Hindawi
BioMed Research International
Volume 2019, Article ID 8427042, 15 pages
https://doi.org/10.1155/2019/8427042

Review Article

Application of Computational Biology and Artificial Intelligence Technologies in Cancer Precision Drug Discovery

Nagasundaram Nagarajan,[1] Edward K. Y. Yapp,[2] Nguyen Quoc Khanh Le,[1] Balu Kamaraj,[3] Abeer Mohammed Al-Subaie,[4] and Hui-Yuan Yeh[1]

[1]*School of Humanities, Nanyang Technological University, 14 Nanyang Dr, Singapore 637332*
[2]*Singapore Institute of Manufacturing Technology, 2 Fusionopolis Way, Singapore 138634*
[3]*Department of Neuroscience Technology, College of Applied Medical Sciences, Imam Abdulrahman Bin Faisal University, Jubail 35816, Saudi Arabia*
[4]*Department of Clinical Laboratory Sciences, College of Applied Medical Sciences, Imam Abdulrahman Bin Faisal University, Dammam, Saudi Arabia*

Correspondence should be addressed to Nagasundaram Nagarajan; naga25_sundar@yahoo.co.in and Hui-Yuan Yeh; hyyeh@ntu.edu.sg

Received 30 August 2019; Accepted 14 October 2019; Published 11 November 2019

Academic Editor: Ferdinando Chiaradonna

Artificial intelligence (AI) proves to have enormous potential in many areas of healthcare including research and chemical discoveries. Using large amounts of aggregated data, the AI can discover and learn further transforming these data into "usable" knowledge. Being well aware of this, the world's leading pharmaceutical companies have already begun to use artificial intelligence to improve their research regarding new drugs. The goal is to exploit modern computational biology and machine learning systems to predict the molecular behaviour and the likelihood of getting a useful drug, thus saving time and money on unnecessary tests. Clinical studies, electronic medical records, high-resolution medical images, and genomic profiles can be used as resources to aid drug development. Pharmaceutical and medical researchers have extensive data sets that can be analyzed by strong AI systems. This review focused on how computational biology and artificial intelligence technologies can be implemented by integrating the knowledge of cancer drugs, drug resistance, next-generation sequencing, genetic variants, and structural biology in the cancer precision drug discovery.

1. Introduction

Personalized or precision cancer therapy involves the identification of anticancer medicine for individual tumor molecular profiles, clinical features, and associated microenvironment of cancer patients [1, 2]. Precision medicine also aims to treat cancer more effectively with less adverse effects. According to a report by the International Agency for Research on Cancer (IARC), approximately 18.1 million of new registry on cancer cases and 9.6 million cancer-related deaths have been reported worldwide in 2018 [3]. Combined with classical cancer treatment methods, recent innovations in cancer treatment such as targeted chemotherapy, antiangiogenic agents, and immunotherapy were adapted by physicians on a case-to-case basis for better results [4]. In a number of instances, cancers such as hepatocellular carcinoma, malignant melanoma, and renal cancer often show intrinsic resistance to drugs without prior dosage of anticancer drugs [5]. In other cases, the initial response to the chemotherapy is remarkable. However, such a period is followed by a poor outcome, as cancer responds well to chemotherapy initially but later shows resistance due to development of resistance. Millions of cases regarding adverse drug resistance in cancer treatments are reported every year, which translates to a possibility of thousands of avoidable deaths. Such a dire situation thus calls for the designing of potential drugs. However, it is a time-consuming and complex process since each cancer patient responds differently to chemotherapy agent and its harmful effects are often unpredictable [6].

