process of the system are called Collective Variables (CV) or Reaction Coordinates (RC). For example, in the chemical reaction shown in the figure below, the distance between O and C $d(\mathrm{C}-\mathrm{O})$ can be regarded as the reaction coordinate, and the distance between C and Br $d(\mathrm{Br}-\mathrm{C})$ can also be regarded as the reaction coordinate.

Given that the reaction coordinates are well defined, methods such as umbrella sampling can be used to estimate the free energy difference between different reaction coordinates through molecular simulation, and then the free energy change along with the reaction coordinates during the transforming process can be described, which is the basis of kinetic and thermodynamic research.

Slow Degrees of Freedom

In the process of dynamic simulation, some degrees of freedom change rapidly with time (such as bond length, bond angle, etc., usually on the order of fs or ps). And some degrees of freedom change slowly with time (such as the dihedral angle, usually on the order of ns, μ s , or even ms).

Enhanced Sampling

Enhanced sampling refers to accelerating the sampling of slow degrees of freedom in the simulation process by some technical means, which are classified as collective variable-based (e.g. umbrella sampling), and collective variable-free (e.g. replica exchange).

Quantum Mechanics

Quantum Mechanics is a branch of physics that studies microscopic systems. By describing the motion and interaction of microscopic particles (such as electrons, protons, etc.), quantum mechanics can explain many experimental phenomena that cannot be explained under the framework of classical mechanics, including blackbody radiation and the spectrum of the hydrogen atom.

Operator

Generally, an operator acts on the state space of a physical system, making the physical system transform from one state to another. Within the context of quantum mechanics, the state of a system can be described by a state vector. Physical observables (such as position, momentum, Hamiltonian, etc.) all correspond to a (Hermitian) operator.

Schrödinger Equation

In quantum mechanics, the Schrödinger equation is a partial differential equation that describes the time evolution of the quantum state of a physical system and is the fundamental equation of quantum mechanics. The Schrödinger equation can be divided into two types: the “time-dependent Schrödinger equation”

and the “time-independent Schrödinger equation” (also known as the steady-state Schrödinger equation)

The time-dependent Schrödinger equation describes how the wave function of a quantum system evolves over time, while the time-independent Schrödinger equation describes the physical properties of a stationary quantum system.

First Principle

First Principle, also called ab initio, refers to derivation and calculation based on the basic laws of physics without additional assumptions and empirical fitting. For example, the of use the Schrodinger equation to solve electronic structure.

Wave Function

In quantum mechanics, the state of a quantum system can be described by a wave function. The wave function Ψ(**r**,t) is a complex-valued function. According to Bonn's statistical interpretation, $||^2$ is the probability density of finding a particle at position **r**, time t.

Born-Oppenheimer Approximation

The Born-Oppenheimer approximation refers to the approximate variable separation of the nuclear coordinates and the electron coordinates when solving quantum mechanical equations containing the nucleus and electrons, to decompose the wave function of the whole system into separately solving the nuclear wave function and the electron wave function, which are two relatively simple processes. The basis of this approximation is that the mass of the nucleus is 3 to 4 orders of magnitude larger than that of the electron, and the speed of the nucleus is much smaller than that of the electron, so the electron can be regarded as being in the potential field formed by the stationary nucleus, and the nucleus won't be affected by the specific position of the electron, only the average force of electrons counts.

Density Functional Theory

Density functional theory (DFT) is a quantum mechanical method to study the electronic structure of multi-electron systems, and it is one of the most commonly used methods in the fields of condensed matter physics and computational chemistry. Since the classical method of electronic structure theory needs to solve the multi-electron wave function with a higher dimension ($3N$ for a system containing N electrons), the basic idea of the density function is to use the electron density instead of the wave function as the basic amount of research, thereby reducing the computational complexity. The most common application of density functional theory is implemented with the Kohn-Sham method.

1.5.2 Chemistry

Atomic Orbitals

In Quantum Mechanics, Atomic Orbitals are mathematical functions that describe the wave-like behavior of electrons in atoms. This function can be used to calculate the probability of electrons appearing around the nucleus, and the meaning of "orbital" refers to the probability of electrons appearing in a specific area. According to the "shape" of the track, it can be classified into s, p, d, f, etc.

Electronegativity

Electronegativity describes the ability of atoms of an element to attract electrons in a compound. The greater the electronegativity of an element, the stronger the ability of its atoms to attract electrons in the compound. In a period of the periodic table, the electronegativity of the element atom increases from left to right; and it decreases from top to bottom in a group. Therefore, the elements at the upper right of the periodic table (O, N, F, Cl, etc.) have higher electronegativity values. The element with the greatest electronegativity is fluorine.

Chemical Bond

A chemical bond refers to the strong interaction between atoms, ions, and other particles. Through chemical bonds, particles can form polyatomic compounds (such as organic molecules, inorganic molecules, ionic compounds, etc.). Simply put, for a polyatomic system, the most stable configuration between positively charged nuclei and negatively charged electrons is that when electrons are located between nuclei, electrons are attracted between different nuclei, and using this force the nuclei are "attracted" together, forming a chemical bond.

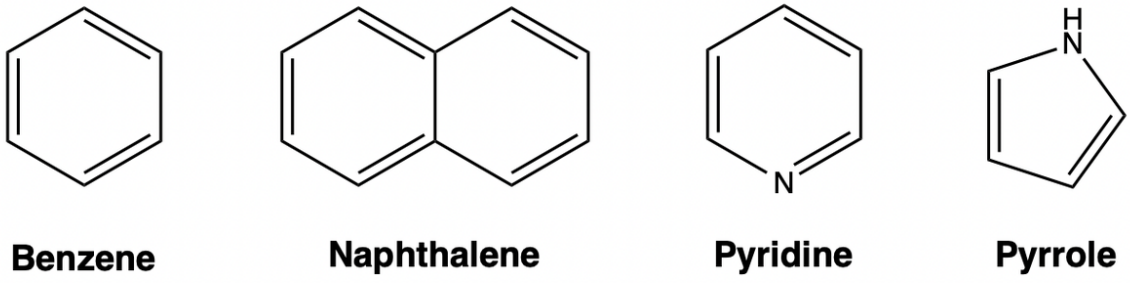
- **Ionic Bond**: A chemical bond formed by electrostatic interaction between oppositely charged anions and cations, without directionality, such as sodium chloride (salt), calcium carbonate.
- **Covalent Bond**: A chemical bond formed by sharing electron pairs between atoms. Two atoms with similar electronegativity are equally attracted to electrons, so they mainly form chemical bonds by sharing each other's outer valence electrons. Covalent bonds are directional, resulting in complex molecular structures. For example, in the methane molecule, carbon atoms and hydrogen atoms are connected by covalent bonds to form a regular tetrahedron, the carbon atom is located at the center of the tetrahedron, and the hydrogen atom is located at the vertex of the tetrahedron. According to the number of shared electron pairs, it can also be classified into a single bond, double bond, and triple bond.
- **Hydrogen Bond**: When a hydrogen atom forms a covalent bond with an atom with high electronegativity X (usually O, N, F), if it bonds with another atom with high electronegativity. When Y (usually also O, N, F) is close, using hydrogen as the medium between X and Y, a special form of interaction like X-H$\cdots$Y is generated, known as a hydrogen bond. Hydrogen bonds widely exist in biological macromolecules such as water and proteins and DNA. It plays a crucial role in stabilizing the conformation of biological macromolecules.

