Artificial Intelligence and Big Data Analysis of Electron Cloud Density

Lu Xu[1,4]*, Qin Yang[2,3]*

[1]School of Sports and Health Science, Tongren University, Tongren 554300, PR China

[2]School of Physics and Optoelectronic Engineering, Yangtze University, Jingzhou 434023, PR China

[3]Neonatal Screening Center, Shanghai Children's Hospital, Shanghai Jiao Tong University, Shanghai, 200040, PR China

[4]College of Material and Chemical Engineering, Tongren University, Tongren 554300, PR China

Review Article

Received: 08-May-2023, Manuscript No. JCHEM-23-98053; **Editor assigned:** 10-May-2023, Pre QC No. JCHEM-23-98053(PQ); **Published:** 07-Jun-2023, DOI: 10.4172/2319-9849.12.2.001.

***For Correspondence:**

Lu Xu, College of Material and Chemical Engineering, Tongren University, Tongren 554300, PR China

E-mail: lxchemo@163.com

Qin Yang, School of Physics and Optoelectronic Engineering, Yangtze University, Jingzhou 434023, PR China

E-mail: yqbioinformatics@126.com

ABSTRACT

This review tries to explain about a new Quantitative Structure-Activity Relationship (QSAR) method, Deep Electron Cloud-Activity Relationships (DECAR) and Deep Field-Activity Relationships (DFAR). It also outlooks the prospects of the international Lab for Molecular Deep Electron Cloud-activity Relationships (LabMolDECAR) under development in Tongren University. It is one of the core tasks for chemists to investigate the properties and activities for millions of chemical molecules with diverse structures.

Keywords: Quantitative structure-activity relationship; Deep electron cloud-activity relationships; Deep field-activity relationships; Theoretical chemistry; Electron cloud density analysis

INTRODUCTION

A great deal of attempts and efforts have been paid on modelling and revealing the relationships between molecular structures and their properties by theoretical and computational chemists [1]. However, there is still a lack of widely applicable laws or rules when it comes to predicting the properties/activity of molecules [2].

In order to develop general models predicting various molecular properties/activities, we recently proposed the concepts and methods of Deep Electron Cloud-Activity Relationships (DECAR) and Deep Field-Activity Relationships (DFAR) by combining Artificial Intelligence (AI), rigorous quantum chemistry and big data analysis [3]. The corresponding Quantitative Structure-Activity Relationship (QSAR) models have been established, which theoretically are possible to achieve reliable prediction of various molecular properties. Based on DECAR/DFAR, the international Lab for Molecular Deep Electron Cloud-activity Relationships (LabMolDECAR) is currently under development in Tongren University (Tongren, Guizhou, China), which is expected to further promote the integration of AI and computational chemistry.

