AI-Driven Creation of a New Paradigm for Scientific Research: Postprint

Authors: Weinan E

Date: 2024-03-27T00:00:00+00:00

Abstract

The purpose of scientific research is to discover fundamental principles and solve practical problems. Although humanity has achieved tremendous success in discovering fundamental principles and solving practical problems, the lack of effective tools and effective scientific research organizational models remains the main bottleneck constraining research efficiency. The rapid development of artificial intelligence (AI) provides new possibilities for changing this situation. In recent years, deep learning methods have demonstrated remarkable success in the field of scientific research, not only helping to solve some core scientific problems and expanding scientific methodology, but also beginning to drive scientific research from the traditional "cottage industry model" to a "platform model". Currently, China has laid a good foundation in the field of AI-driven Science (AI for Science), and should seize the opportunity to strive to lead scientific and technological innovation and contribute to the technological development of humanity.

Full Text

Preamble

Special Issue: Vigorously Promoting Scientific Research Paradigm Transformation

Citation Format: E Weinan. AI helps to establish a new paradigm for scientific research. *Bulletin of Chinese Academy of Sciences*, 2024, 39(1): 10-16, doi: 10.16418/j.issn.1000-3045.20231224001.

This article draws upon the author's paper "AI for Science" published in SIAM News on December 1, 2023, and has been revised and expanded based on that work.

Received: December 27, 2023 | Vol. 39, No. 1, 2024

Abstract

The purpose of scientific research is to discover fundamental principles and solve practical problems. Although humanity has achieved tremendous success in both discovering fundamental principles and solving practical problems, the lack of effective tools and efficient organizational models remains the primary bottleneck constraining research productivity. The rapid development of artificial intelligence (AI) offers new possibilities for changing this situation. In recent years, deep learning methods have demonstrated remarkable success in scientific research, not only helping to solve some core scientific problems and expanding scientific methodology, but also beginning to drive a transformation of scientific research from the traditional "craftsman model" to a "platform model." Currently, China has established a solid foundation in the field of AI-driven science (AI for Science) and should seize this opportunity to strive for leadership in technological innovation and contribute to humanity' s scientific and technological advancement.

Keywords: AI-driven scientific research, scientific computing, Android paradigm

1. Fundamental Problems in Scientific Research

Scientific research has two main purposes: discovering fundamental principles, such as the laws of planetary motion and quantum mechanics, and solving practical problems arising in engineering and industry. There are also two primary approaches: the Kepler paradigm, which is data-driven, and the Newton paradigm, which is driven by fundamental principles. The best example of the former is the discovery of the three laws of planetary motion, where Kepler identified these patterns through analysis of observational data. The best example of the latter is Newton' s explanation and application of these three laws. Newton proposed the second law of mechanics and the law of universal gravitation, which reduced the problem of planetary motion to an ordinary differential equation and led to the derivation of the three laws of planetary motion. While Kepler made the original scientific discovery, he did not understand the underlying reasons; Newton further discovered the fundamental principles behind these phenomena, which could then be applied to many other problems.

From a practical application perspective, the task of finding fundamental principles was largely completed after the establishment of quantum mechanics. As early as 1929, Dirac [1] declared that "the underlying physical laws necessary for the mathematical theory of a large part of physics and the whole of chemistry are completely known, and the difficulty is only that the exact application of these laws leads to equations much too complicated to be solved." His assertion applies not only to chemistry but also to biology, materials science, and all other natural sciences and engineering disciplines that do not involve high-energy physics. In practice, it is often unnecessary to delve down to the quantum mechanics level; instead, simplified fundamental principles can be used, such as the Euler

equations of gas dynamics and the Navier-Stokes equations of fluid mechanics.

However, not all problems in current scientific research have been solved. For example, studying material properties and design, drug design, internal combustion engine design, and many control problems remain far from being solvable using fundamental principles. In these fields, theoretical work is often far removed from the real world, and real-world problems must be solved through trial and error or empirical experience. This leads to low efficiency in scientific research and slow progress in technological advancement in related fields.

All these "difficult" problems share a common characteristic: they depend on multiple independent variables. Thus, these difficulties actually stem from the curse of dimensionality. Taking the Schrödinger equation in quantum mechanics as an example, ignoring symmetry, the number of independent variables in the wave function is three times the number of particles. Therefore, although a system of ten electrons is very simple, its corresponding partial differential equation in 30-dimensional space is already extremely complex!

2. AI Provides New Solutions for Scientific Computing

Deep learning has achieved tremendous success in image classification, image generation, and Go. These are standard AI problems, but from a mathematical perspective, they are actually problems of function approximation, probability distribution approximation and sampling, and solving Bellman equations—all of which are typical problems that applied mathematics, particularly computational mathematics, has long faced. The difference is that these AI problems have much higher dimensions than those encountered in applied mathematics. For instance, in image classification, the independent variable is the image, with each pixel representing one degree of freedom. Thus, a 32$\times$32 color image has 3,072 degrees of freedom; in other words, the problem dimension is 3,072.

