The New Answer to Drug Discovery: Quantum Machine Learning in Preclinical Drug Development

Yew Kee Wong, Yifan Zhou, Yan Shing Liang, Hai Chuan Qiu, Yu Xi Wu, Bin He

BASIS International School Guangzhou, Guangzhou 510653, China

Abstract. The Research & Development (R&D) phase of drug development is a lengthy and costly process, usually spanning from six to nine years [1] and costing four hundred to fourteen hundred million USD [2]. To revolutionize this process, we introduce our new concept-the combination of Quantum-based Machine Learning network (QML) and Quantum Computing Simulation (QS)-to shorten the whole R&D phase to three to six months and decrease the cost to merely fifty to eighty thousand USD. Our program takes the inputs of the target protein/gene structure and the primary essay [3]. For Hit Generation [3], the QML network generates possible hits [4] according to the molecular structure of the target protein while the QS filters molecules from the primary essay based on the reaction and binding effectiveness with the target protein. Then, For Lead Optimization [3], the resultant molecules generated and filtered from QML and QS are compared, and the ones that appear as a result of both processes will be made into dozens of molecular variations, while others will only undergo simple modifications. Lastly, all optimized molecules would undergo multiple rounds of QS filtering with a high standard for reaction effectiveness and safety, creating a few dozen pre-clinical-trail-ready drugs. Our concept of the combination of QML and QS can also prove revolutionary in many other fields, such as agriculture research, genetic editing, and even aerospace engineering.

Keywords: Quantum Computer, Hit to Lead, Lead Optimization Simulation, Machine Learning.

1 Introduction

1.1 Current Drug Development Issues/ Traditional Industry

The costs of traditional drug development methods are increasing and are accompanied by extremely low productivity. Many chemical molecules are eliminated at the development stage and those that remain fail in clinical trials, with only a tiny fraction of them eventually being used in drug development. [5]

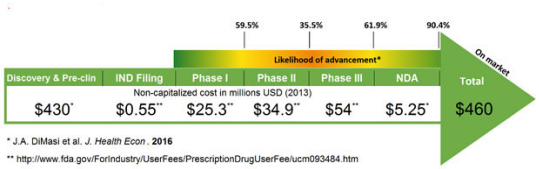

Figure 1. Cost of drug development by stages: Seqens CDMO-APIs [6]

Each new conventional drug takes 10 to 15 years to develop, while costing roughly US$4.6 billion, with discovery and pre-clinical development taking up 93.4% of the total cost [6]. Traditionally, the process of drug discovery has been through the identification of ingredients that are effective in traditional therapies or through accidental discovery. Furthermore, the drugs are screened for results and optimized to increase stability and affinity. Once the drug has been identified as effective, it can then be tested in clinical trials. The traditional way of developing drugs is therefore considered to be expensive and inefficient [7].

1.2 The use of Machine Learning Applied with Quantum Computing

Quantum computing is 158 million times faster compared to traditional computing, which can greatly increase the speed of machine learning [8]. The advantages of quantum computing are that its algorithms are the fastest known and the complexity of the computation is relatively low.

Drug development through quantum computing can be compressed to a cost of around $10,000 to $50,000 and reduced to a few weeks [9]. Specifically, quantum computing can greatly improve the accuracy and speed of data when processing molecule-to-molecule interactions. For initial experimental data, quantum computing will also filter in a way that is superior to traditional computers. Unlike conventional computers, quantum computing can provide superior molecular hits in the shortest possible time. This provides the pharmaceutical industry with a significant reduction in cost and time to develop drugs.

1.3 Approach Overview

Our QML will be based on a self-supervise ML training system that takes the inputs of the target protein/gene structure and the primary essay and generates possible hits according to the molecular structure of the target protein, making up 50% of the Hit Generation pool. The other 50% would be generated by using our QS that filters molecules from the primary essay based on the reaction and binding effectiveness with the target protein. Next, For Lead Optimization, the resultant molecules generated and filtered from QML and QS are compared, and the ones that appear as a result of both processes will be made into dozens of molecular variations, while others will only undergo simple modifications. Lastly, all optimized molecules would undergo multiple rounds of QS filtering with a high standard for reaction effectiveness and safety, creating a few dozen pre-clinical-trail-ready drugs.