Functional Group

Functional groups are atoms or groups of atoms that determine the properties of organic compounds. Common functional groups include hydroxyl (-OH), carboxyl (-COOH), ether bond (C-O-C), carbonyl (C=O), halogen atom (-F, -Cl, -Br, -I), etc.

Aromatic

Aromaticity is a chemical property that exists in cyclic planar molecules co ntaining π bonds composed of delocalized electrons, which can provide molecules with stability that cannot be explained by conjugation alone. The number of electrons in the delocalized π of an aromatic molecule needs to satisfy the Huckel rule (also called the "4n+2" rule). Molecules with aromaticity are called aromatic compounds, and molecules without aromaticity are called aliphatic compounds. Aromatic compounds can be roughly classified into simple aromatic compounds (such as benzene), polycyclic aromatic compounds (such as naphthalene, and anthracene), and heterocyclic compounds (such as pyridine, and pyrrole).

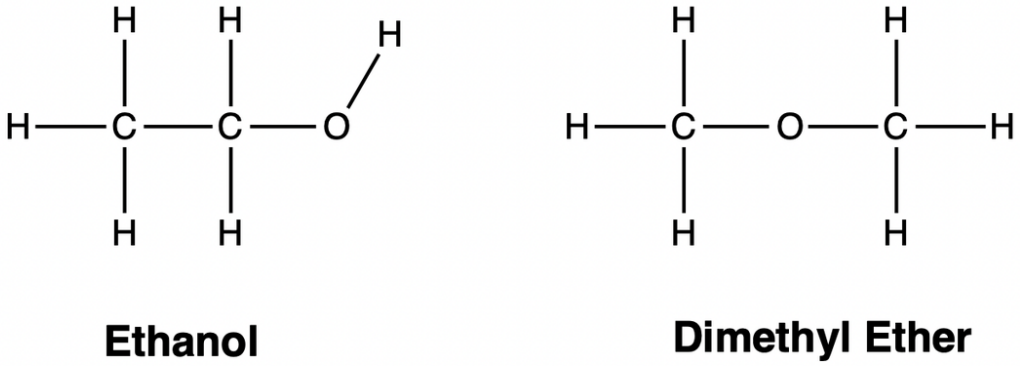


Conformation

Conformer usually refers to three-dimensional conformation, which refers to the structure that a molecule has in three-dimensional space. For organic molecules, their conformations cannot be randomly generated due to the limitation of the directionality of covalent bonds.

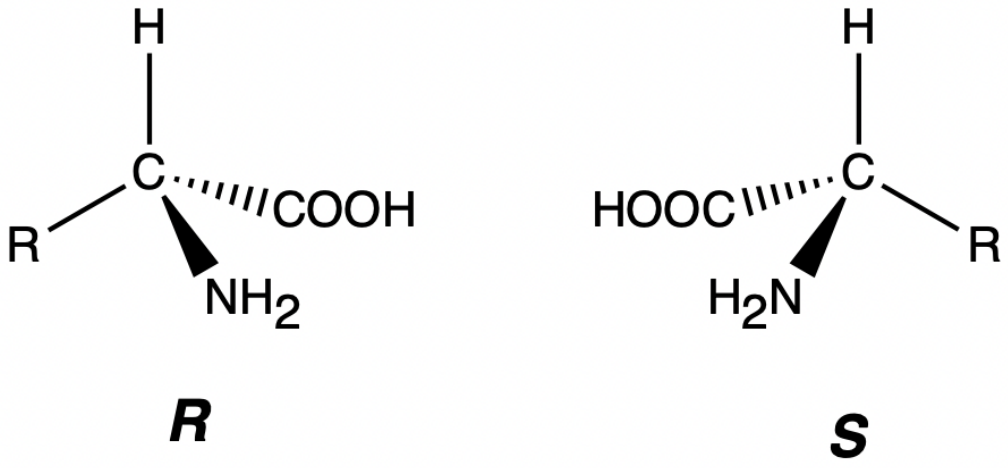
Isomers

In Organic Chemistry, substances with the same chemical composition (molecular formula) but different structures are called isomers of each other. For example, the compositions of ethanol and dimethyl ether are both C_2H_6O, but their structures are different:



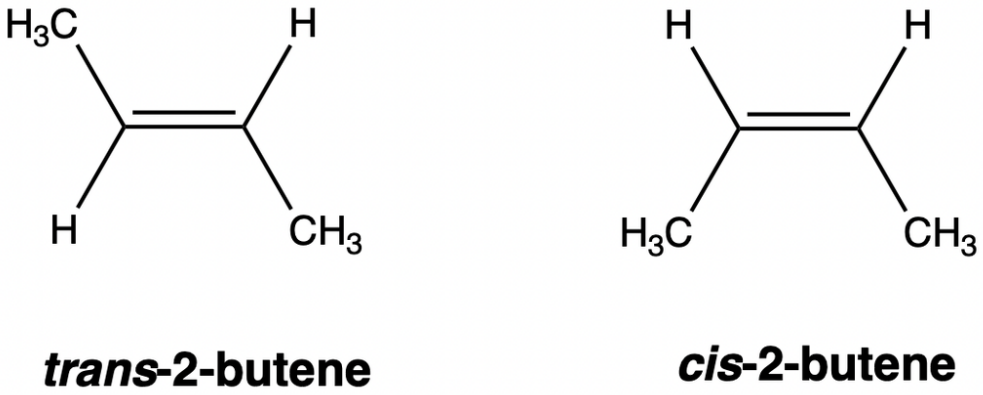
Stereoisomerism

Stereoisomers refer to molecules in which atoms are topologically connected in the same way but spatial arrangement of the atoms are different. For example, a molecule is likeley to have stereoisomers when it contains carbon atoms to which four different functional groups are bonded. Such atom is called chiral atoms, and usually R/S are denoted to distinguish two different them. In terms of biomolecules, such as peptides, amino acids and sugar, L/D are frequently used to denote different type of stereoisomers. The two amino acid configurations shown in the figure below are stereoisomers of each other. All natural amino acids are in the L configuration, and their carbon atoms are in the S configuration.