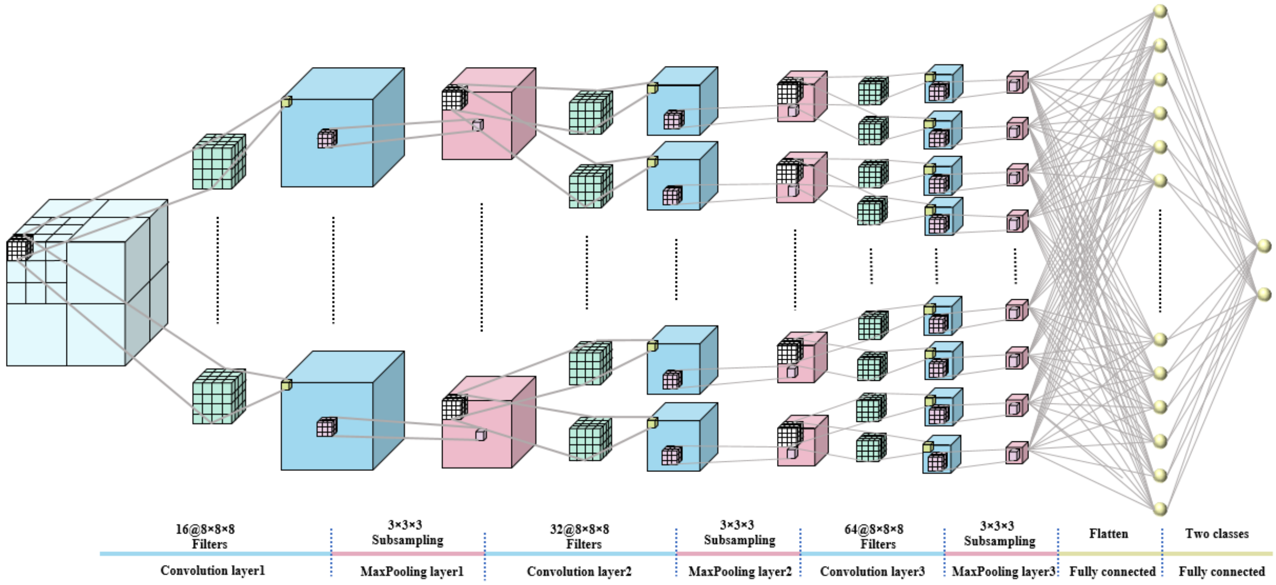
LITERATURE REVIEW

In order to explain the principles of DECAR/DFAR and revealing the prospects of LabMolDECAR, we have tried to describe concepts involved on DECAR/DFAR.

Relationship between DECAR/DFAR

Briefly, DECAR/DFAR is a QSAR method, which is based on strict Density Functional Theory (DFT), deep learning, and very large molecular data sets. In DECAR/DFAR, high-quality three-dimensional (3D) Electron Cloud Density (ECD) or field data are calculated to learn and infer the properties or activities for a large number of molecules using deep neural networks (Figure 1). DECAR/DFAR consists of three essentials. Firstly, a large number of molecular entities (thousands, tens of thousands or even millions, and possibly as many as to challenge the limits of our computational resources) and known molecular activity data are required as the inputs of machine learning; secondly, a rigorous quantum chemistry approach (e.g., density functional theory) is needed to calculate high-quality 3D ECD (or associated field data) as the accurate Molecular structure descriptors; lastly, a machine learning method or model is required that is sufficiently powerful and flexible to learn very large data e.g., a 3D convolutional deep neural network as shown in Figure 1.

Figure 1. Principle of deep electron cloud-activity relationships.



Relationship between DECAR/DFAR and previous works

For example, Linus Pauling's hybrid orbital theory and Kenichi Fukui's frontier molecular orbital theory are classic and landmark works. Undoubtedly, DECAR/DFAR are inspired by previous works, especially hybrid orbital theory [4] and frontier molecular orbital theory [5]. These works can be regarded as the pioneers of ECD analysis. For either hybrid orbitals or frontier molecular orbitals, the essence of orbitals is an approximate and local ECD distribution.

DECAR/DFAR goes further in the direction indicated by these classic works by analysis of high-quality, global-local ECD (field) distributions for a large number of molecules, so the corresponding models can explain and predict more types of molecular properties. We believe that hybrid orbital theory and frontier molecular orbital theory represent the best intuitive intelligence and the most profound insights into chemistry of our human minds, while DECAR/DFAR represents the AI of computers in the era of big data analysis. The former is of great benefit in understanding the essence of molecular structures and properties intuitively, while the latter might provide a useful tool for systematic interpretation and prediction of molecular properties by analysis of big data.

Use of AI or machine learning methods in DECAR/DFAR

Quantum mechanics is one of the most fundamental theories for chemistry, and is also a basic tool for calculating and explaining molecular properties. The basis of theoretical or computational chemistry is quantum mechanics, which plays a basic role in calculating and explaining the properties of molecules. However, there are many problems by using quantum mechanics to investigate molecular properties [6,7]. It may be related to the intrinsic empiricism and ambiguity of chemistry, which means that quantum mechanics is often used to calculate molecular properties with clear and definite physical actions/relationships. However, many of the chemical properties that we are interested in are defined empirically or observed experimentally. For example, does a molecule have anti-cancer or anti-COVID-19 activity? Or, is a molecule toxic and carcinogenic? There is no doubt that these properties are essentially caused by physical (including chemical) interactions, but it is difficult to describe or explain them in terms of precise and clear physical interactions/relationships. The mentioned empiricism or fuzziness in chemistry is largely caused by the complexity and fuzziness of molecular activity mechanisms. For instance, the variety of molecular structures is innumerable, and their targets are often unknown and not unique, therefore, the involved physical (including chemical) interactions are very complicated. It is thus difficult for quantum mechanics to directly calculate or accurately explain these molecular properties.

