DOI: 10.14529/jsfi200302

Perspectives on Supercomputing and Artificial Intelligence Applications in Drug Discovery

Jun Xu[1,2]***, Jiming Ye***[1]

This review starts with outlining how science and technology evaluated from last century into high throughput science and technology in modern era due to the Nobel-Prize-level inventions of combinatorial chemistry, polymerase chain reaction, and high-throughput screening. The evolution results in big data accumulated in life sciences and the fields of drug discovery. The big data demands for supercomputing in biology and medicine, although the computing complexity is still a grand challenge for sophisticated biosystems in drug design in this supercomputing era. In order to resolve the real-world issues, artificial intelligence algorithms (specifically machine learning approaches) were introduced, and have demonstrated the power in discovering structure-activity relations hidden in big biochemical data. Particularly, this review summarizes on how people modernize the conventional machine learning algorithms by combing non-numeric pattern recognition and deep learning algorithms, and successfully resolved drug design and high throughput screening issues. The review ends with the perspectives on computational opportunities and challenges in drug discovery by introducing new drug design principles and modeling the process of packing DNA with histones in micrometer scale space, an example of how a macrocosm object gets into microcosm world.

Keywords: drug discovery, big data, artificial intelligence, HPC.

1. Big Data and Supercomputing Challenges in Drug Discovery

In the last century, three cutting-edge inventions, which were combinatorial chemistry (CC), polymerase chain reaction (PCR), and high-throughput screening (HTS), significantly changed biomedical science and technology. CC was invented by Robert Bruce Merrifield who won 1984 Nobel Prize for Solid Synthesis [26], and made high throughput syntheses (a method for scientific experimentation using robotics, data processing/control software, liquid handling devices, and sensitive detectors allows a researcher to quickly make millions of chemicals for biological tests) become possible [19]. PCR was invented by Kary Banks Mullis who won 1993 Nobel Prize [31], and expedited human gene project. HTS was invented by Donald J. Cram, Jean-Marie Lehn and Charles J. Pedersen, who jointly won 1987 Nobel Prize in chemistry for their development and use of molecules with structure-specific interactions of high selectivity. HTS significantly accelerated screening huge number of compounds against biological targets. These inventions trigged high throughput science and technology and revolutionized pharmaceutical discovery and development. Because people now can make chemical compounds, biopolymers and validate their biological properties in high throughput manner. Consequently, human being is facing big data and supercomputing challenges in modern time.

Drug discovery and development involve in the following major processes: molecular design, biological or chemical syntheses, molecular structural elucidations and pharmaceutical analyses, pharmaceutical target identification and validation, drug screening, preclinic experiments and clinic trials, pharmacokinetics (PK) and pharmacodynamics (PD) analyses, disease diagnoses, and clinic drug applications. Each process involves instrumental measurements that result in big data. These data are not only "big" (volume from GB to PB), but stored in many different formats (variety) and required prompt analyses (velocity).

[1]School of Biotechnology and Health Sciences/Center for Biomedical Data Research, Wuyi University, China
[2]Research Center for Drug Discovery, School of Pharmaceutical Sciences, Sun Yat-Sen University, China

There are mainly four sources contributing to the big data in drug discovery:

1. **High throughput experiments.** High throughput syntheses can generate many data describing molecular structures and properties and high throughput screening campaigns can generate many data regarding the relations of the compounds and their biological targets.
2. **Health information / office automation.** These resources contain patients information regarding demographic, administrative, health status / risks, medical history, current management of health conditions, and outcomes data.
3. **Scientific publications, patents, and databases.** Publications in life sciences grow rapidly. PubMed collects more than 30 million biomedical articles from more than 7,000 journals; by August 2020, American Chemical Abstracts (ACS) collects more than 100 millionth compound, 64 million gene sequences. The number of US patent applications reaches 15394762 since 1963. Databases ChEMBL, PubChem, ChemSpider, ZINC, and SureChEMBL collect 2, 90, 63, 980, and 17 million compounds [34]. These data cannot been utilized or digested without computational or artificial intelligence approaches.
4. **Simulations**. Along with the increasing computing power, we can simulate greater and more complicated biological systems. Taking molecular dynamics simulation as examples, each nanosecond simulated conformational trajectories resulted in 2 GB data averagely; it would generate millions of conformations ($\sim$ 2 TB data) for micro-second time scale.

These big data bring in following challenges:

1. **Data storage.** Petabyte (10^{15} bytes) of digital information relies on cloud storage; chemical and biological data annotation/curation and quality assurance are challenging.
2. **Visualization.** Small molecules or biopolymers are described in graphs *per se*. These objects are usually converted into numbers (descriptors). Thus, a molecule is defined as a point in multi-dimensional space that requires dimension reduction approaches (such as principal component analysis (PCA), and nonlinear dimensionality reduction techniques), metadata generation techniques.
3. **Data mining.** Based on the high dimensional data, scientists are facing classification problems. Molecules are classified into two or more clusters corresponding to their phenotypes. Moreover, people need to understand the relations between the key factors/features/chemotypes and a specific phenotype(s). The real challenges are (a) the relations between the features and phenotypes are not of classical analytic function relations; (b) the features for an entire molecule are not related to its phenotypic property in the most of situation; (c) the local feature(s)/substructure(s) for an molecule can be the key to a phenotypic property, but there are uncountable ways to partition a molecular structure into substructures. That is why so may data mining tools have been developed (such as clustering algorithms, decision trees, supporting vector machines (SVM), artificial neural networks (ANN).
4. **Computational complexity.** The most precise theory to study a molecular system is quantum chemistry. However, the computational complexity of different quantum chemistry algorithms is so difficult that even a quantum computer will be unable to solve [38]. When we deal with a huge number of molecules interacting a protein, the situations are worse. To identify drug targets for a drug lead, multisequence alignment techniques are required. The computational complexity of sequence alignment algorithms ranges from $O(m * n)$ to $O(n^2)$ [5]. To identify privileged substructures responsible for a biological activity, sub-

structure match algorithms are applied. The computing complexity of these algorithms are usually polynomial [39].