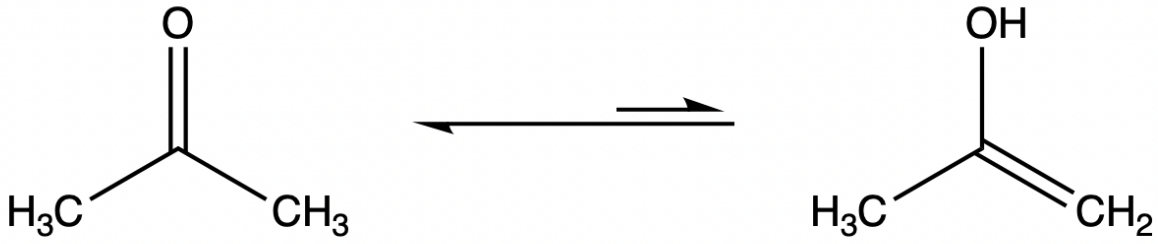
Cis-trans Isomerism

Cis-trans isomerism refers to isomerism that occurs due to the hindered free rotation in the compound molecule, which is commonly found in compounds with double bonds or rings.



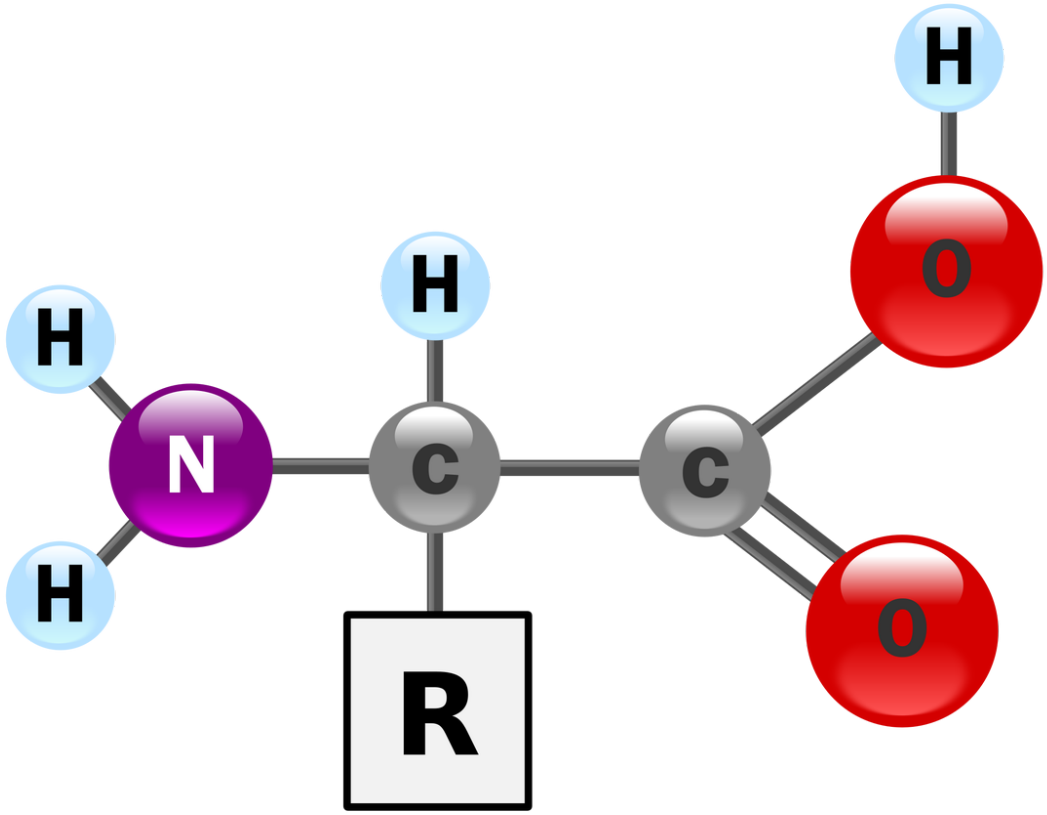
Tautomerism

Tautomerism means the structure of some organic compounds is converted between two functional isomers. Most tautomerisms involve the transfer of hydrogen atoms or protons, and the conversion of single bonds to double bonds. The distribution of tautomers in equilibrium depends on specific factors, including temperature, solvent, and pH, etc. The diagram below shows the keto (left) and enol (right) tautomers present in carbonyl compounds, with the keto structure predominating in the usual case.



Amino Acids

Amino acids are biologically important organic compounds consisting of amino (-NH2) and carboxyl (-COOH) functional groups and side chains attached to each amino acid. Amino acids are the basic units that make up a protein. In nature, there are 20 genetically encoded amino acids.



Protein Structure

Protein structure refers to the spatial structure of a protein biomolecule, which can be divided into four levels to describe different aspects.

- **Primary structure**: the linear amino acid sequence that makes up the polypeptide chain of a protein.
- **Secondary structure**: a stable structure formed by hydrogen bonds between C=O and N-H groups between different amino acids, mainly α-helix and β-sheet.
- **Tertiary structure**: the three-dimensional structure of a protein molecule is formed by the arrangement of multiple secondary structural elements in three-dimensional space.
- **Quaternary structure**: used to describe the interaction of different polypeptide chains (subunits) to form functional protein molecules.

Ligand

In biochemistry or pharmacology, a ligand refers to a compound that can bind to a receptor and then lead to some physiological effect. In medicinal chemistry, ligands are usually small organic molecules or short peptides composed of several amino acids. The forces between ligands and receptors are usually non-covalent interactions: such as hydrogen bonds, electrostatic interactions, van der Waals interactions, etc.