Ultimately, there is a crucial need to identify the primary mechanism with an ability to predict resistance to cancer therapies. The incorporation of tumor genetic profiling into clinical practice has improved the existing knowledge regarding the complex biology of tumor initiation and progression. Next-generation sequencing (NGS) is a platform commonly utilized by researchers to decode the genetic pattern of cancer patients, which allows for precision antitumor treatment based on their respective genomic profiles. It is clear that NGS plays a major role in treating diseases; however, it faces many technical challenges in its implementation. The highly accurate data obtained from NGS lead to the identification of a large set of genomic variations, in order to further identify the harmful variations of diseases. As such, specific modern computational algorithms are required to analyze and interpret the data. A number of computational tools have been developed to analyze the dataset that are integrated with genomic sequence and biochemical data on genetic polymorphism. Such tools will allow the prediction of functional consequences of deleterious polymorphism. Most of the tools were design followed by the combination of physicochemical properties of amino acids, protein structure information, and evolutionary sequence conservation analysis. Analyzing the functional consequence of genetic variation is not the limit; hence, directing such a analysis towards precision drug discovery and the structural attributes of drug interaction will bring about a new dimension in the cancer treatment. NGS technology usually produces huge set of data, and it is very difficult to analyze the data with the current existing tools. However, AI approaches have the capability to analyze NGS data in favor to identify suitable drug for individual patients.

Artificial intelligence (AI) proves to have an enormous potential in many areas of healthcare, including biomedical data analysis and drug discovery. The modern supercomputers and machine learning systems are able to explore the genetic data in order to identify the precision drugs. The key reason for applying AI in genetic data analysis is the completion of the human genome projects, which have reported huge amounts of genetic information. Over the last few years, the idea of using AI to accelerate precision drug identification to process and boost the success rates of pharmaceutical research programs has inspired a surge of activity in this area. Nowadays, biomedical studies can access extensive data sets due to the advancement of sequencing techniques and the accumulation of information on genetic variations. As such, there are currently greater prospects for precision medicine to come into the foreground of cancer treatment. As artificial intelligence makes use of the genetic profile for each patient, the right drug can be identified to cater to the patient's needs. Moreover, the artificial intelligence system is able to refine the key information in a short span of time. In this review, we aim to discuss about the integration of recent computational and biological techniques in order to develop a more effective cancer treatment. This will allow the fabrication of a precision drug identification platform through the application of artificial intelligence.

2. Literature Survey on Next-Generation Sequencing Technologies and Variant Calling Algorithms

In the early 1970s, a new technology was established to sequence the DNA molecule. However, its technical complexity, working cost, and limited availability of radioactive reagent made it difficult for the researchers to use this technology in the laboratory. Following this, the first-generation automated DNA sequence technology designed by Sanger and colleagues adopted a chain termination method [7]. Maiden et al. in 1990 used the DNA sequencing technology in the multilocus sequence-typing scheme for *Neisseria meningitidis* [8]. *Haemophilus influenzae* is the first environmental living microorganism that was sequenced in 1995 with the use of the Sanger sequencing methodology [9]. However, it is very expensive and time-consuming to sequence the whole human cell genome with this technology. In 1990, the human genome project was initiated with a goal to decode 3.2 billion base pairs of human genomes for biomedical research in disease diagnostic and treatment. Initially, the Sanger sequencing technology was used in this project worth 3.8 billion with international collaboration [10, 11]. Later in the early 2000s, another new technology emerged, namely, next generation sequencing (NGS) technology, which truly revolutionized the DNA sequencing process by reducing the time, cost, and labor. After 2010, genome sequencing was done on bacterial pathogens, which transfers the usage of technology from within the laboratory to public health practice. The sequencing technologies were used in several events of the critical infectious disease outbreak. Some examples include the cholera outbreak after a massive earthquake in Haiti during 2010 and the *E. coli* O104 : H4 disease outbreak, which was associated with consumption of fenugreek sprout in 2011 [12, 13]. In both cases, it was important to understand the virulent characteristic immediately, in order to reduce the progress of the disease, which will create massive morbidity and mortality. In these events, both academic and government research laboratories reacted quickly with NGS technology using crowd sourcing and open sharing of data. After these outbreaks, more public health laboratories have started to utilize NGS technology. Standardized NGS tests have been adopted in many countries' public laboratories for surveillance and in addition, NGS rated highly in specialized hospital laboratories [14, 15].