Based on the above considerations, we believe that machine learning methods are essential to explain or predict molecular properties, which could be learned and inferred from a large number of known molecular structure-property examples. Hence, DECAR/DFAR is essentially a kind of machine learning-based QSAR.

Main criticism as lack of interpretability

The interpretability of a model is of great importance. Predictive capability and interpretability are two essential aspects of a QSAR model. According to the strict tradition of theoretical physics, either a model, or a law/rule, if it does not have broad or general predictive powder in itself, the explanation based on it for a certain phenomenon or fact must be vague or approximate. A significant problem in chemistry is that there are no general principles or laws that can be widely used to predict molecular properties. DECAR/DFAR would provide a feasible method for widely predicting molecular properties. We believe that a more accurate interpretation for molecular properties is possible only if a widely applicable QSAR (or other) model is achieved.

DISCUSSION

Actually, for all DECAR/DFAR models, a general assumption or interpretation is that “the shape and distribution of the global-local ECD/field determines a certain property of a molecule”. As for the specific property and how they correspond to the shape and distribution of an ECD, this can be explained and understood from the shapes of the 3D convolutions. This is also reflected in our previous paper, which gives a way to explain the 3D convolutions.

Advantages of DECAR/DFAR and the traditional QSAR models

Firstly, DECAR/DFAR requires as many input molecular entities as possible. We anticipate that tens of thousands, hundreds of thousands, or even millions of molecules, would be used by DECAR/DFAR as data accumulates. However, traditional QSAR mostly includes dozens and hundreds of molecules for machine learning. Obviously, various molecular properties can be broadly predicted only with the full consideration of the diversity of molecular structures and the learning of sufficient molecular entities. In general, traditional QSAR cannot reliably predict molecular properties that have different parent structures from those in the training set.

Secondly, the molecular descriptors of DECAR/DFAR are high-quality 3D ECD or related field data calculated by strict quantum chemistry method, which are more accurate and informative than most traditional molecular descriptions. So far, at least thousands of molecular descriptors have been developed, the vast majority of which are approximate and empirical. Traditional molecular descriptors have played a certain role in different periods and even up to now. Some of them are very intuitive and practical, and can be calculated conveniently and quickly. However, in the field of machine learning, there is a saying called "garbage in, garbage out". This does not mean that traditional molecular descriptors are worthless, but rather that such approximate molecular descriptors may be difficult to generate reliable and accurate QSAR models and prediction results, at least for molecules with significant differences in atomic composition and structure. As known to all, the actual structures of molecules are very complex. According to the Hohenberg-Kohn theorem [8], the 3D ECD of the ground-state molecule determines all its properties. As long as the quantum computing method is accurate enough, the 3D ECD or field data may be the most informative and accurate chemical molecular descriptors so far. Therefore, according to the Hohenberg-Kohn theorem, DECAR/DFAR can theoretically predict many molecular properties simultaneously (if not all).

Finally, DECAR/DFAR adopt 3D convolutional deep neural networks. Of course, there are some QSAR studies that use deep learning now, but we have not found deep learning based on a large number of strict and high-quality data or precise descriptors in the literature. The accumulation of molecular data can only be a gradual process. In our recent DECAR/DFAR paper, the number of molecules processed was approximately 3000. Considering the data have been expanded six times, the final amount of data for learning reached around 18000, with each data consisting of 8 million data points (with an ordinary accuracy). It can be foreseen that the data volume will continue to increase in the future. Therefore, one possible advantage of DECAR/DFAR comes from AI and deep learning, which have shone brightly in the classification of massive images and point clouds [9]. Moreover, the more effective data it learns, the better the model predicts. Then, if the ECD is regarded as an accurate image of the molecular structure, DECAR/DFAR is also very likely to achieve such success in the chemical field.