The success of deep learning on these high-dimensional problems suggests that deep neural networks may be more effective tools for approximating high-dimensional functions. Although a complete mathematical theory of deep learning has not yet been established, important progress and intuitive understanding have been achieved. First, neural networks are a special class of functions. If one uses piecewise linear functions on a regular grid to approximate a function, the error is proportional to the square of the grid size. This is precisely the source of the curse of dimensionality: as the dimension increases, the number of grid points required for the same grid size grows exponentially. This applies not only to approximations based on piecewise linear functions but to all approximation methods based on fixed basis functions. However, if neural network functions are used to approximate general functions, it can be proven that in at least some cases, the approximation accuracy does not deteriorate with increasing dimension, much like the Monte Carlo method for numerical integration [2].

This observation has broad implications. Since functions are one of the most

basic mathematical objects, a new high-dimensional function approximation tool will have profound impacts on many different fields. In particular, deep learning should help solve those problems previously discussed that suffer from the curse of dimensionality. This is the starting point for AI-driven science (AI for Science).

The most successful example in this regard is the AlphaFold algorithm for predicting protein structures. Protein structure is one of the most fundamental problems in biology. The conventional approach to studying protein structure involves minimizing the total potential energy of the entire protein-solvent system. However, two major difficulties limit the success of this method: obtaining a sufficiently accurate potential energy function and the complexity of the function landscape. Scientists have also attempted data-driven methods, but their success was limited to predicting secondary structures such as α-helices and β-sheets. By fully leveraging protein sequence datasets and state-of-the-art deep learning models, DeepMind developed the AlphaFold2 algorithm, which elegantly and essentially solved the protein structure problem [3]. This research shocked the world.

AlphaFold2 is a purely data-driven method, but this does not mean that AI for Science represents a purely data-driven research paradigm. In fact, scientific research follows fundamental or first principles as described earlier, and a major component of AI for Science is using AI methods to develop more efficient algorithms or approximate models for these fundamental principles. The most famous example in this regard is molecular dynamics. Molecular dynamics is a fundamental tool in biology, materials science, and chemistry. The idea is to study the properties of molecules and materials by computing the dynamic trajectories of atoms in a system. Atomic motion follows Newton' s laws, and the difficult part comes from simulating the interatomic interaction forces or potential energy functions.

The empirical potential approach attempts to guess the functional form of interatomic potential energy functions and then fit parameters using data from experiments or first-principles calculations. While this approach can provide some assistance, it is unreliable as a quantitative tool for studying specific systems. In 1985, Car and Parrinello developed the first AI method based on first principles: using quantum mechanical models (such as density functional theory) to calculate interatomic forces in real time. This approach enables simulations of specific systems with first-principles accuracy, but in practice, efficiency is the bottleneck. Due to efficiency limitations, this method can only handle systems with thousands of atoms.

Machine learning proposes a new paradigm. In this new paradigm, quantum mechanics is used only to provide data. Based on this data, machine learning methods can derive accurate approximations of interatomic potential energy functions. The resulting model is called Deep Potential Molecular Dynamics (DeePMD). It is a reliable atomic simulation tool with first-principles accuracy. Combined with high-performance computing, it has extended the capa-

bility of first-principles-accuracy molecular dynamics simulations from systems of thousands of atoms to 17 billion atoms [7]. The DeePMD software package, DeePMD-kit, has also greatly lowered the barrier to using DeePMD [8].

Similar ideas can be applied to other physical models. For example, more general and accurate density functional models can be trained using highly accurate quantum chemistry calculation data. More accurate and reliable coarse-grained molecular dynamics models can be developed, as well as more accurate matrix models for kinetic equations. In fact, machine learning is precisely the tool that was missing from past multiscale, multiphysics modeling efforts [6].

Beyond models of fundamental principles, AI methods can also provide more efficient and accurate inversion algorithms, thereby enhancing experimental characterization capabilities. The AI-based algorithms discussed previously can provide more realistic and accurate data for forward problems, while the differentiable structure in neural networks can help design optimization or sampling algorithms for inverse problems. This work is still in its early stages, but it has enormous potential for development.

AI methods may also change how people utilize literature and existing scientific knowledge. Literature and existing scientific knowledge are among the primary sources of scientific inspiration, but effectively using these resources is also a very challenging task: it requires mining relevant literature and knowledge from vast amounts of information and spending considerable time reading and studying them. However, AI databases and large language models can be used to collect, integrate, and more effectively query this information. In principle, for any research topic of interest, AI tools can be used to quickly summarize relevant information from the literature and its sources. AI technology can even help suggest further research directions. This will greatly improve the efficiency of scientific research.