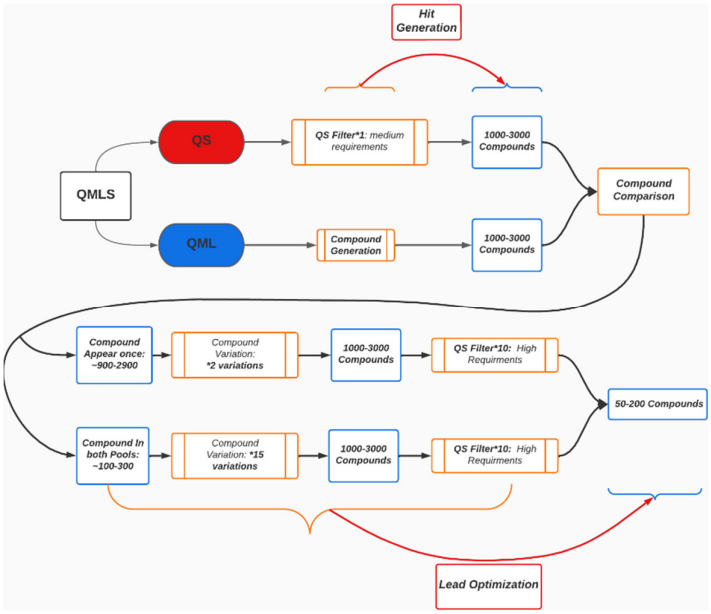

Figure 2. Chart showing the whole QML & QS System

2 Quantum Computing Assisted Machine Learning Design and Predicted Results

2.1 Molecular Generation & Pre-QML Processing

Our Molecular Generation Algorithm is based on the relationship between the protein structure and how it will interact with other molecules [10]. Specifically, the amino acid chain sequences, α&β folding patterns, R-Group characteristics, and bonding 3D shape determines how the protein will interact with other proteins [11]. Using this established relationship, we will generate amino acid sequence, bonding patterns, and 3D shape based on the inputs of desired interaction of protein with the target molecule. But why QML? For two reasons; first, current mathematical models of protein interactions using even the best supercomputers can take days and can only detect 70% of protein interactions [12] but with Machine Learning's large training database and Quantum Computer's parallel processing we hope to greatly improve the accuracy and shorten the time to hours even minutes. Secondly, the ML offers the ability to generate completely new protein structures, bringing previously undiscovered possible drug solutions to reality.

Before the activating QML, our program will filter out amino acid bases, parings, and chains that would defiantly not produce the desired inaction with the target protein, based on their functional group composition [13] to decrease the number of possible

sequences, producing higher quality hits in less time. To put into perspective, for 20 amino acid bases, there are 20^{100} possible sequences if the length is just 100 (Average lengths being between 283 and 340 [14]). Moreover, our program will also lay the foundation for QML protein generation by providing the QML with 2-5 sets of amino acid chain consisting of 5-10 amino acids [15] that has the potential of producing hits to start with as the basis for the QML to build on, to further decrease the possible sequences.

2.2 Machine Learning Algorithm Design

Compound Generation. First, corresponding to the Hit Generation, the QML will intake the target protein structure, in the form of a matrix converted from the Protein Data Bank (PBD) [16], and amino acid sets base generated prior. Then, the QML will start the process of compound generation, generating 500-3000 compounds, which is about half the average number of compounds of the secondary assay [17], or the amount prior to Lead Optimization.

Compound Variation Generation. After the hits generated and filtered from the QML and QS are separated into compounds that are resultant of both the QML and QS and compounds that are not repeated. Then, to perform Lead Optimization, the QML will make 15 variations for each repeated compounds and 3 for non-repeated ones as repeated compounds are more likely to be drug candidates. QML performs variation generation by adding/deleting amino groups, altering bonds, and changing folding sequences. This process is intended to create safer and more effective compounds by making more variations of the compound and is based on the framework of the popular Lead Optimization strategy of utilizing in vitro and in vivo [18].

Finally, there will be approximately 3000-12,000 compounds generated in total that will proceed to the next stage where they will be filtered again by the QS until only 50-200 pre-clinical-ready drugs are left.