Calculation, data and computation of DECAR/DFAR

As DECAR/DFAR requires the ECD or field data for tens, hundreds of thousands or even millions of molecules. The computational cost of DECAR/DFAR is really very large, but its computational efficiency is high and the return is high as well. At present, the main computational cost of ECD or field data lies in the geometric structure optimization of molecules and wave function calculation. Although the time (mainly determined by molecular size and computational accuracy) is much longer than that of traditional molecular descriptors, the computation of millions of molecules can be run in parallel (by different countries, research institutions, researchers and computers) and it is entirely possible to complete and implement. Why is the calculation highly rewarding? As long as the calculated ECD is accurately enough, the corresponding molecular data can be used for learning and inference of other properties without recalculating for the same molecule, so that the data is actually accumulative, shareable, reusable and highly rewarding.

Of course, the ECD data of each molecule includes at least several million data points (depending on the molecular size and model accuracy), and thousands of molecules are indeed a considerable computational burden for deep learning at present. Similar to the computation of ECD, deep learning offers high returns and high utilization. Firstly, the network is portable and updatable, and the trained network can also be used for learning more molecular data; secondly, the network can be shared widely among researchers. So, the computational cost of deep learning is worthwhile. The more effective data are used to train, the more reliable and robust the network is, and therefore the more time it takes is completely reasonable. Lastly, the hardware for deep learning, especially GPU clusters, is updated quickly, so the computational problem for deep learning can also be solved.

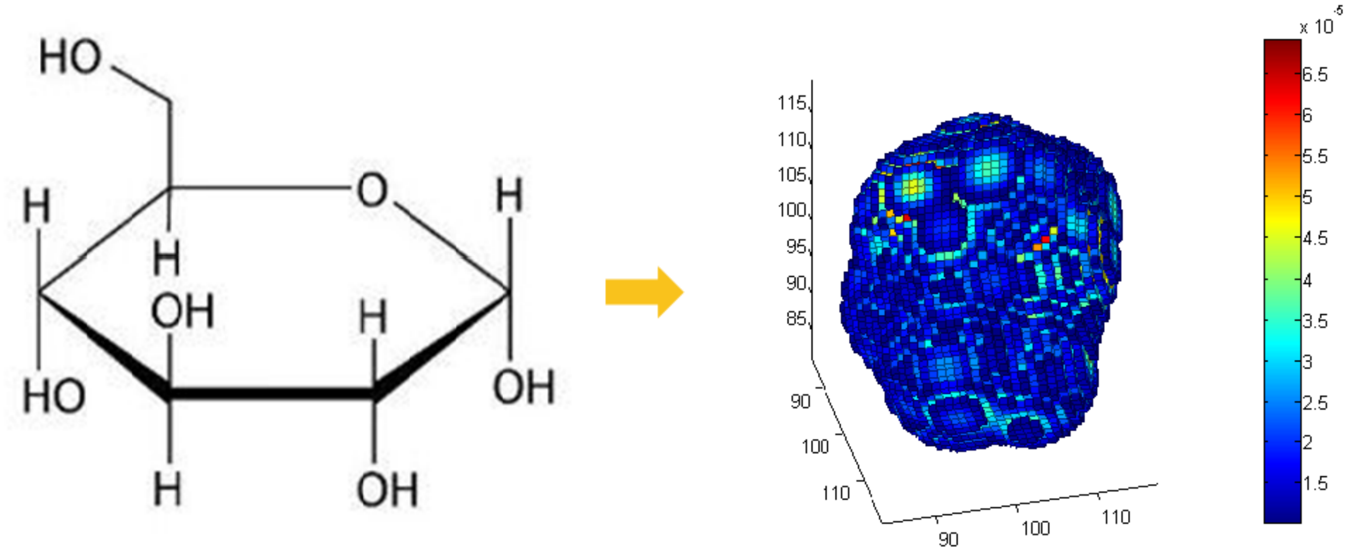
Operation of international lab for molecular deep electron cloud-activity

International cooperation is essential for data accumulation and updating/sharing of networks. Firstly, DECAR/DFAR need to use a large amount of molecular ECD/field data. As mentioned earlier, this should be done in parallel on the servers of different countries, research teams, and researchers. One of the roles of LabMolDECAR (Tongren University) is to generate a database of ECD/field data for millions of molecules in the same way as Wikipedia does (academics jointly provide, supervise, review, update, annotate, and share the data). Secondly, the network of DECAR/DFAR deep learning will be shared and updated among different researchers. Generally speaking, the database of molecular ECDs and fields is like a huge image bank of chemical molecules, and the task of DECAR and DFAR is to learn and predict the type (molecular activity) of these massive images.