Traditional drug discovery did not generate big data, however, modern instrumentation and automation changed the situations. With micro-chip technology, people can collect *in vivo* data from model animals 24 hours a day to monitor a drug action *in sito*. New drug discovery technologies such as Surface Plasmon Resonance (SPR, measuring protein-ligand affinity with optics) [18], Isothermal titration calorimetry (ITC, measuring protein-ligand affinity with entropy and enthalpy) [29], and Saturation Transfer Difference NMR spectroscopy (STD-NMR, measuring protein-ligand affinity with nuclear magnetic resonance) [27] allow us to acquire unprecedent protein-ligand interaction data. Omicses (such as Pharmaco-proteomics, Pharmaco-metabonomics), High performance computing, and cloud technologies are indispensable components for modern drug discovery service platform (Fig. 1).

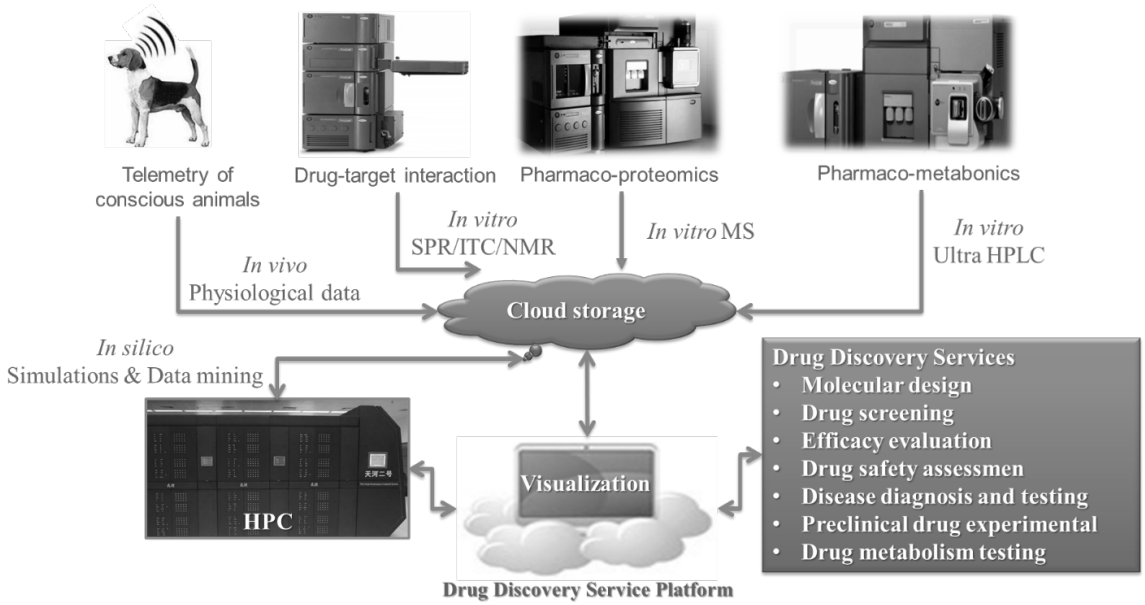

Figure 1. Drug discovery service platform with HPC and cloud storage

There are millions of available compounds from vendors and billions of virtual compounds for virtual screening. For a structure-based virtual screening approach (such as molecular docking), billions of conformations have to be enumerated for millions molecular structures requiring petabytes (PB) storage. If the docking results were validated by 50 ns MD simulations, each ligand-protein complex would need $\sim$ 25 GB space.

To constantly monitor the pharmaceutical efficacy of a compound *in vivo*, 2–3 months continuous administration will result in 3.5 PB physiological and pharmacological data, from which the efficacy, dosage, and toxicity can be determined.

Constantly monitoring cell changes (such as the effect of drugs on cell activity, the drug distribution, the alter cell behavior, proliferation or apoptosis) while cells incubated with a compound will result in 1 PB data for tracking 10 traits in 1000 cells for 24-hours.

Drug discovery processes involve in the data derived from patients to various devices in many different formats; these data require different search engines and approaches to retrieve and elucidate; and eventually result in personal diagnosis and treatment scenarios (Fig. 2).