2.1 The ELT Framework

To make this strategy truly effective, two important issues must be addressed. First is network architecture: it should be scalable and follow fundamental physical laws. Scalability enables machine learning on small systems with application to larger ones. This problem was solved in the classic work of Behler and Parrinello. Following physical laws means preserving symmetries, conservation laws, invariances, and other physical constraints. For potential energy functions, the main considerations are translational, rotational, and permutational invariance. This can be achieved using an embedding network that maps information about atomic positions to a set of symmetry-preserving functions [4]. Then an approximation network is used to fit the potential energy function.

Second is data-related issues. On one hand, if the potential energy function generated by machine learning methods is to be as accurate and reliable as the original quantum mechanical model in all practically relevant scenarios, the training dataset must be sufficiently representative of all these different scenarios.

On the other hand, since annotated data is computed using quantum mechanical models, which are relatively expensive, we want the dataset to be as small as possible. This requires an adaptive data generation algorithm that can help AI dynamically generate the "optimal" dataset during the learning process.

The algorithm consists of three components: exploration, labeling, and training —hence the name ELT. ELT can start with no data and a rough initial potential energy function. During exploration, sampling algorithms (such as some molecular dynamics method) are used to explore different atomic conformations. For each encountered conformation, an indicator value can be computed to determine whether it needs labeling. The labeled data is then added to the training dataset, and the approximation of the potential energy function is updated regularly based on it.

The key to the algorithm lies in the sampling scheme and how to compute the indicator value. The basic idea of the sampling scheme is to explore only the conformational space that is practically interesting and lacks sufficient training data. The key to the indicator value is to identify which conformations lack sufficient nearby training data. For the latter, the ELT scheme adopts the approach of training a set of approximate potential energy functions. The standard deviation among these approximate potential energy functions is defined as the indicator function. For a currently sampled conformation, if its indicator function value exceeds a threshold, the conformation is labeled. The logic behind this is that if there were sufficient training data near this conformation, all network-predicted potential energy function values would be very accurate and close to each other. A large standard deviation indicates that there is not enough training data nearby, so the current conformation should be annotated and added to the training dataset. For the sampling algorithm, biased molecular dynamics is chosen, where the bias potential is defined by the current approximation of the potential energy function, with weights defined by the size of its confidence interval. The logic is that if the obtained potential energy function is sufficiently accurate in a certain region, we should leave this region and sample elsewhere [6].

With these main components, it is indeed possible to provide potential energy functions with first-principles accuracy for a large class (if not all) of atomic systems. As these new possibilities emerge, we can explore a new research paradigm, which we call the "Android paradigm" for scientific research. In this new paradigm, the scientific community will work together to build a new infrastructure, including AI algorithms for fundamental principles, AI-enabled experimental facilities, and new knowledge databases. These platforms constitute the "Android platform" for scientific research. Whether searching for catalysts in specific chemical reactions or designing new batteries, these application-specific research efforts can be conducted on this "Android platform." This will undoubtedly accelerate the process of scientific research.

This perspective of horizontal integration will also help break down disciplinary barriers and strengthen interdisciplinary research and education. The perspec-

tive of horizontal integration itself is not new, but due to the lack of effective tools, it has been difficult to achieve substantial progress in the past. As mentioned earlier, AI methods provide significant room for improving these horizontal tools. These new horizontal tools, such as platforms for accessing literature and existing scientific data, as well as automated and intelligent experimental platforms, enable researchers to view different research scenarios more effectively from a horizontal perspective. For atomic systems, biology focuses on biological macromolecules, materials science on condensed matter systems, chemistry on small molecules, and chemical engineering on polymers. From the perspective of theoretical tools, all these systems rely on electronic structure methods and molecular dynamics methods. Experimental tools include spectroscopy and microscopy imaging techniques at different scales. Although different fields focus on different systems, the tools and knowledge from these different fields should be shared to the maximum extent possible. In this framework, the boundaries between disciplines naturally disappear.

3. Current State of AI for Science Development in China

With this vision, our team launched the DeepModeling open-source platform in 2018. The purpose of this platform is to invite the scientific community to work together to build infrastructure for physical modeling and data analysis based on AI methods. To date, it has generated tremendous impact and attracted many developers. In China, the development of AI for Science presents a promising and encouraging landscape. All these developments have laid a good foundation for AI for Science in China.

1. In just a few years, the importance of AI for Science and the enormous development space it brings have gained widespread recognition. A large number of leading scholars in various fields attach great importance to this opportunity. The organization of the special issue "Vigorously Promoting Scientific Research Paradigm Transformation" by *Bulletin of Chinese Academy of Sciences* in early 2024 is one example.