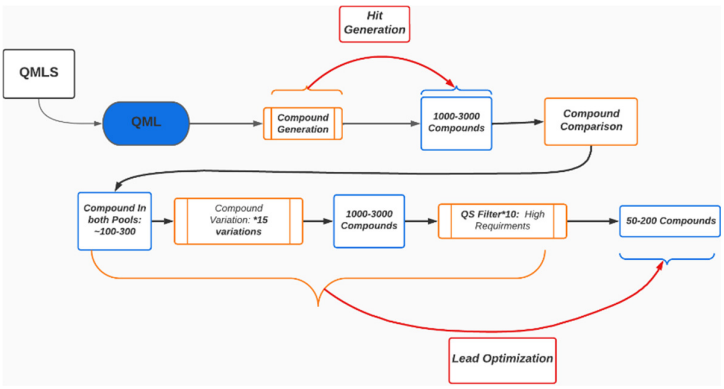

Figure 3. Chart showing the QML system

2.3 Machine Learning Algorithm Training

Since our QML not only need to generate possible hits but also make these hits constructable in the real world and safe for human consumption, simple reinforcement learning or supervised learning would not create an QML as effective as we need. As a result, we deiced to implement Self Supervised Pre-training, which introduces a progression, from simpler to harder, of related tasks to the ML, then optimize the QML to fit our initial requirements. Self-Supervised Pre-training in many fields such as visual analysis and language processing has already proved more effective than conventional ML methods [19][20], especially with complex, simulative MLs. Moreover, Pre-Training produces better results for smaller datasets [21], which might be the case for some rare amino acid groupings.

Specifically, we plan to train our QML to first perform 3-4 simple tasks such as taking away a functional group or bond from an amino acid then asking the QML to predict the missing piece, to gasp the natural structure and grouping of functional groups. Then, we will implement 2-3 more tasks to train our QML to predict the reaction/bonding patterns and reaction hazardless. Finally, we will optimize our QML to perform the functions of protein generation and protein variation generation.

For the implementation of our QML, we plan to use Microsoft's deepspeed.ai software for its memory efficiency, support of long input sequence length, and speed [22].

3 Drug Development Simulation Software Design

3.1 Overview, Goals, and Process of Simulation

After Creating a list of (numerical #) potential protein options for producing the final medical, there needs to be a process to prune and filter the list based on the targets the developing team needs. These targets can include functions such as blocking off specific body receptors [23], catalyzing particular chemical reactions, or creating body-needed elementary bio-molecules as an enzyme [24]. During the filtering process, it is also essential to note the possible unknown side-effects that could be potentially malicious, for example, Osteoporosis, diabetes, or dehydration, as detected in protein diet consumers that did not ingest the correct number of proteins [25]. To avoid causing severe harm and to ensure the effectiveness of the drug developed, the filtering process thus needs to achieve the goals of simulating and predicting all the protein's functions and reactions and ridding off all undesired proteins with inadequate structures, R groups, or other attributes. Usually, such a filtering process done by humans subjectively is time-consuming [26], but as technologies of Quantum computation evolved into the fields of drug creation, drug simulation time and cost have increased lowered [27]. A similar idea can be used in this filtering process of drugs to enhance efficiency greatly.

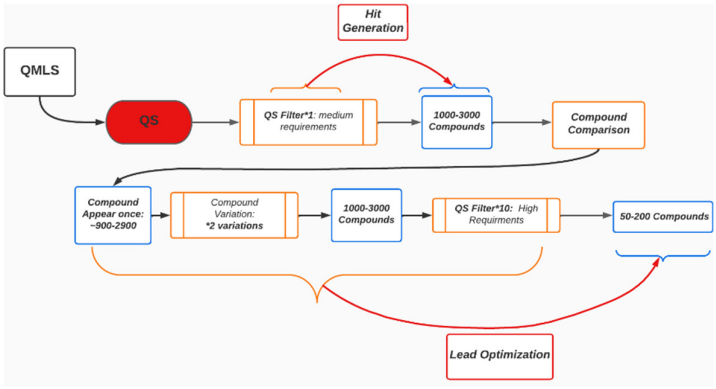

Figure 4. Chart showing the QS system

3.2 Methods of Filtration

As discussed, the amino acid chain sequences, α&β folding patterns, R-Group characteristics, and bonding 3D shape plays determining roles in how the proteins will interact with their environment, so the simulation software should design similarly to an algorithm. This algorithm-like software can be fed in with a list of PDB (protein data bank) files containing complete information about a protein [28], then it will return a list of PDB files that meets the target criteria. The criteria are manually determined by humans, possibly including amino acid chain sequences, R-groups, folding patterns, and 3D bonding shapes.