Intrinsically, modern drug discovery is to discover macrocosmic solutions by simulation microcosmic phenomena with many experimental data. Therefore, this is a multi-scales simulation

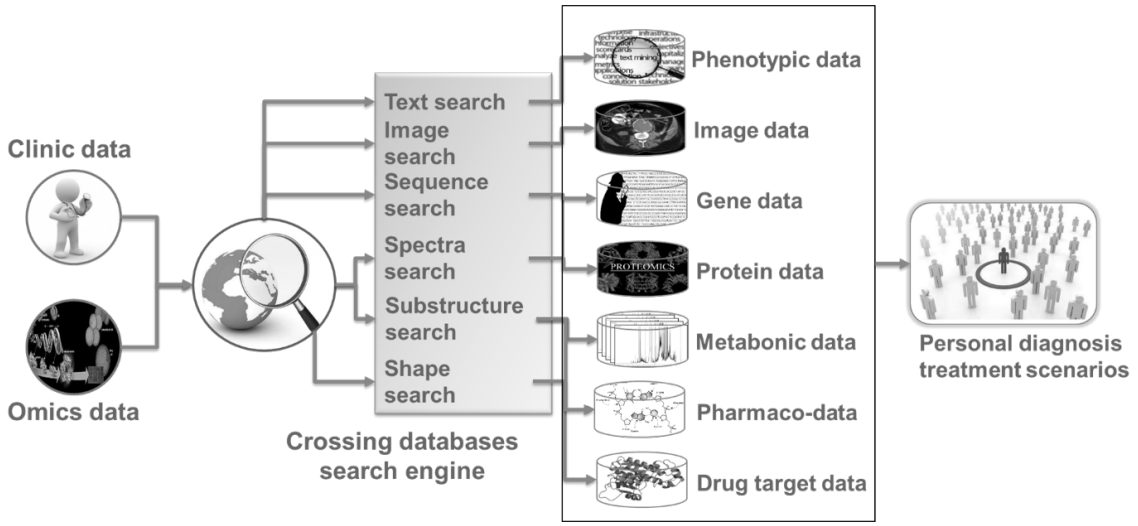

Figure 2. Data, search engine and mining tools involved in drug discovery process

process, which covers time scale (from femto-seconds to hours/days), space scale (from nanometers to meters) at changing resolutions (from electron orbitals to molecular machines) and various theories/methods (from density function theory to biopolymer physics) [11].

Therefore, the computing complexity in drug discovery is due to the complexity of the molecular systems. In many cases, the computing complexity issue can be reduced by parallel computing technology (*aka* high-performance computing) if the problem is parallelizable. For example, employing molecular dynamics-based virtual screening (MDBV), a state of the art HPC can be 600 times faster than an eight-core PC server is in screening a typical drug target (which contains about 40 K atoms). Also, careful design of the GPU/CPU architecture can reduce the HPC costs [15].

A successful virtual drug screening campaign relies on a properly selected compound library. Brutal random virtual screening can lead failure even one has the highest performance computing facility. Therefore, we desperately develop artificial intelligence (AI) applications in pharmaceutical studies.

2. Artificial Intelligence and Drug Discovery

The essence of drug discovery is to identify a molecule that interacts its designated biological target from a compound library that have millions of molecules. To do this, we have to understand the relation of molecular structure and activity (SAR). Here, the structure in SAR is actually substructure. A drug molecule can be considered as a molecular machine consists of various functional parts (also termed as substructures, fragments, or chemotypes). How to define the functional parts has been puzzling for many years. Many methods, such as empiric-based method [13], and computational rule based methods [7, 12, 28, 40] were proposed. There is no perfect way to partition substructures from a compound library. Therefore, People also explored other methods, such as molecular descriptors [30], atomic pairs [8], and fingerprints [6].

Conventionally, in order to predict whether a species (for example, a natural substance) has a biological activity, scientists have to extract moieties from the substance, determine chemical structures (represented in topologies, 3D shapes or static surfaces) of the active ingredients; then to covert the chemical structures into a numeric array (called as molecular descriptors or fingerprints). Then, various mathematic models are applied on the data to generate predictive

models. Finally, the models result in the prognosis whether the substance is a candidate to become a drug (Fig. 3).

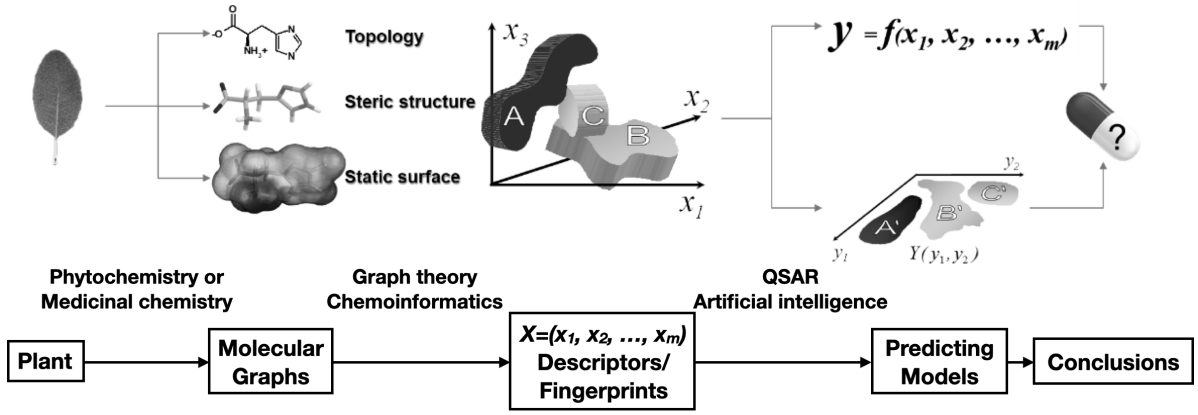

Figure 3. Flow-chart for conventional structure-activity predictions

Molecular structure representations can be converted into various molecular descriptors such as sub-structural fragments, scaffolds, atom pairs (paths), topologic indexes, physical/biological/chemical properties, and fingerprints (bit-maps). The combinations of the descriptors will be figured out based on two principles: (1) a descriptor in the combination has to be significantly associated with the property to be predicted; (2) descriptors within the combination should be orthogonal to each other. Based upon the descriptor combination data, one can build predictive models with learning methods as shown in Fig. 4.