Between 1975 and 2005, the Sanger method was the predominant sequencing methodology. It has been considered as the gold standard for sequencing DNA that can produce 500–1000 bp long high-quality DNA reads. In 2005, 454 Life Science corporations introduced a revolutionized pyrosequencing technology referred to as "next generation sequencing (NGS) technology" [16]. This massive DNA sequencing technology is capable of reading and detecting thousand to millions of short DNA fragments in a single machine run without the need of cloning. Later versions of DNA sequencing technology were able to generate short reads (50–400 bp) and long reads (1–100 kb). The working mechanism and performance have been extensively discussed in many review articles [17, 18]. The MiSeq and MiniSeq technologies offer low to mid sample processing, moderate instrumentation cost and user-friendly working methods with automated and affordable cost per sample around $120 per 5 MB genome sequencing. Therefore, they have been the primary choice of technology for public health and disease diagnostic laboratories. The technologies HiSeq, NextSeq, and NovaSeq are considered as more suitable for core sequencing facility, irrespective of their high instrumentation cost since its cost per sample is low throughout the sequencing. However, they require automation for library preparation. By utilizing the full capacity of a sequencing machine, the cost can be effectively further reduced. In addition, the real-time testing is critical since the laboratory specific samples are sequenced in the laboratory-owned sequencing machines, which are highly tuned for the routine samples. For example, around 4000 isolates can be processed annually with a single MiSeq instrument and the use of v3 reagent, which would cover real-time testing in a laboratory. Amongst the NGS sequencing platforms, HiSeq as a product of Illumina generates the best quality of base call data. Ion Torrent, as a product of thermos fisheries, also performs sequencing by synthesis and its detection based on the hydrogen ions released during DNA polymerization that can be measured by the solid-state pH meter [19]. The PGM and S5 instruments are the IonTorrent equivalents for the Illumina MiniSeq and MiSeq; the ion proton is equivalent of Illumina NextSeq. The performance, the strength, and the weakness of prominent genomic sequencing platform have been compared and tabulated in Table 1.

Mutation/variation in the genetic code is considered as an important cause of cancer and thus it is the major focus in cancer research and treatment. The recent advancement in the sequencing technology can generate a huge set of data that can be explored by computational methods to identify the de novo mutation. Theoretically, all mutations including in the genomic region or variant allele frequency (VAF) can be identified with sufficient read depth. However, the noise in the files makes it difficult to identify them with confidence. A number of computational methods have been designed to identify the genetic variation or mutation from the complex DNA sequence reads (Table 2). The process involves a procedure with three features: read processing, mapping and alignment, and variant calling. As a first step, the read processing algorithms such as NGS QC Toolkit [20], Cutadapt [21], and FASTX Toolkit have been used to trim out the low quality and exogenous sequences such as sequencing adapter. During the library preparation of targeted sequencing, some of the protocol uses unique molecular identifiers (UMI) and PCR primers. In order to trim and remove the oligonucleotide, a customized read processing script must be developed. Second, the processed reads are mapped with the reference genome to identify the sequence, which is followed by base-by-base alignment. Most common applying, mapping, and alignment tools for DNA sequence include NovoAlign, BWA [22], and TMAP (for Ion Torrent reads) and as for RNA sequencing, splice-aware aligner tools such as STAR [23] and TopHat [24] are used. Genome Analysis Toolkits (GATKs) are the widely used tool for variant calling; following the procedures generally is important in this step such as PCR de-duplication, indel-realignment, and base quality recalibration [25, 26]. The final process is the variant calling, which is an important step for identifying correct variants/mutations from artifacts stemming from the prepared library, sequencing, mapping or alignment, and sample enrichment. A number of germ line and somatic variant calling tools have been developed which are freely available for analysis. The underlying knowledge is quite vary for somatic and germline variant calling tools. The rate of allele frequency in germline variants calling algorithms is expected to be 50 or 100%, and hence germline variant calling algorithms have accurately identified AA or AB or BB among these three genotypes, which fit the best [26–29]. Most artifacts occur in less frequency rate and are less likely to create a problem since in this case homozygous reference would be the most likely genotype. However, neglecting this type of artifact is not recommended in somatic variant calling because some original variants may also occur in very low frequencies in situations such as impure sample, rare tumor subclone, and in circulating DNA. Hence, the greatest challenge of the somatic variant calling algorithm is to accurately identify the low-frequency variants from artifacts, which can be done using advanced error correction technology and a more sensitive statistical model. Genetic variants can be classified into three major groups: insertion and deletion (indel), structural variant (such as duplication, translocation, copy number variation, etc.), and single nucleotide variant (SNV). Currently, only minimum number of variant caller algorithms is available to predict all these type variants, as they need specific trained algorithms. For single nucleotide variation and short indels (typically size ≤10 bp), the primary procedure is to check for nonreference nucleotide bases from the stack of sequence that cover each position. To evaluate the genotypic variants, mostly probabilistic modeling tools are used or to classify the artifact from the odds of variant. For structural variants and long indels, since the reads are too short to span over any variant, the focus is to identify the break points based on the patterns of misalignment with paired end reads or sudden change of read depth. Split reads assembly and de novo methods are frequently used for somatic variant analysis and long indel detection. GATK Unified Genotyper/Haplotype Caller, GAP, and MAQ are some of the tools used for germline variant calling [25, 26, 30, 31]. For somatic variant calling unified haplotype and genotype calling algorithms have been used, but the core algorithms are not formulated for this analysis following that it performs poorly for low-frequency somatic variants, and this information is highlighted in some independent studies as well as in the GATK documentation [32, 33]. Some other variant callers such as thunder and CRISP that are mainly used for pooled samples are also used for variant analysis [34].