Receptor

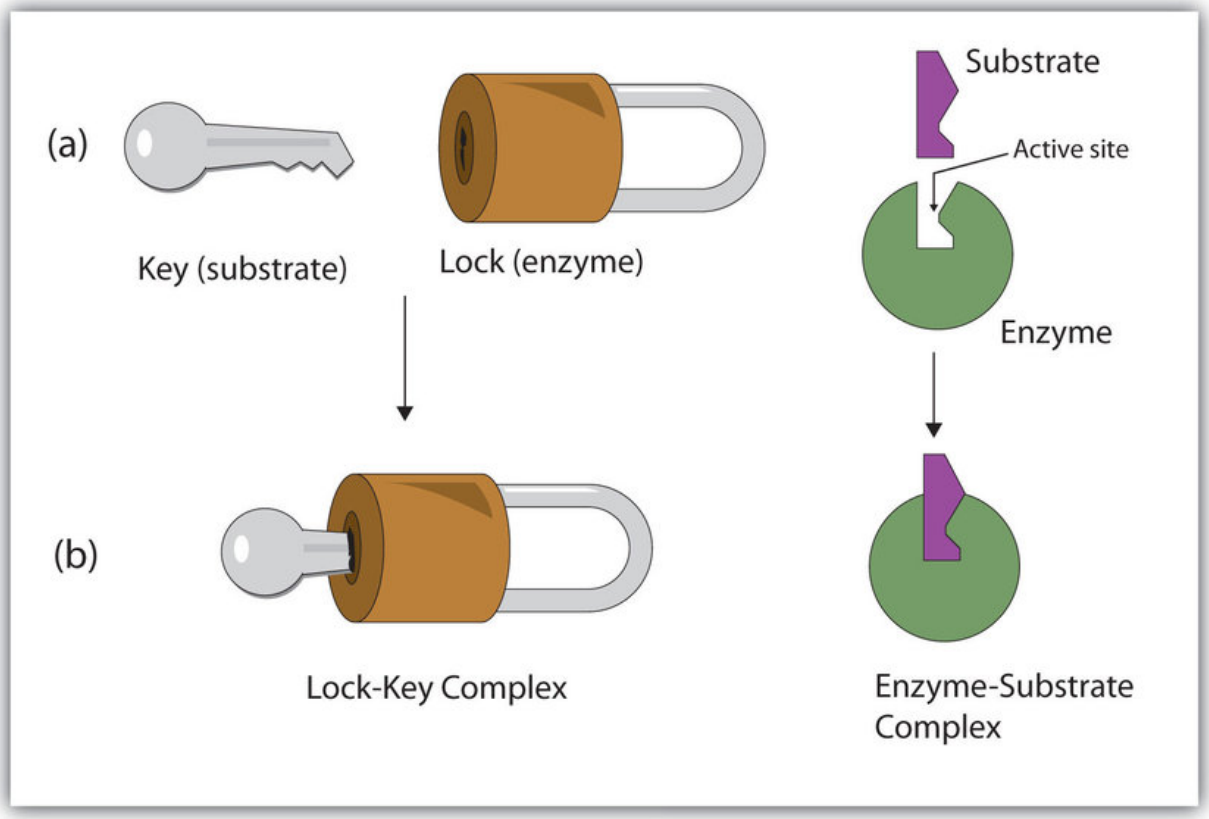
Signal transduction is responsible for intracellular communication via series of molecular events (protein phosphorylation) upon chemical/physical signal outside cell , where receptor function in the central role as transmit signals outside cells and produce specific effects within cells. It is usually biological macromolecule such as protein. After the receptor binds to a specific stimuli, the structure will change to a certain extent, and the corresponding effect will be induced in the cell. In medicinal chemistry, receptors usually refer to target proteins able to bind with ligands.

Lock and Key Model

The lock-and-key model is a theory proposed by E. Fischer in 1890 to explain the specific binding between enzymes and substrates (or between ligands and receptors). The model believes that the structures of enzymes and substrates at their binding sites should be strictly matched and highly complementary, just like the structural complementarity and matching of a lock and its original key. The disadvantage of this model is that the model treats the structure of the enzyme and the substrate as rigid structures, which is inconsistent with the fact that the conformation of the enzyme and the substrate changes during the catalytic reaction.

Induced Fit Model

The Induced-Fit Model is a model proposed by Koshland in 1958 to describe the enzyme-substrate (ligand-receptor) binding interaction. This model believes that in the process of binding the enzyme to the substrate, the substrate can induce a certain change in the structure of the enzyme, and finally form an active conformation that can bind to the substrate.



Molecular Docking

Molecular Docking is a technique that simulates the interaction between ligands and receptors. The technology predicts ligand binding modes and ligand-receptor binding forces by physically modeling intermolecular interactions and applying optimization algorithms such as the Monte Carlo method.

Reversible Reaction

A reversible reaction is a chemical reaction that can proceed in both the forward and reverse directions under the same conditions. When the degree of the reverse reaction direction is much smaller than that of the forward reaction direction, the reaction can be considered irreversible. Most of the reactions are reversible, such as the dissociation of weak acid/base, ligand-receptor binding, etc.

Chemical Equilibrium

Chemical Equilibrium refers to a state in which the forward and reverse reaction rates of a chemical reaction are equal in a reversible reaction with certain macroscopic conditions, and the concentrations of the reactants and the components of the products do not change. Take the following reaction as an example:

When the equilibrium is reached, the concentrations of A, B, C are respectively $[A]$,$[B]$,$[C]$, then the equilibrium constant K can be defined:

Given the reaction conditions, the equilibrium constant for a reaction with a fixed stoichiometric ratio is the same, and is related to the free energy change of the reaction as follows:

van der Waals force

van der Waals (vdW) force refers to the non-directional, unsaturated, weak interaction force between atoms. Van der Waals interactions are much weaker than chemical bonds, but they will significantly affect the melting point, boiling point, and many other properties. Van der Waals interactions have 3 major contributions:

- **Attractive or replusive interactions** are between permanent charges, dipoles, quadrupoles, etc.
- **Induction** (also known as polarization), which is the attractive interaction between a permanent multipole on one molecule with an induced multipole on another. This interaction is sometimes called Debye force.
- **Dispersion** (usually named London dispersion interactions after Fritz London), which is the attractive interaction between any pair of molecules, including non-polar atoms, arising from the interactions of instantaneous multipoles.

In molecular simulations, van der Waals forces are usually described in terms of the Lanner-Jones potential function, which has the following form:

Where r is the distance between two atoms, C is a parameter, usually obtained by fitting physical quantities such as density and the enthalpy of evaporation.

Hydrophobic interaction

Hydrophobic interaction, also known as a hydrophobic effect, is a chemical phenomenon that which groups with hydrophobicity in an aqueous solution (such as alkyl groups without polarity) are close to each other to reduce the contact area with water. Hydrophobic interactions are the main driver of protein folding.

Thermodynamics

Thermodynamics focuses on the interaction of heat and work between chemical reactions and system states under the laws of thermodynamics. Generally speaking, the problems (equilibrium state) that do not involve the study of the chemical reaction process belong to the category of chemical thermodynamics, such as phase transition, and the balance of sodium and potassium ions on two sides of the cell membrane.

Kinetics

Kinetics, also known as reaction kinetics and chemical reaction kinetics, is a branch of physical chemistry that studies the rate and mechanism of chemical reactions. Chemical kinetics is different from chemical thermodynamics. It does not care about the equilibrium state, but studies the chemical reaction dynamically, and studies the time required for the transformation of the reaction system, as well as the microscopic process involved.