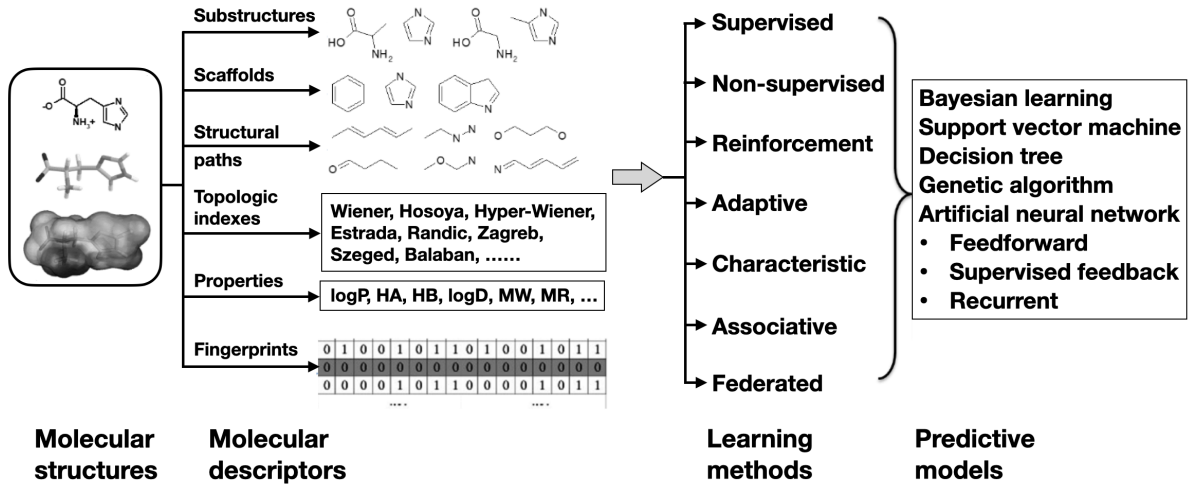

Figure 4. Conventional flow-chart for applying AI methods in predicting pharmaceutical properties for molecules in drug discovery process

The cores of AI are pattern recognitions that are divided into numeric and non-numeric pattern recognitions. Markush structure or substructure recognitions are non-numerical; self-organizing map (SOM) (*aka* Kohonen network), support vector machine, hierarchical cluster tree, or random forests (*aka* random decision forests) are numerical. The common defect of the conventional machine learning algorithms is that the model performance highly relies on how a modeler selects and combines the molecular descriptors. Unfortunately, there is not rational rules to choose and combine molecular descriptors. In order to make up for this defect, people tried

many approaches, such as rule-embedded naive Bayesian learning [24], multiple machine learning models [23], and combining recursive partitioning with Nave Bayesian learning approaches [35].

Now, people realize that deriving substructures that related to activities from a molecule or molecular library depends on related drug target. In the earlier time of chemoinformatics, a number of molecular structure linear notions were developed due to the lack of computer graphic terminals in that period. Weininger developed the linear notations system called as SMILES (simplified molecular-input line-entry system) that are well accepted internationally [37]. SMILES is an accurate language for molecules, a SMILES notation/sentence precisely describes the atomic connectivity in a molecule. Thus, a compound library can be "written" as an article composed in SMILES sentences. A focused compound library for a specific biological target can be viewed as an article written in SMILES sentences under the same title.

This concept is important because we can derive substructures and activities relations (SAR) without predefining substructures. With deep learning approaches, we can figure out the SAR or predict drug targets with syntax pattern recognition techniques [25].

As shown in Fig. 5, a chemical structure is converted into a SMILES sentence, which is then transformed to a reduced vocabulary, eventually a word embedding matrix is calculated and finally sent to recurrent neural network (RNN) to train a learning model.