2. A number of research teams dedicated to AI for Science are emerging and showing good momentum. After more than three years of preparation, the AI for Science Institute, Beijing was officially established in September 2021 with the support of the Beijing municipal government. This is the first research institution in the world dedicated to AI for Science, committed to building the infrastructure for the AI for Science era. In addition, there are the Machine Chemist team at the University of Science and Technology of China and the AI for Electrochemistry team at the Xiamen University Tan Kah Kee Innovation Laboratory.

3. A number of enterprises are also actively deploying in the AI for Science direction. This reflects the tremendous confidence of industry in AI for Science. Under the banner of AI for Science, a large group of capable, determined, and motivated young industrial practitioners has gathered.

4. National agencies such as the Ministry of Science and Technology and the National Natural Science Foundation of China, as well as local governments including Beijing and Shanghai, are actively formulating policies to support AI for Science research. In 2022, the Interdisciplinary Science Division of the National Natural Science Foundation of China first launched the "Major Research Plan for Interpretable and Generalizable Next-Generation Artificial Intelligence," with AI for Science being an important component.

4. Recommendations

Today' s solid foundation does not mean that the healthy development of AI for Science in China is guaranteed. For the development of any field, becoming a hot topic is a double-edged sword. The hotter the topic, the more likely bubbles are to form. How can we ensure that we seize this opportunity and enable AI for Science to lead China to the forefront of the next wave of technological innovation and industrial transformation? This article proposes the following four specific recommendations.

1. **Highly forward-looking top-level design is needed.** Top-level design must prioritize infrastructure construction. Infrastructure construction has long cycles, heavy tasks, and great difficulties, but its importance for long-term development is unquestionable. In recent years, we have witnessed examples where long-term superficial prosperity in some fields was shattered overnight. When compared with advanced countries, a huge gap emerges. The root cause is always insufficient effort in infrastructure.
2. **A rational resource allocation mechanism is needed.** Resources should be allocated to capable, motivated researchers who are truly active on the front lines. The negative impact of an irrational resource allocation system is not just waste of resources but also the root cause of unhealthy academic practices. We must completely break the resource allocation system based on seniority, publicity, connections, and "dividing the cake."
3. **The concept of openness and win-win cooperation should be actively promoted.** Scientific research has always been a common cause for all researchers. Under the new framework of AI for Science, the "self-sufficient, small-scale craftsman" research model will hardly meet future development needs. Only through win-win cooperation can we fully mobilize the potential and enthusiasm of researchers and accelerate the improvement of overall scientific and technological innovation capabilities.
4. **Academic atmosphere construction should be strengthened.** Academic atmosphere is one of the most important factors determining whether China' s scientific and technological innovation can succeed and whether AI for Science can develop smoothly in China. We should actively encourage young people to propose new ideas and concepts, encourage questioning and challenging various academic viewpoints,

and actively advocate a pragmatic and truthful atmosphere. Academic conferences and discussions should return to their original goals. The practice of making false propaganda and painting rosy pictures for leaders should lose its space for survival in China.

It is hoped that Chinese scientists will cherish the current good development momentum of AI for Science, cooperate closely, firmly seize this once-in-a-lifetime opportunity of AI for Science, strive to take the lead in the next wave of scientific and technological innovation, and make due contributions to humanity' s scientific and technological development.

References

1. Dirac P A M. Quantum mechanics of many-electron systems. *Proceedings of the Royal Society A*, 1929, 123(792): 714-733.
2. E W N, Wang Q. Understanding neural network-based machine learning: What we know and what we don' t. 2020, doi: arXiv:2009.10713v3.
3. Jumper J, Evans R, Pritzel A, et al. Highly accurate protein structure prediction with AlphaFold. *Nature*, 2021, 596: 583-589.
4. Zhang L F, Han J Q, Wang H, et al. End-to-end symmetry preserving inter-atomic potential energy model for finite and extended systems//Proceedings of the 32nd International Conference on Neural Information Processing Systems. New York: Curran And Associates, Inc., 2018: 4441-4451.
5. Zhang L F, Wang H, E W N. Reinforced dynamics for enhanced sampling in large atomic and molecular systems. *The Journal of Chemical Physics*, 2018, 148: 124113.
6. E W N. Machine learning and multiscale modeling. *Physics Today*, 2021, 74(7): 36-41.
7. Lu D H, Wang H, Chen M G, et al. 86 PFLOPS deep potential molecular dynamics with ab initio accuracy to 10 billion atoms//Proceedings of the 27th ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming. Seoul: ACM, 2022: 205-218.
8. Zeng J Z, Zhang D, Lu D H, et al. DeePMD-kit v2: A software package for deep potential models. *Journal of Chemical Physics*, 2023, 159(5): 054801.

Note: Figure translations are in progress. See original paper for figures.