Table 1: Comparison of performance, strengths and weaknesses of promising sequencing platforms.

Platform\instrument	Throughput range (Gb)	Read length (bp)	Strength	Weakness
Sanger sequencing				
ABI 3500/3730	0.0003	Up to 1 kb	Read accuracy and length	Cost and throughput
Illumina				
MiniSeq	1.7–7.5	1 × 75 to 2 × 150	Low initial investment	Run and read length
MiSeq	0.3–15	1 × 36 to 2 × 300	Read length, scalability	Run length
NextSeq	10–120	1 × 75 to 2 × 150	Throughput	Run and read length
HiSeq (2500)	10–1000	1 × 50 to 2 × 250	Read accuracy, throughput, low per sample cost	High initial investment, run length
HiSeq 3000/HiSeq 4000	105–1500	2 × 50 to 2 × 150	Read accuracy, throughput, low per sample cost	High initial investment, run and read length
NovaSeq 5000/6000	2000–6000	2 × 50 to 2 × 150	Read accuracy, throughput, low per sample cost	High initial investment, run and read length
IonTorrent				
PGM	0.08–2	Up to 400	Read length, speed	Throughput, homopolymers
S5	0.6–15	Up to 400	Read length, speed, scalability	Homopolymers
Proton	10–15	Up to 200	Speed, throughput	Homopolymers
Ion GeneStudio S5 prime System (ion 550″ chip)	10–50	Up to 200 (2 runs in one day)	Read length, speed, scalability	Homopolymers
Oxford nanopore				
MInION	0.1–1	Up to 100 kb	Read length, portability	High error rate, run length, low throughput
GridION X5	50–100	Up to 1000 kb	The GridION X5 offers real time, long-read, high-fidelity DNA and RNA sequencing.	High error rate
Pacific BioSciences				
PacBio RSII	0.5–1	Up to 60 kb(Average 10 kb, N50 20 kb)	Read length, speed	High error rate and initial investment, low throughput
Sequel	5–10	Up to 60 kb(Average 10 kb, N50 20 kb)	Read length, speed	High error rate
Sequel II	9–13	Up to 160 Gb	Read length, speed	Initial investment