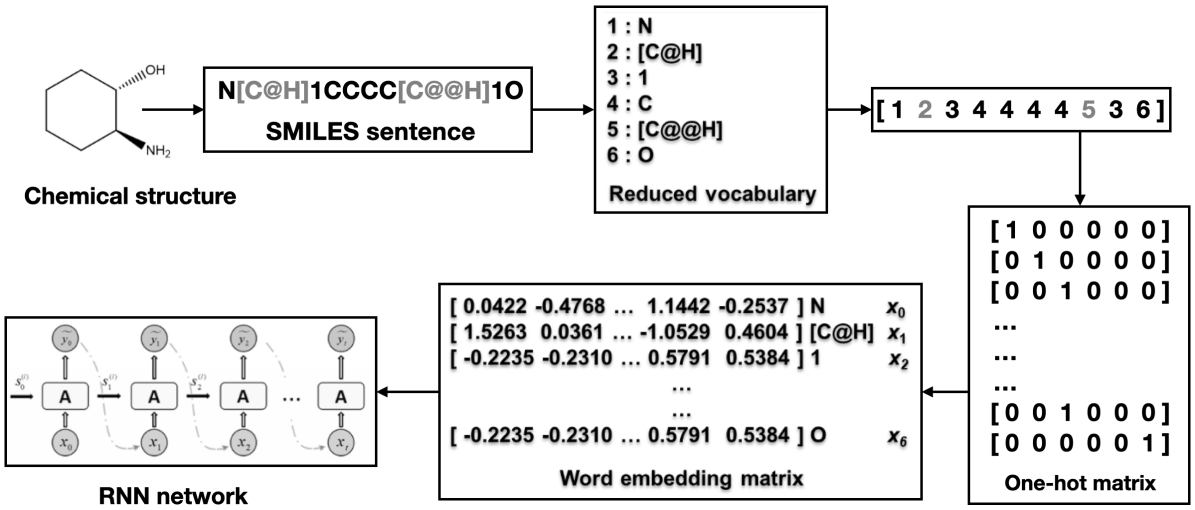

Figure 5. Training deep learning models using SMILES without predefining substructures

With self-attention mechanism, structure-activity/property relations (SAR/SPR) can be discovered through chemical linear notation (for example, SMILES) syntax analyses using an interpretable deep learning architecture. The syntax pattern recognition approach has been applied in predicting chemical properties, toxicology, and bioactivity from experimental data sets [2, 3, 9, 10, 17, 36, 44].

With the syntax pattern recognition protocol, drug-like, lead-like, or quasi-biogenic molecules can be proposed by a deep learning program. A quasi-biogenic molecule generator (QBMG) to compose virtual quasi-biogenic compound libraries by means of gated recurrent unit recurrent neural networks has been reported. The library includes stereo-chemical properties, which are crucial features of natural products. QMBG can reproduce the property distribution of the underlying training set, while being able to generate realistic, novel molecules outside of the training set. The proposed compounds were associated with known bioactivities. Therefore,

with a given focused compound library for a biological target, a computer can generate novel compounds that are promising to be active against the target [43].

A property of a molecule can associate with one or more substructures in its structure. For chemical structure stability prediction, if one substructure is found responsible for the instability, it will be enough to conclude the molecule is instable. A model (DeepChemStable) [22] employing an attention-based graph convolution network based on the COMDECOM data (experimental chemical compound instability data set [45]) was implemented to predict a compound instability. The main advantage of this method is that is an end-to-end model, which does not predefine structural fingerprint features, but instead, dynamically learns structural features and associates the features through the learning process of an attention-based graph convolution network. The previous ChemStable program (with conventional machine learning approach) [24] relied on a rule-based method to reduce the false negatives. DeepChemStable, on the other hand, reduces the risk of false negatives without using a rule-based method minimizing the rate of false negatives, which is a greater concern for instability prediction (Fig. 6).

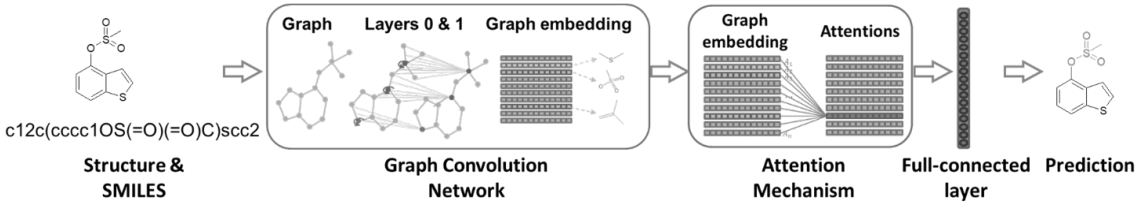

Figure 6. Flow-chart of an attention-based graph convolution network that predict a compound chemical instability

Fragment-based drug design (FBDD) [16] gains great achievements these years. Linking fragments to generate a focused compound library for a specific drug target is still puzzeling. A program named SyntaLinker that is based on a syntactic pattern recognition with deep conditional transformer neural networks was reported recently. The state-of-the-art transformer links molecular fragments automatically by learning from known structures in medicinal chemistry databases (such as ChEMBL). Linking the fragments was viewed as connecting substructures that were predefined by empirical rules in the past. In SyntaLinker, however, the rules of linking fragments can be learned implicitly from the known chemical structures by recognizing syntactic patterns embedded in SMILES notations. With deep conditional transformer neural networks, SyntaLinker can generate molecular structures based on a given pair of fragments and additional restrictions [41].

Syntactic pattern has also been applied in predicting chemical reaction feasibility. Copper(I)-catalyzed alkyneazide cycloaddition (CuAAC) reaction is a main click chemistry reaction [20] and widely employed in drug discovery. However, the success rate of the CuAAC reaction is not satisfactory as expected. A recurrent neural network (RNN) model was reported to predict its feasibility. Authors designed and synthesized a structurally diverse library of 700 compounds with the CuAAC reaction to obtain experimental data. Then, a bidirectional longshort-term memory with a self-attention mechanism (BiLSTM-SA) model was built. The model achieved total accuracy of 80%. Density functional theory investigations were conducted to provide evidence for the correlation between bromo-α-C hybrid types and the success rate of the reaction [32].