1.5.3 Biology

Cell Biology

A branch of biology that studies the structure, corresponding function and subsequent behaviour of components within a cell.

Biochemistry

Solving biological issues with chemical perspectives and techniques.Focusing on the intracellular entities, treating them as chemical blocks and studying the their functionality thus map out the landscape about how life works.

Molecular Biology

Molecular biology studies the composition, structure, function and behaviour of bio-active and/or bio-significant molecules, such as nucleic acids and proteins.

Genetics

Study heredity in the perspective of elemental blocks from DNA and their temporal/spatial distribution/variation in organism. Originality of diseases (abnormality) and driving force of evolution could be derived from thorough understanding of genetics.

X-omics

Source of large-scale and comprehensive biological data assembled from Genomics Transcriptomics, Proteomics, Metabonomics, Microbiomics.

Systems biology

Analysis and modeling of complex biological systems based on data acquired by X-omics.

Synthetic biology

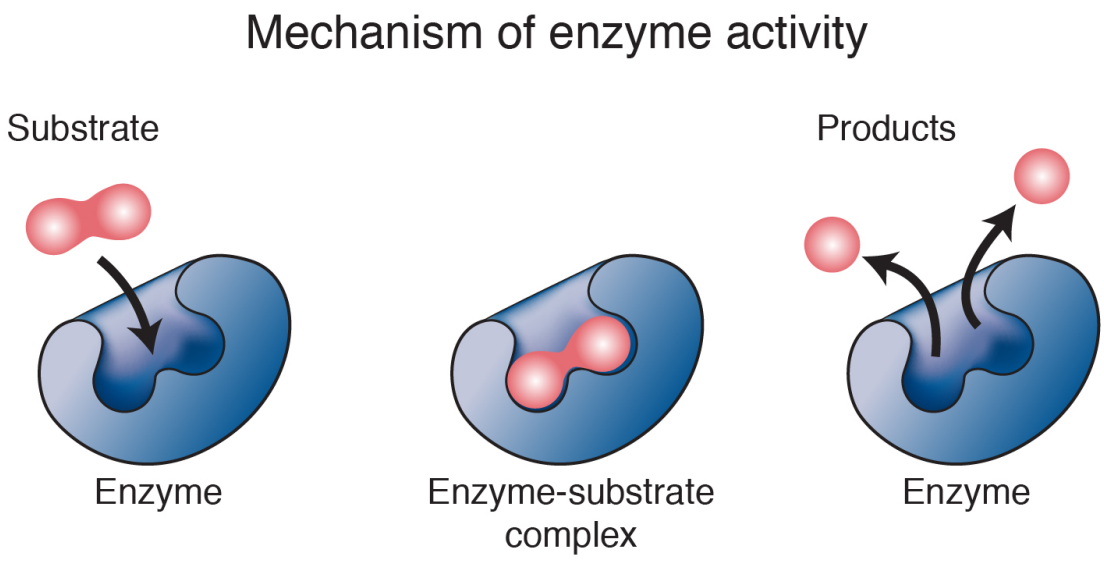
Design of new device and circuits based on biological components.

Epidemiology

Distribution and determinants of disease in population.

Enzyme

An enzyme is a biological catalyst that is capable of accelerating a specific chemical reaction in cells. The enzyme is not destroyed during the reaction process and could be used again and again (under the sustained condition). In most cases, enzymes are proteins.



Phase I/II biotransformation

metabolism of a drug can be divided into 2 phases. Phase I mainly involves the breakdown (mainly by hydrolysis and oxidation). Phase II mainly involves the conjugation of chemical groups (polar in most cases) to make drug more soluble and suitable for excretion.

Cytochrome P450

A family of key enzymes contain heme as the cofactor to function as mono-oxygenases. It is the typical phase I drug metabolizing enzyme and are involved in so many components' metabolism from drug and food. They can be easily induced and inhibited by their substrate thus have a outstanding role when studying the drug-drug interaction (DDI). e.g. Patients who are taking Alvastatin are not allowed to eat grapefruit.

Drug targets

Molecules that are intrinsically associated with particular diseases and could be specifically addressed by a drug to take action. Most of the known drug targets are proteins.

Active site

Catalytic center of enzymes that bind substrate(s) and initiate reactions. For enzymes that are proteins, side chains along the backbone of key amino acids constructing the active site, shape it into specific size with specific chemical behavior.

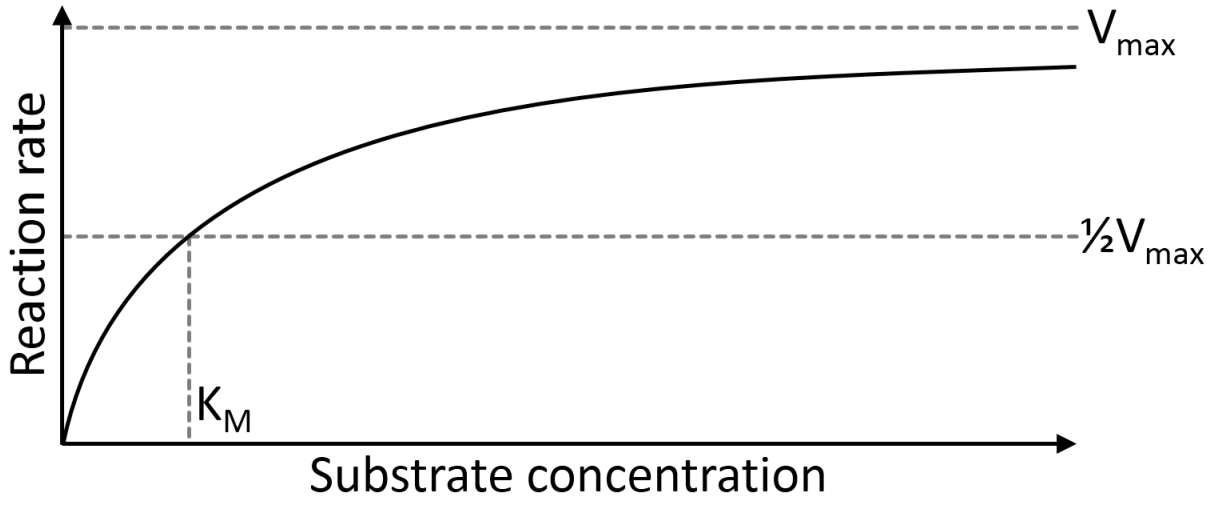
Cofactor/Prosthetic group/Coenzyme

Cofactors are necessary non-peptide components required for enzymes to function properly. Cofactors can either be inorganic metal ions or organic molecules. The assistance of cofactors for enzyme function is achieved by binding to the inactive form of enzyme (apo-enzyme) to produce the catalytic active form (holo-enzyme). A prosthetic group is a type of cofactor that tightly bind to the assisted enzyme and is not easily to be removed. A coenzyme is a specific type of cofactor as they are organic small molecules.