3.2.1 Filtering by Amino acid chain sequences and R-groups. If the goal is to identify certain amino acid chains or R-groups within the PDB file, then this process can be achieved with a simple conversion tool: *PDB2fastA* tool [29]. This tool will yield a *fastA* file containing needed information on the sequence of amino acids and the protein's name, which can be analyzed.

Example of a *fastA* file [30]:

>7R98_1|Chains A, B, C|Nucleoprotein|Severe acute respiratory syndrome coronavirus 2 (2697049)

ATASWFTALTQHGKEDLKFPRGQGVPINTNSSPDDQIGYYRRATRRIRGGDGKMK DLSPRWYFYYLGTGPEAGLPYGANKDGIIWVATEGALNTPKDHIGTRNPANNAAIV LQLPQGTTLPKGFYAE

>7R98_2|Chains D, E, F|Nanobody B6|Lama glama (9844)

MAEVQLQASGGGLVQAGDSLRLSCVAVSGRTISTFAMGWFRQAPGKEREFVATIN WSGSSARYADPVEGRFTISRDDAKNTVYLEMSSLKPGDSAVYYCASGRYLGGITSYS QGDFAPWGQGTQVTVSSAAALEHHHHHH

This amino acid chain can also be analyzed manually and by a computer to determine the R-groups attached to the chain. As one amino-acid only corresponds to one distinct R-groups [31], a dictionary of corresponding amino-acid and R-groups can be constructed and used in the software.

A simple pseudo-code of this general implementation:

```
function test ( desired_R_groups, desired_Amino_Acids ):
# The PDB must be converted to fastA before executing
  r_Group_Dic = { amino_acid1 : R_group1, amino_acid2 : R_group2, ... }

Read generated_protein.fasta as GP
# Only the amino acid sequence should be left in Sequence after stringify function
  Sequence = GP.stringify
   if GP contains desired_Amino_Acids:
      for i in GP.length/3:
      r_group = r_Group_Dic (GP[i])

      if r_group contains desired_R_groups:
         Print ("This GP qualifies")
      else:
         Print ("This GP do not qualify")
```

This test takes in 1 *fastA* file and returns whether this particular protein sequence contained in this *fastA* file contains this list of R-groups or amino acid sequences inputted as parameters. This test can be executed n times, and every execution will contain *i/3* tests with n denoting the number of *fastA* files and *i* denoting the exact length of the amino acid sequences formatted by 3 letters representing 1 amino acid. A full test of a whole list of PBD generated will therefore have a complexity of *O(n*i)*, an acceptable range that a standard computer or supercomputer could do. Running this test first could narrow the lead optimization range in the most accurate way as this test is based only on factual comparisons, meaning there is no uncertainty with what might be outputted.

3.2.2 Filtering by structures of protein. The PDB file has already contained enough information to produce a good-looking protein model. The PDB file can be viewed in tools such as *Jmol* [32] or Swiss-PDB viewers [33]. These tools will calculate and provide interfaces and information such as animations, vibrations, surfaces, and orbitals of the structure for the viewer to further determine whether they want to retain or discard that protein. This step in the filtering process is just as paramount as the first step in filtering amino-acid and R-groups as it clears out a large proportion of unwanted noise in the data for the development team to find the desired protein eventually.

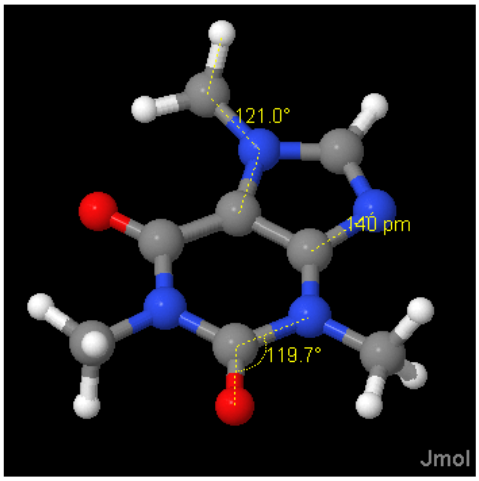

Figure 5. Example model on *Jmol's* official page: jmol.sourceforge.net/screenshots/measurements-1.png [32]