3. Perspectives on Computational Opportunities and Challenges in Drug Discovery

The Nobel Prize in Chemistry 2016 was awarded jointly to Jean-Pierre Sauvage, Sir J. Fraser Stoddart and Bernard L. Feringa for the design and synthesis of molecular machines [33]. This can be viewed as an overture for artificial molecular machine era. So far, chemists focus on the mechanical aspects artificial molecular machines [1, 42]. Actually, a drug molecule can also be viewed as an artificial molecular machine that consists of a number of parts (fragments) for regulating biological targets. Thus, the essential questions for drug design methodology becomes (1) what are the fragments for a drug molecule for its target? (2) how to assembly the fragments to make (synthesize) a drug molecule? (3) how to biologically validate the assembled molecules. FBDD, click chemistry (combinatorial chemistry), and HTS are the current answers to these questions respectively. Drug discovery process is similar a machine invention process (Fig. 7).

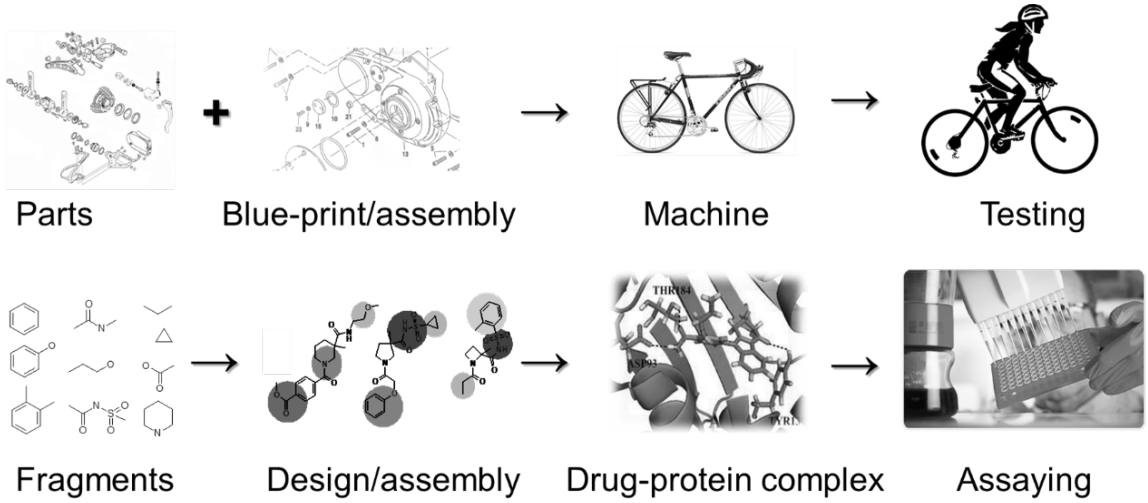

Figure 7. Comparison of processes of a machine discovery and a drug discovery

Drug discovery is much more sophisticated than design and make a machine in macrocosm due to a drug molecule has to regulate even more complicated biological machines in microcosm [14, 21]. The main challenges to a drug designer are: (1) the designed plan for assembling fragments is not necessarily chemically feasible; and (2) the designed molecules against a target is not necessarily functioned as expected. Because most of the mechanisms of actions in life are not well understood to us. Therefore, new drug design approaches and *in silico* experiments are demanding to deal with the big data and computing complexity problems.

Artificial intelligence (AI) techniques will continue to demonstrate their power in drug discovery. Especially, deep learning (DL) techniques have shown the usefulness in deriving SAR from big biochemical data. However, DL assumes the positives and negatives are evenly distributed in a training set, and the number of the samples is big enough. However, typical medicinal chemistry data mainly contain positives with no or minor negatives.

Drug discovery involves multi-scale computation issues. For example, the length of a typical human DNA molecule is about 1.8 meters (visible in macrocosm) has to be tightly packed up to fit in the micro-meter-scale space of cell nucleus (in microcosm). We dont have a convincing theory to explain how a DNA enters microcosm world from macrocosm world with the help of histone proteins. It is a grand computational challenge to generate a model and simulate this

process. Interestingly, recent report claimed that the histones are not just used for packing DNA, they are enzymes that may have helped power eukaryote evolution [4].

Acknowledgements

The author would like to thank the National Key R&D Program of China (2017YFB0203403), the National Natural Science Foundation of China (81870608), the science and technology program of Guangzhou (201604020109), the Guangdong Provincial Key Lab. of New Drug Design and Evaluation (Grant 2011A060901014) for funding.