Michaelis-Menten equation

Michaelis–Menten kinetics describe the typical kinetic behaviour of enzymes. The name was given after German biochemist Leonor Michaelis and Canadian physician Maud Menten. The Michaelis–Menten kinetics model describes the rate of enzymatic reactions v in the form of Michaelis–Menten equation showing bellow:

Here, enzyme reaction rate v, the rate of forming product $[P]$, is related with substrate concentration $[S]$ V_{max} describes the maximum reaction rate achieved by the studied system. It would be reached when the substrate concentration is saturated under a given enzyme concentration. The Michaelis constant K_M is numerically equal to the substrate concentration where half V_{max} is reached. In most of the enzyme catalyzing single-substrate reactions, their kinetics behaviours are assumed to fit Michaelis-Menten equation, regardless of further assumptions.



Kinase

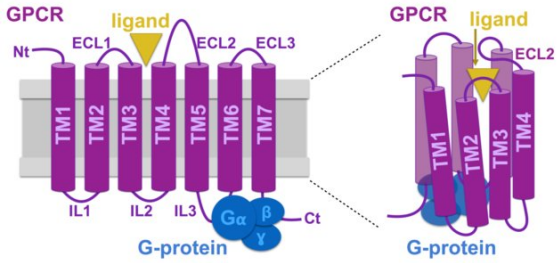
Types of enzyme responsible for substrate phosphorylation.

Receptor tyrosine kinase

Tyrosine kinase is a type of kinase for tyrosine phosphorylation. It functions as an “on” or “off” switch in many cellular signalling process. Receptor tyrosine kinase is a subclass of tyrosine kinase that serves as cell surface receptor with high-affinity for many polypeptide growth factors, cytokines, and hormones.

G protein coupled receptors (GPCR)

A large group of evolutionarily-related proteins serve as cell surface receptors to produce cellular response activation upon signal outside cell. The transmembrane domain of GPCRs pass through the cell membrane seven times (typical structure characteristics of GPCRs). Ligands can either bind at extracellular N-terminus and loops or within the transmembrane helices of GPCR. Effective binding of ligand would cause conformational change. Subsequent dissociation of α subunit from the conjugated G-protein would further facilitate intracellular signal processing.



Catalytic receptor

Type of cell surface protein with the ligand binding site localized at the extracellular surface of the plasma membrane and the functional region possessing catalytic activity on the intracellular face of the plasma membrane. The two parts are linked by a single transmembrane-spanning domain consisting of 20–25 hydrophobic amino acids. It commonly exists and functions as a dimer. Endogenous ligands for catalytic receptor are often peptides or proteins.

Transport protein

A transmembrane protein which function to allow selective passage of specific molecules from the external environment and is able to translocate ions, small molecules, or macromolecules. Transport proteins may be divided into subgroups as channels and carriers.

Carrier protein

Active carrier proteins function in the energy-consumed manner and are able to translocate the substances against concentration gradient. Passive carrier proteins assist the substance by facilitated diffusion.

Ion channel

An ion channel is a type of transmembrane protein that mediates the passage of ions through the membrane. The major differences between ion channels are ion carriers are: (1).high efficiency, usually 10^6 per second (or higher); (2).translocation of ions down their electrochemical gradient in an energy conservation way.

Nuclear hormone receptors (NHR)

A class of transcriptional factors to regulate gene expression regulated by their binding ligands. The ligand binding domain (LBD) is capable of recognizing specific ligands to stimulate conformational change (dimerization) of NHR. The DNA binding domain (DBD) mediates the receptor towards its hormone response elements (HRE). DBD functions in the form of a dimer with each monomer recognizing a six base pair sequence of the targeted DNA.

Ubiquitination

A biological process of protein degradation (intracellularly). The protein is first labelled with a ubiquitin, a 76-amino-acid protein, through a three-step process with help of ubiquitin-activating enzyme (E1), ubiquitin-conjugating enzyme (E2), and ubiquitin-protein ligase (E3), facilitating mono-ubiquitination. The labelled ubiquitin chain could be extended by adding more ubiquitin, resulting in polyubiquitination. The 26S proteasome recognizes the polyubiquitination as a signal to initiate proteolysis and process the protein for degradation.

Proteasome

Huge protein complex to break peptide bonds for unneeded or damaged proteins.

Heat shock proteins (HSP)

Molecular chaperones (proteins) to assist protein functioning in response to stressful conditions (eg. exposure to cold and/or UV light, wound healing etc). HSPs are named according to their molecular weight. HSP90 refers to HSPs which are 90 kilodaltons in size. Ubiquitin (8 kilodaltons) also possess heat shock protein features.

Tubulins

Structural unit for living cell skeletal system. Tubulins are proteins that can be polymerized into long chains or filaments to assemble into microtubules - hollow fibers that serve as cell skeletal system.

Binding Site Detection for Receptors

Not all functional components in our body can be drug targets. However, this doesn't mean they cannot be modulated. Sometimes they are just too hard to be accessed accurately due to their distribution in tissue or a structural factor, while in other cases inhibition of these components cannot trigger the expected downstream reaction due to intrinsic homeostasis / ignorance of its mechanism. In most cases, orthosteric binding sites (the pocket to binding endogenous ligand) can be easily determined by sequence / structure alignment. These site may lack selectivity, rendering growing interest in allosteric site detection. (sites not directly binding the endogenous ligand, but modulate its binding behavior) Traditional methods for allosteric site detection rely on MD simulation. See: Investigating Cryptic Binding Sites by Molecular Dynamics Simulations

- **Orthosteric/Allosteric Regulation** A protein can have endogenous ligands and protein-protein binding partners. If a drug binds the protein in areas directly involved in endogenous binding, its effect on the protein is called orthosteric regulation. If the drug binds other areas (far away) but can affect the behavior in this area, its effect on the protein is called allosteric regulation. Orthosteric regulation is easier to study: such binding can at least compete with endogenous partner, affecting target behavior. Allosteric regulation is much harder to research, requiring dynamic insight to determine the relationship between orthosteric site and the potential allosteric site.
- **Covalent Regulation** Traditionally, a drug molecule binds to the target without a reaction with it. It can bind and dissociate, resulting in a chemical equilibrium. However, some novel types of drugs try to form a chemical bond

with the target, binding to them permanently. Giving the obvious Sequelae effect (the drug effect can maintain a long time after the drug's blood concentration becomes low), this kind of regulation can be both effective and risky.