3.2.3 Filtering by predicting reaction of protein. The last of the filtration is predicting the interaction between protein and its environment. This is the hardest step, as there are numerous uncertainties in protein interactions and the only way to determine an interaction's result is to experiment with it in real life. However, it is not ideal for testing out every possible protein due to the high cost of laboratory preparations and time. To circumvent these obstacles, the developing team can utilize existing protein interaction databases or predict via a simulator. Some interactions might be found right off the database or be synonymous with other prevalent interactions in the database, so it would be helpful to first run filter search tests on platforms such as GPS-port [34] or Wiki-Pi [35]. These databases hold millions of verified interactions that can be used for identification and can better narrow down the range of search for the desired proteins.

Unfortunately, not all reactions are recorded in these bases, so the developing team would need to manually test out and predict the given protein's final suitability by finding a close neighbor reaction. This method can yield high accuracy in predicting the protein's function [36] and thus allows the developing team to entirely rule out unwanted proteins.

4 Combination of Simulation and Quantum Computer and Estimated Results

4.1 Using quantum computing for medicine overview

Doing drug development usually involved testing a variety of molecular compounds versus numerous diseases and see how the compounds react [37]. However, trying

these compounds in real life is usually very expensive, and developing a drug through testing hundreds, or even thousands of such compounds can lead to the cost of such a process to skyrocket, resulting in the drug being priced high on market for making up this development cost, and this can have a lot of bad effects. Thus, a computer simulation is usually used to simulate the ways molecules react. In this case, a quantum computer will be much more suitable to do the simulation than the classic transistor-based computer in doing a large quantity of little math operations that bond together to form a computer simulation [38].

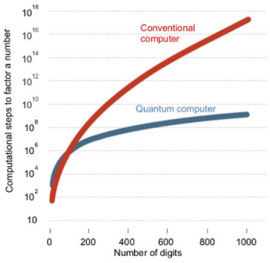

Figure 6. Comparison between speed of Quantum vs. Conventional computer.

4.2 Enhance calculation speed by quantum computer

The computer simulation used for medical development includes many calculations to model how the drug prototypes react with biological organisms, therefore the faster the device running the simulation is, the sooner each cycle of the simulation will finish, and when dealing with the large variety of molecular permutation, a small difference in speed can go a long way in reducing the total simulation time by days, months, or even years. From the image shown above, when the size of the number used increases, the conventional computer have its computational time exponentially increasing, while a quantum computer scales nearly linearly with this increase of data due to the principle of superposition that it is working with, and such scaling separates the speed between the two further due to the large datasets and long multi-digit numbers that represents molecular properties, which thus reduces the time taken to try out thousands of compounds in this phase of development.

4.3 Accommodation to Quantum Analysis

A drug is developed through multiple phases, from the initial drug discovery, pre-clinical research, clinical trials, and regulatory approval. The whole process could take around 15 years, with 3 to 5 years taken during drug discovery [39]. The other steps are human related, and therefore can be controlled separately, but the drug discovery process is the one that is computer-based and take up a large section in this whole process. Shortening the discovery section from years to just minutes with the speed of quantum computers will reduce the time a full process takes, thus alleviating the possible bottleneck in drug development.

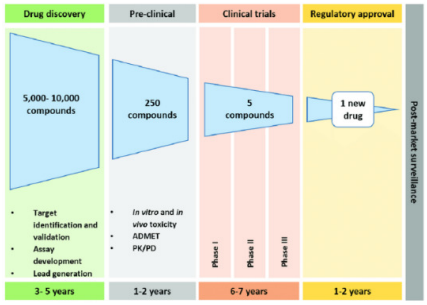

Figure 7. Compounds tested in drug development by stages: source

As mentioned before, the quantum computer can alleviate one bottleneck in drug discovery, but there are other things that scientists can do to speed up the process [40]. For one, they can select a certain type of compound that they see as the one with potential of becoming a drug, and go through a series of processes with that one compound before moving on to something else, which can be represented as a neural network. On the other hand, they can select the ones that will likely have unwanted side effects and get rid of them from the testing list sooner with the help of some algorithm, which again decreases the bottleneck, from a fair process to a selective process.