References

1. Aprahamian, I.: The future of molecular machines. ACS Central Science 6(3), 347–358 (2020), DOI: 10.1021/acscentsci.0c00064

2. Arús-Pous, J., Johansson, S.V., Prykhodko, O., et al.: Randomized smiles strings improve the quality of molecular generative models. Journal of Cheminformatics 11(1), 71 (2019), DOI: 10.1186/s13321-019-0393-0

3. Arús-Pous, J., Patronov, A., Bjerrum, E.J., et al.: SMILES-based deep generative scaffold decorator for de-novo drug design. Journal of Cheminformatics 12(1), 38 (2020), DOI: 10.1186/s13321-020-00441-8

4. Attar, N., Campos, O.A., Vogelauer, M., et al.: The histone H3-H4 tetramer is a copper reductase enzyme. Science 369(6499), 59–64 (2020), DOI: 10.1126/science.aba8740

5. Baichoo, S., Ouzounis, C.A.: Computational complexity of algorithms for sequence comparison, short-read assembly and genome alignment. Bio Systems 156-157, 72–85 (2017), DOI: 10.1016/j.biosystems.2017.03.003

6. Banegas-Luna, A.J., Cern-Carrasco, J.P., Pérez-Sánchez, H.: A review of ligand-based virtual screening web tools and screening algorithms in large molecular databases in the age of big data. Future Medicinal Chemistry 10(22), 2641–2658 (2018), DOI: 10.4155/fmc-2018-0076

7. Bemis, G.W., Murcko, M.A.: The properties of known drugs. 1. Molecular frameworks. Journal of Medicinal Chemistry 39(15), 2887–2893 (1996), DOI: 10.1021/jm9602928

8. Carhart, R.E., Smith, D.H., Venkataraghavan, R.: Atom pairs as molecular features in structure-activity studies: definition and applications. Journal of Chemical Information and Computer Sciences 25(2), 64–73 (1985), DOI: 10.1021/ci00046a002

9. Chen, H., Engkvist, O., Wang, Y., et al.: The rise of deep learning in drug discovery. Drug Discovery Today 23(6), 1241–1250 (2018), DOI: 10.1016/j.drudis.2018.01.039

10. Chen, J., Cheong, H.H., Siu, S.W.I.: Bestox: A convolutional neural network regression model based on binary-encoded SMILES for acute oral toxicity prediction of chemical

compounds. In: Martín-Vide, C., Vega-Rodríguez, M.A., Wheeler, T. (eds.) Algorithms for Computational Biology. pp. 155–166. Springer International Publishing, Cham (2020), DOI: 10.1007/978-3-030-42266-0_12

11. Dans, P.D., Walther, J., Gómez, H., Orozco, M.: Multiscale simulation of DNA. Current Opinion in Structural Biology 37, 29–45 (2016), DOI: 10.1016/j.sbi.2015.11.011

12. Dehaspe, L., Toivonen, H., King, R.D.: Finding frequent substructures in chemical compounds. In: Agrawal, R., Stolorz, P.E., Piatetsky-Shapiro, G. (eds.) Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining, KDD-98, 27-31 August 1998, New York City, New York, USA. pp. 30–36. AAAI Press (1998)

13. Durant, J.L., Leland, B.A., Henry, D.R., Nourse, J.G.: Reoptimization of MDL keys for use in drug discovery. Journal of Chemical Information and Computer Sciences 42(6), 1273–1280 (2002), DOI: 10.1021/ci010132r

14. García-López, V., Chen, F., Nilewski, L.G., et al.: Molecular machines open cell membranes. Nature 548(7669), 567–572 (2017), DOI: 10.1038/nature23657

15. Ge, H., Wang, Y., Li, C., et al.: Molecular dynamics-based virtual screening: Accelerating the drug discovery process by high-performance computing. Journal of Chemical Information and Modeling 53(10), 2757–2764 (2013), DOI: 10.1021/ci400391s

16. Hajduk, P.J., Greer, J.: A decade of fragment-based drug design: strategic advances and lessons learned. Nature Reviews Drug Discovery 6(3), 211–219 (2007), DOI: 10.1038/nrd2220

17. Hirohara, M., Saito, Y., Koda, Y., et al.: Convolutional neural network based on SMILES representation of compounds for detecting chemical motif. BMC Bioinformatics 19(19), 526 (2018), DOI: 10.1186/s12859-018-2523-5

18. Homola, J.: Surface plasmon resonance sensors for detection of chemical and biological species. Chemical Reviews 108(2), 462–493 (2008), DOI: 10.1021/cr068107d

19. Itoh, H., Tokumoto, K., Kaji, T., et al.: Development of a high-throughput strategy for discovery of potent analogues of antibiotic lysocin E. Nature Communications 10(1), 2992 (2019), DOI: 10.1038/s41467-019-10754-4

20. Kolb, H.C., Finn, M.G., Sharpless, K.B.: Click chemistry: Diverse chemical function from a few good reactions. Angewandte Chemie International Edition 40(11), 2004–2021 (2001), DOI: 10.1002/1521-3773(20010601)40:11¡2004::AID-ANIE2004¿3.0.CO;2-5

21. Lancia, F., Ryabchun, A., Katsonis, N.: Life-like motion driven by artificial molecular machines. Nature Reviews Chemistry 3(9), 536–551 (2019), DOI: 10.1038/s41570-019-0122-2

22. Li, X., Yan, X., Gu, Q., et al.: DeepChemStable: Chemical stability prediction with an attention-based graph convolution network. Journal of Chemical Information and Modeling 59(3), 1044–1049 (2019), DOI: 10.1021/acs.jcim.8b00672

23. Li, Y., Wang, L., Liu, Z., et al.: Predicting selective liver X receptor beta agonists using multiple machine learning methods. Mol Biosyst 11(5), 1241–1250 (2015), DOI: 10.1039/c4mb00718b

24. Liu, Z., Zheng, M., Yan, X., et al.: ChemStable: a web server for rule-embedded naïve Bayesian learning approach to predict compound stability. Journal of Computer-Aided Molecular Design 28(9), 941–950 (2014), DOI: 10.1007/s10822-014-9778-3