Copying of current deep learning techniques in the field of AI or machine learning

We assume DECAR/DFAR also poses higher requirements and challenges for deep learning techniques. Firstly, in image recognition, both the pictures and the objects to be recognized in the picture are scalable (scaling does not affect the type of objects). But, the chemical properties of molecules are both related to their sizes and ECD distributions. That is, two molecules (ECDs) that are geometrically similar can never be considered to have the same chemical properties. Therefore, molecular ECD (field) data are not scalable. For the given accuracy of the ECD, the size of the training network data depends on the largest molecule. The transfer or inference of trained networks of new data requires consideration of this problem. Secondly, the 3D rotation of molecules may require a much larger scale of data expansion, and each rotation and translation of a molecule may theoretically yield an infinite number of equivalents. Developing a reasonable data expansion strategy is therefore necessary, taking into account representativeness, adequacy and feasibility. For the molecules or activities with high complexity, it is necessary to develop computational operations or transformations invariant to the rotation (translation) of the 3D ECD, which is still one of the research frontiers in deep learning. Thirdly, it involves the training problem of the network. Considering that the data are very large for molecules including tens of atoms, an ECD of moderate accuracy may require the acquisition of millions or even tens of millions of data points, which is much larger than the vast majority of current image data, as shown in Figure 2. There may be redundant information in the data. Consequently, the problems such as gradient disappearance and overfitting encountered in the training of deep neural networks may be more serious, which are still the frontiers of deep learning and need to be further explored and studied.

Figure 2. Characterization of a molecular structure by electron cloud density.



DECAR/DFAR in computational chemistry and cheminformatics in the future

It is an incremental process. With the accumulation of a large number of molecular ECDs (field data) and related activity data, hundreds of deep learning networks based on big data will emerge, each of which acts as an AI machine that can predict a molecular property or activity of interest [10].

CONCLUSION

Finally, considering that the famous AlexNet network has already learned and classified 1000 classes of images, could we anticipate a single network in the future that can simultaneously learn and predict various (hundreds or thousands of) molecular properties/activities? We feel optimistic about this. This study explained about the relationships between molecular structures and their properties in both theoretical manner and computational manner, it is necessary to develop computational operations. Also, the frontiers of deep learning and need to be further explored and studied. Further studies are under development which enable us to identify future aspects involved in DECAR/DFAR.

REFERENCES

1. Wasielewski MR, et al. Exploiting chemistry and molecular systems for quantum information science. Nat Rev Chem. 2020;4:490-504.
2. Dral PO. Quantum chemistry in the age of machine learning. J Phys Chem Lett. 2020;11:2336-2347.
3. Xu L, et al. Deep electron cloud-activity and field-activity relationships. Research Square.

4. Harris ML. Chemical reductionism revisited: Lewis, pauling and the physico-chemical nature of the chemical bond. Stud Hist Philos Sci. 2008;39:78-90.
5. Lowdin PO. Advances in quantum molecular sciences-a tribute to Kenichi Fukui. J Mol Struc-THEOCHEM. 1983; 103:3-24.
6. Butler KT, et al. Machine learning for molecular and materials science. Nature. 2018;559:547-555.
7. Fedik N, et al. Extending machine learning beyond interatomic potentials for predicting molecular properties. Nat Rev Chem. 2022;6:653-672.
8. Hohenberg P, et al. Inhomogeneous electron gas. Phys Rev. 1964;136:864-871.
9. Ghahremani P, et al. Deep learning-inferred multiplex immunofluorescence for immunohistochemical image quantification. Nat Mach Intell. 2022;4:401-412.
10. Krizhevsky A, et al. Imagenet classification with deep convolutional neural networks. Commun ACM. 2017;60:84-90.

Citation: Xu L and Qin Yang, et al. Artificial Intelligence and Big Data Analysis of Electron Cloud Density. RRJ Chemist. 2023;12:001.