5 QML&QS: Its Future in Pharmaceutical R&D

5.1 Quantum computing save investigation fee

With the decreased development cycle of medicine, the investigation fee of medicine can be reduced. The total cost of drug development is mainly consisting with the expenses for conducting clinical trials and capital costs which is used for developing the drug. Even though the need for clinical trials trails cannot be eliminated, quantum computer can still reduce the capital costs. The time to calculate the correct molecule can be decreased from 10 to 15 years to below few months, which averagely cost about 10,000 to 50000$ for renting the quantum computer and this cost is below 1% of the average capital cost of 1.8 billion [41] now. The capital cost is inarguably consisting of other fee such as money to hire researchers, but because the decrease in production cycle, the HR cost will also be reduced, leading to a decrease in total investigation fee.

5.2 Pandemic prevention in the early phase

Quantum computer can stop the spread of large-scale pandemic by developing the sovereign remedy in the first few months before the germs spread to the whole world. With the increase in population and the increase of long-distance transportation, disease, especially ones which can transmitted through air or water, can easily become worldwide pandemic which mutates in high frequency based on the population infect-

ed. To stop such illness, sovereign remedy needs to be made before a new variant resistant to it spread out. According to a typical example of global Pandemic, covid-19, new variant of concern appears with a maximum integral of a year and a minimum of one month [42]. traditional research method needs about 9.1[43] years to develop a medicine. With quantum computer, the time it takes to make a new month can be decrease to few months or even weeks which is lower than the average time for a variant to develop. Therefore, quantum computer-based drug development is a powerful or even the only solution to global pandemic.

5.3 Personalize treatment and prevention for new diseases

Quantum computer's high speed allows to detect specific wrong gene and stimulate modification needed for that gene. Genetic diseases are mostly caused by wrong genetic sequence and can be treated by using gene editing techniques such as CRISPR and ZFNs, which commonly face the problem of off target [44]. off target effect might cause cancer and other abnormal symptom and limits the use of genetic edition treatment in human. Quantum computer can stimulate the editing process, proving data to modify the gene editor's DNA scissors, other protein and gRNA in advance before the actual treatment to make genetic edition treatment safer, and also reduced the time and money required. With the use of quantum computer based on stimulation, genetic editing in human can be apply in a greater scale and genetic based rare disease such as Wilson's Disease and glycosaminoglycan can be cure entirely.

5.4 Quantum computing in other industry

Quantum computer also have wide use in other industry such as food industry, artificial intelligence and in traffic optimization. Food industry can be improved by new kind of crop and animal. Quantum computer stimulation can help to invent more beneficial gene and edict gene of crops to improve production rate in this industry. Traffic optimization algorithm would perform better with greater amount of data. Traffic optimization can be improved by quantum computer by supporting using big data from theoretically all streets to preform large scale calculation to reduce traffic jam. Artificial intelligence can also become better with quantum computer based on its requirement for intense calculation. With more computing recourses provided, AI can do better in completing task such as image recognition and even make Strong AI possible to exist.

6 Conclusions

Our concept of the combination of QML and QS has great potential not only in the pharmaceutical R&D stages but also in fields such as genetic editing, agricultural development, and chemical optimization and creation in general. Therefore, we plan to conduct two more researches, one focusing on the specific implementation of the QML & QS system discussed in this paper and one on a further investigation and analysis of the application of our concept in other fields.