25. Mayr, A., Klambauer, G., Unterthiner, T., et al.: Large-scale comparison of machine learning methods for drug target prediction on ChEMBL. Chem. Sci. 9(24), 5441–5451 (2018), DOI: 10.1039/C8SC00148K

26. Merrifield, R.B.: Solid phase synthesis (Nobel lecture). Angewandte Chemie International Edition in English 24(10), 799–810 (1985), DOI: 10.1002/anie.198507993

27. Meyer, B., Peters, T.: NMR spectroscopy techniques for screening and identifying ligand binding to protein receptors. Angewandte Chemie International Edition 42(8), 864–890 (2003), DOI: 10.1002/anie.200390233

28. Peng, H., Liu, Z., Yan, X., et al.: A de novo substructure generation algorithm for identifying the privileged chemical fragments of liver X receptorbeta agonists. Scientific Reports 7(1), 11121 (2017), DOI: 10.1038/s41598-017-08848-4

29. Rajarathnam, K., Rösgen, J.: Isothermal titration calorimetry of membrane proteins – progress and challenges. Biochimica et Biophysica Acta (BBA) – Biomembranes 1838(1, Part A), 69–77 (2014), DOI: 10.1016/j.bbamem.2013.05.023

30. Sahoo, S., Adhikari, C., Kuanar, M., Mishra, B.K.: A short review of the generation of molecular descriptors and their applications in quantitative structure property/activity relationships. Current Computer-Aided Drug Design 2(3), 181–205 (2016), DOI: 10.2174/1573409912666160525112114

31. Saiki, R.K., Bugawan, T.L., Horn, G.T., et al.: Analysis of enzymatically amplified beta-globin and HLA-DQ alpha DNA with allele-specific oligonucleotide probes. Nature 324(6093), 163–166 (1986), DOI: 10.1038/324163a0

32. Su, S., Yang, Y., Gan, H., et al.: Predicting the feasibility of copper(i)-catalyzed alkyne–azide cycloaddition reactions using a recurrent neural network with a self-attention mechanism. Journal of Chemical Information and Modeling 60(3), 1165–1174 (2020), DOI: 10.1021/acs.jcim.9b00929

33. Van Noorden, R., Castelvecchi, D.: World's tiniest machines win chemistry Nobel. Nature 538(7624), 152–153 (2016), DOI: 10.1038/nature.2016.20734

34. Walters, W.P.: Virtual chemical libraries. Journal of Medicinal Chemistry 62(3), 1116–1124 (2019), DOI: 10.1021/acs.jmedchem.8b01048

35. Wang, L., Chen, L., Liu, Z., et al.: Predicting mTOR inhibitors with a classifier using recursive partitioning and naïve Bayesian approaches. PLOS ONE 9(5), 1–15 (2014), DOI: 10.1371/journal.pone.0095221

36. Wang, S., Guo, Y., Wang, Y., et al.: SMILES-BERT: Large scale unsupervised pretraining for molecular property prediction. In: Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics. pp. 429–436. BCB '19, Association for Computing Machinery, New York, NY, USA (2019), DOI: 10.1145/3307339.3342186

37. Weininger, D.: SMILES, a chemical language and information system. 1. introduction to methodology and encoding rules. Journal of Chemical Information and Computer Sciences 28(1), 31–36 (1988), DOI: 10.1021/ci00057a005

38. Whitfield, J.D., Love, P.J., Aspuru-Guzik, A.: Computational complexity in electronic structure. Phys. Chem. Chem. Phys. 15(2), 397–411 (2013), DOI: 10.1039/C2CP42695A

39. Xu, J.: GMA: A generic match algorithm for structural homomorphism, isomorphism, and maximal common substructure match and its applications. Journal of Chemical Information and Computer Sciences 36(1), 25–34 (1996), DOI: 10.1021/ci950061u

40. Xu, J.: A new approach to finding natural chemical structure classes. Journal of Medicinal Chemistry 45(24), 5311–5320 (2002), DOI: 10.1021/jm010520k

41. Yang, Y., Zheng, S., Su, S., et al.: Syntalinker: automatic fragment linking with deep conditional transformer neural networks. Chem. Sci. 11(31), 8312–8322 (2020), DOI: 10.1039/D0SC03126G

42. Zhang, L., Marcos, V., Leigh, D.A.: Molecular machines with bio-inspired mechanisms. Proceedings of the National Academy of Sciences 115(38), 9397–9404 (2018), DOI: 10.1073/pnas.1712788115

43. Zheng, S., Yan, X., Gu, Q., et al.: QBMG: quasi-biogenic molecule generator with deep recurrent neural network. Journal of Cheminformatics 11(1), 5 (2019), DOI: 10.1186/s13321-019-0328-9

44. Zheng, S., Yan, X., Yang, Y., Xu, J.: Identifying structure–property relationships through SMILES syntax analysis with self-attention mechanism. Journal of Chemical Information and Modeling 59(2), 914–923 (2019), DOI: 10.1021/acs.jcim.8b00803

45. Zitha-Bovens, E., Maas, P., Wife, D., et al.: Comdecom: predicting the lifetime of screening compounds in DMSO solution. J Biomol Screen 14(5), 557–565 (2009), DOI: 10.1177/1087057109336953