Immunology

A branch of physiology raising huge interest recently; studies the immune system of human body.

- **Lymphocyte** A type of white blood cell that plays a vital role in immune responses. There two types of lymphocyte: B-cells and T-cells.
- **B-cells and T-cells** B-cells are a type of lymphocyte that are able to produce antibodies. T-cells are involved in cell-killing (directly kill the virus-infected cells), immune response amplification (via cytokines, a signal protein secreted from T-cells) and cell memory that enable an organism to respond to the same infection more quickly and efficiently if infection happen again.
- **Antigen and antibody** The term antigen originally referred to a substance that may trigger an immune response and serves as a antibody generator. Antibodies (or immunoglobulins) are large, Y-shaped protein secreted from B-cells to recognize and neutralize antigens.
- **Complement system** The complement system functions via the cascade involving distinct plasma proteins that react with one another to opsonize pathogens and induce a series of inflammatory responses to fight infection. It works as enhancing and or complementing the effects of antibody activity and is firstly evolved as part of the innate immune system.
- **Cluster of differentiation antigen (CD)** Surface proteins on leukocytes, reflecting differentiation stage or activation state of the cell and can be recognized by specific monoclonal antibodies.
- **Epitope** Epitope is the antigenic determinant lying on the antigens to simulate immune responses. Binding and subsequent reaction of immune cells and antibodies with antigens is initiated via the recognition of epitope.
- **Antigen-presenting cell (APC), Major histocompatibility complex (MHC) and Human leukocyte antigen (HLA)** APCs are cells possessing the ability to present an antigen for T-cell recognition. The heterogeneous group (protein complex) on the APC surface for antigen presentation is called major histocompatibility complex (MHC). There are two type of MHC, class I and class II, differed by structure and expressed cell types. MHC in human is also called human leukocyte antigen (HLA). There is significant work aiming to solve the recognition pattern issues of MHC with presented antigen. AI models have achieved rather ideal accuracy for the prediction task to define whether an antigen (mainly short peptide sequence) could be presented by MHC (thus stimulate the immune reaction from T-cells with much possibility) to design more efficient immune regulators (neoantigen).
- **Cytokines** Cytokines are messenger proteins released from immune cells to regulate immune responses. Abnormal activities of cytokines could induce "cytokine storm" which has lethal impact.

Antigenicity and immunogenicity

When a foreign material (antigen) enters, an organism would initiate a barrier system to fight against and eventually eliminate this intruder. Antigenicity describes the ability of an antigen bind to, or interact with the products of the final cell-mediated response (such as B-cell or T-cell receptors). Immunogenicity measures the ability of the antigen to activate the immune response (including innate immune response and the subsequent adaptive (acquired) immune response). Immunogens possess antigenicty, while antigens may not always have immunogenicity. Metal ions are typically haptens, which are antigens, but would not trigger immune responses.

Monoclonal antibody, vaccine and neoantigen

Monoclonal antibodies are engineered antibodies that typically recognize the same epitope, and thus possesses high specificity towards the targeted antigen. Vaccines are the biological preparation containing an agent to initiate the immune responses to form a barrier thus protect the body from certain disease derived from infection. The agent of vaccine resembles the disease-causing microorganism and is often made from weakened or killed forms of the microbe, its toxins, or the surface proteins. Neoantigens are the translation product (protein) of mutated DNA in cancer cells. They are different from the original protein under physiological condition and may thus play a significant role in stimulating immune response against cancer cells.

Prediction and design of protein-protein interaction

Protein-protein interaction (PPI) is the basis for many biological processes to function properly. Specific recognition between the interacting proteins is established on the basis of physical contacts. The forces driving stable/favourable interaction come from electrostatic interaction, hydrogen bonding and/or the hydrophobic effects etc. Based on the forces performed by atom/atom groups, there exist recognition patterns in the aspects of protein sequence as conserved region formed by amino acids that possess similar physicochemical properties have been observed in certain type of PPI. With the understanding of the interaction forces and their corresponding protein sequences, recognition/interaction patterns of PPIs should be reasonably summarized in relation with their biological outcomes. These summarized patterns in forms of models could be further applied for biological effect prediction with the protein sequences as input. Further, one could design functional protein sequences to achieve the desired bio-activity.

Yield, Solubility, Stability of therapeutic Macromolecules

Therapeutic macromolecules are compounds with large molecular weight possessing therapeutics effects and are typically derived from biological processes. The commonly applied therapeutic macromolecules (macromolecular drugs) include peptides, proteins, antibodies, polysaccharides and nucleic acids. Procedures to collect therapeutic macromolecules include bio-synthesis, recombinant protein expression, conjugation and modification etc. In comparison with small-molecule drugs, stable pipeline construction to yield therapeutic macromolecules requires much more effort. It is also worth noting that, as therapeutic macromolecules are typically derived from biological process in organism, some of them possess favourable intrinsic properties such as being lipophilic/hydrophobic and many of them would be easily degraded when recognized by cell metabolizing systems. Thus, enhancing the solubility of therapeutic macromolecules to facilitate desirable distribution properties as well as to sustain the intact entities thus gain stability to occupy therapeutic window wide enough to take action are another important issues for future discovery.

Immuno-therapy

Immuno-therapy functions by activating/mediating/enhancing immune responses of the patients to fight against diseases (cancer).

Cell therapy

Engineered cells (isolated from patients) with therapeutic effects are injected/grafted/implanted (back) into the patient's body to treat the disease. The most well known cell therapy is CAR-T, where chimeric antigen receptor T cells are genetically engineered to produce an artificial T cell receptor to take action in the way of immunotherapy.

1.5.4 Pharmacy

Pharmaceutics

How chemical entities are transferred to medication.

Pharmacokinetics

Pharmacokinetics studies how organisms process drugs.

- **Absorption (A)**: how drugs get into the bloodstream
- **Distribution (D)**: how drugs are reversibly transferred from one location to another within the body. some drugs tend to concentrate in part of the body like adipose tissue, raising potential risks for clinical usage.
- **Metabolism (M)**: how drugs are broken down and modified inside the body.
- **Excretion (E)**: how drugs and their metabolites (a metabolised form of drugs) are removed from the body.
- **Toxicity (T)**: a pharmacodynamic property of drugs. Since its assessment protocol shares something similar with ADME, they are referred to as a whole in many cases.

Pharmacodynamics

Pharmacodynamics studies what a drug does to the body. Pharmacodynamics focus on the molecular, biochemical and physiological effects or actions of the studied drug.