7 References

1. https://www.researchgate.net/figure/Drug-discovery-and-development-timeline-The-current-drug-approval-pipeline-can-take-15_fig1_308045230
2. https://www.frontiersin.org/articles/10.3389/fmed.2021.760762/full#B6
3. https://www.nebiolab.com/drug-discovery-and-development-process/
4. https://www.sigmaaldrich.com/HK/zh/technical-documents/technical-article/research-and-disease-areas/pharmacology-and-drug-discovery-research/hit-discovery-and-confirmation-for-early-drug-discovery
5. https://ascpt.onlinelibrary.wiley.com/doi/abs/10.1038/sj.clpt.6100235
6. https://cdmo.seqens.com/api-manufacturing/want-to-know-why-early-drug-development-costs-so-much/
7. https://www.fda.gov/patients/drug-development-process/step-1-discovery-and-development
8. https://www.nature.com/articles/s41586-019-1666-5
9. https://books.google.com/books?hl=zh-CN&lr=&id=92hzAwAAQBAJ&oi=fnd&pg=PR3&dq=the+use+of+machine+learning+in+quantum+computer&ots=SMFDhkUeXi&sig=dk0oDDV5o9FcEuxUVuVhrD-NVJMM#v=onepage&q=the%20use%20of%20machine%20learning%20in%20quantum%20computer&f=false
10. https://portlandpress.com/biochemist/article/26/4/13/2035/Relationship-between-protein-structure-and
11. https://pubmed.ncbi.nlm.nih.gov/10329133/
12. https://www.pellegrini.mcdb.ucla.edu/pellegrini/publication_pdfs/Pellegrini.pdf
13. https://www.chem.fsu.edu/chemlab/chm1046course/functional.html
14. https://proteopedia.org/wiki/index.php/Amino_acid_composition
15. https://www.ncbi.nlm.nih.gov/books/NBK26830/
16. https://www.rcsb.org/
17. https://www.news-medical.net/life-sciences/Primary-vs-Secondary-Assays-in-Preclinical-Testing.aspx
18. https://pubmed.ncbi.nlm.nih.gov/16181128/
19. https://openaccess.thecvf.com/content/CVPR2021/html/Wang_Dense_Contrastive_Learning_for_Self-Supervised_Visual_Pre-Training_CVPR_2021_paper.html
20. https://ieeexplore.ieee.org/abstract/document/9814838
21. https://arxiv.org/abs/2112.10740
22. https://www.deepspeed.ai/training/
23. https://sphweb.bumc.bu.edu/otlt/MPH-Modules/PH/PH709_BasicCellBiology/PH709_BasicCellBIology7.html
24. https://www.healthline.com/health/why-are-enzymes-important
25. https://www.medican-health.com/the-disadvantages-of-the-protein-diet/
26. http://prodata.swmed.edu/QCS/QCS_manuscript.pdf
27. https://www.mckinsey.com/industries/life-sciences/our-insights/pharmas-digital-rx-quantum-computing-in-drug-research-and-development

28. https://pdb101.rcsb.org/learn/guide-to-understanding-pdb-data/introduction
29. https://bitbucket.org/pierrepo/pdb2fasta/src
30. https://www.rcsb.org/structure/7R98
31. https://www.sigmaaldrich.com/US/en/technical-documents/technical-article/protein-biology/protein-structural-analysis/amino-acid-reference-chart
32. https://jmol.sourceforge.net/
33. https://spdbv.unil.ch/content.html
34. http://gpsprot.org/
35. http://severus.dbmi.pitt.edu/wiki-pi/
36. https://academic.oup.com/bioinformatics/article/19/suppl_1/i197/227962
37. https://www.fda.gov/patients/learn-about-drug-and-device-approvals/drug-development-process
38. https://www.geoexpro.com/articles/2016/07/supercomputers-for-beginners-part-iv
39. https://www.researchgate.net/figure/Drug-discovery-and-development-timeline-The-current-drug-approval-pipeline-can-take-15_fig1_308045230
40. https://www.elsevier.com/connect/breaking-bottlenecks-in-drug-discovery-and-development
41. (Paul SM, Mytelka DS, Dunwiddie CT, Persinger CC, Munos BH, Lindborg SR, Schacht AL (March 2010). "How to improve R&D productivity: the pharmaceutical industry's grand challenge". Nature Reviews. Drug Discovery. 9 (3): 203–214. doi:10.1038/nrd3078. PMID 20168317. S2CID 1299234.)
42. https://www.who.int/activities/tracking-SARS-CoV-2-variants
43. https://www.nature.com/articles/d41573-021-00190-9
44. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4191047/
45. Understanding the Fundamentals of Quantum Computing, Yew Kee Wong, April 2022, International Journal of Computer Science Trends and Technology, Vol. 10, No. 2.
46. Realization of Quantum Computers for Experimental Purposes, Yew Kee Wong, June 2022, International Journal of Information Technology, Vol. 8, No. 3.
47. Linear Superposition: A New Perspective, Yew Kee Wong, April 2022, International Journal of Information Technology, Vol. 8, No. 2.
48. Practicality of Quantum Computing & AI, Yew Kee Wong, April 2022, International Journal of Engineering Trends and Applications, Vol. 9, No. 2.