Medicinal Chemistry

Medicinal chemistry refers to designing and synthesizing small molecules as a pharmaceutical agent.

- **Hit-lead-candidate**: A hierarchical description to describe the potential precursor of a registered medication entity.
- **Hit**: Promising candidates from preliminary screening, typically with a micro-molar EC50 (median effect concentration) values, or top scores from virtual screening.
- **Lead**: Candidates demonstrating further potential to become drugs. Lead compounds normally require extensive modification and assessment before becoming a drug candidate.
- **Drug candidate**: A well-studied compound, showing sufficient evidence in potency, selectivity, safety, and other drug-like properties. Drug candidates will become registered drugs after thorough clinical trials usually including thousands of volunteers, tens of years, and investment in the order of billions of dollars)

Synthetic Route Design

Given a promising drug candidate, plan/design of the synthetic routine is performed to find more efficient processes with suitable starting materials to finally yield the product.

Retrosynthesis

Retrosynthesis is a recursion method to design a synthesis pathway for target organic molecules. The target is split into simpler precursors, and the precursor is split in the same manner until the building blocks are commercially available.

Bioisostere

Chemical substitutions or groups with similar physical or chemical properties to produce broadly similar biological effects. Bioisostere replacement from one compound to another is mainly applied in the situation where the parent compound is unsuitable for safe use, and/or bio-availability etc while possessing ideal druggability characteristics.

Reaction Output Prediction

Predict the product(s) with given reactants. There are reaction rules and preference reaction sites to be followed and studied in order to achieve this goal. Organic reactions are very dirty: a system can react in a different manner (e.g. substitution reaction & elimination reaction share very similar reagents and reaction condition), and at different sites (e.g. there can be two oxhydryls in one reagent, change the reaction condition can influence the preference during reaction.

Patent recognition/Literature information extraction

Collections of patent filings or literature contain a plethora of information that can guide drug development. On the other hand, novel compounds need to avoid violating existing intellectual property. Chemical patents are quite intricate, requiring sophisticated training and significant time to understand. Thus, automated information extraction via computer vision or natural language processing technology is quite necessary, although the evaluation would be a challenge.

Property Prediction for Ligands

A drug-like ligand requires favourable properties concerning pharmacokinetics issues (which means it can arrive the target properly) and pharmacodynamics issues (which means it can act with the target properly). These properties will be checked closely and separately below. However, medicinal chemists have come up with a few straight-forward and system-independent guidelines for drug design, to filter out risky molecules for downstream development. Below are some classic samples:

- Lipinski rules of five (RO5) / bRO5: The most classic drug-like guideline. But these rules are being questioned all the time. See Rule of five in 2015 and beyond: Target and ligand structural limitations, ligand chemistry structure and drug discovery project decisions
- Pan-assay interference compounds (Pan-assay interference compounds) Some molecules can always show a positive signal in high-throughput screening, but so far no one has successfully turned them into registered drugs. Such molecules share some intricate similarity, but how to describe them sharply remains a problem.
- dosage form: the form drug is marketed for use. e.g. capsule, syrup, injection, etc
- excipient: inactivate component in a drug product. The purpose of excipients may be to: make the drug stable, soluble, absorbable; change the absorption behavior(e.g. Controlled-release technique); some new excipients are suggested to change the distribution behavior of drug (like liposome).
- pharmaceutical formulation design: choose the proper dosage form, find out the suitable combination of excipient and their ratio, decide the protocol of manufacturing
- crystal structure of drugs: the arrangement of the molecules in a crystal. It can affect both physical and chemical (rare) properties of drug product.

Bioactivity

Bioactivity refers to the fraction (%) of an administered drug that reaches systemic circulation.

Druggability

If a protein is suitable to be a drug target. Studies on druggability don't focus on 'determination of a good target', but rather on 'how to filter out the unsuitable/difficult targets'. Two 'tangible' sub-project in this topic: if a target can be modulated by small molecules (containing suitable pocket / covalent modification site / etc for molecule binding); if the inhibition / activation of a target can cause downstream (at least cellular-level) changes (rather than be antagonised and eliminated due to intracellular homeostasis)

Druggability Prediction for Receptors

Define whether certain kind of receptors could serve as a drug target that could be specifically addressed by a drug to take action for certain disease. It is also commonly to detect whether newly identified receptors could be targeted by existing drugs for re-orientated therapeutic purposes (under the circumstances of drug re-purposing/repositioning).

Pharmacophore

(according to IUPAC) an ensemble of steric and electronic features that is necessary to ensure the optimal supramolecular interactions with a specific biological target and to trigger (or block) its biological response. In short: when we perform a QSAR study, we consider a part of a molecule as a group, and describe it by it's chemical property (charge, hydrophobicity, aromaticity, steric hindrance, etc). Such group will be 'scored' and replaced as a whole.

Prediction of structure-activity relationship

structure-activity relationship (SAR) is the relationship between the structure and the biological activity. It tries to answer 2 question: 1.which parts in a bioactive compound / which combination between these parts matters (certain pharmacophore in a certain topology / geometry structure); 2.how to modify a molecule according to information gained above (infer a stronger pharmacophore / scaffold to replace the old one)

Molecule Generation

Design a new chemical entity satisfying all demands above (have ideal property, can be synthesised easily, haven't been patented) is considered as the holy grail of drug discovery. Since the inference of the above properties is still underdeveloped, there is still a long way to go for this ambition. However, today's development of generative chemistry models can also serve a practical role in settings like library generation (generate at least novel and 'drug-like' molecule) and conditional design (generate molecule satisfying certain explicit constraint). For more information see in Generative Models for De Novo Drug Design.

Formulation Design

Design of the optimal form of drugs based on the effective compound. Further reading could be referred to: 1.State-of-the-Art Review of Artificial Neural Networks to Predict, Characterize and Optimize Pharmaceutical Formulation; 2.Crystal structures of drugs: advances in determination, prediction and engineering

Regenerative medicine

Regenerative medicine seeks the way to replace the damaged tissues or organs from disease, trauma, or congenital issues, in contrast to the traditional clinical ideas that focus only on alleviating or treating the symptoms.

1.5.5 References

[1] Jakob Schneider, Ksenia Korshunova, Francesco Musiani, Mercedes Alfonso-Prieto, Alejandro Giorgetti, and Paolo Carloni. Predicting ligand binding poses for low-resolution membrane protein models: Perspectives from multiscale simulations. Biochemical and Biophysical Research Communications, 498(2):366–374, 2018. Multiscale Modeling.

CHAPTER

TWO

INDICES AND TABLES

- genindex
- modindex
- search