3. Global Cancer Report

A reason for the majority of global deaths is the occurrence of noncommunicable diseases (NCDs) [35]. During the 21st century in almost every country of the world, cancer is the primary cause of deaths and this prevalent issue hinders the extension of life expectancy. In 2015, the World Health Organization (WHO) estimated that cancer is a dominant cause of mortality and morbidity before the age of 70 years in 91 of 172 countries, and in the rest of the 22 countries, it ranks as the third or fourth reason for death. Cancer morbidity and mortality are rapidly increasing worldwide. Ultimately, there are complex reasons such as the lack in the disease prevalence and distribution as well as an aging population. In addition, the population increase and its socioeconomic conditions serve as major causes of cancer death [36, 37]. Cancer incidence is mostly reported in developing countries, where the rising number of the disease is parallel by a modification in the genetic profile of common tumor genetic types. A serious observation made regarding the ongoing changes in the poverty-related and infection-related cancers is that they are increasingly common in some developed continents with the highest incomes, such as Oceania, Asia, North America, and Europe. The root cause of these cancers is often the modernized lifestyles [37–39]. However, the differing cancer tumor genetic profiles of various countries and even between specific ethnic zones signify that geographic variation still exists, with a persistence of local factors in populations at vastly different phases of economic and social transition. This is elucidated by the major differences in frequency of infection related to cancers, including stomach, liver, and cervix in the regions at opposite ends of the human development spectrum [38]. With regard to this information, a statistical analysis regarding the cancer burden worldwide in 2018 was made based on the GLOBOCAN 2018 observation of cancer morbidity and mortality analyzed by the International Agency for Research on Cancer (IARC) [40]. The same parameters as used in 2002 [41], 2008 [41], and 2012 [42] were taken into consideration to observe the cancer morbidity and mortality at the global level. As a result, an assessment has been made regarding the geographic differences observed across twenty predefined global regions. In the total number of cases, 11.6% lung cancer has been observed and as for the total number of cancer-related deaths, 18.4% were cause of lung cancer. For females, breast cancer is the next most common cancer at 11.6% followed by colorectal cancer at 10.2% and prostate cancer at 7.1% for incidence. As for mortality, the prominent causes are colorectal cancer at 9.2% followed by both liver and stomach cancer at 8.2%. In males, lung cancer is the most commonly occurring cancer and the primary reason for cancer

Table 2: List of tumor-normal somatic SNV callers and single-sample somatic and germline SNV callers sorted in alphabetical order.

Variant caller	Type of core algorithm	Type of variant	Type of variant caller
BAYSIC	Machine learning (ensemble caller)	SNV	Tumor-normal somatic SNV callers
CaVEMan	Joint genotype analysis	SNV	Tumor-normal somatic SNV callers
deepSNV	Allele frequency analysis	SNV	Tumor-normal somatic SNV callers
EBCall	Allele frequency analysis	SNV, indel	Tumor-normal somatic SNV callers
FaSD-somatic	Joint genotype analysis	SNV	Tumor-normal somatic SNV callers
FreeBayes	Haplotype analysis	SNV, indel	Tumor-normal somatic SNV callers
HapMuC	Haplotype analysis	SNV, indel	Tumor-normal somatic SNV callers
ISOWN	Supervised learning	SNV	Single-sample somatic and germline SNV caller
JointSNVMix2	Joint genotype analysis	SNV	Tumor-normal somatic SNV callers
LocHap	Haplotype analysis	SNV, indel	Tumor-normal somatic SNV callers
LoFreq	Allele frequency analysis	SNV, indel	Tumor-normal somatic SNV callers
LoLoPicker	Allele frequency analysis	SNV	Tumor-normal somatic SNV callers
MutationSeq	Machine learning	SNV	Tumor-normal somatic SNV callers
MuSE	Markov chain model	SNV	Tumor-normal somatic SNV callers
MuTect	Allele frequency analysis	SNV	Tumor-normal somatic SNV callers
OutLyzer	Noise level estimation	SNV	Single-sample somatic and germline SNV caller
Platypus	Haplotype analysis	SNV, indel, sv	Tumor-normal somatic SNV callers
Pisces	Poisson model on read count	SNV, indel	Single-sample somatic and germline SNV caller
PoreSeq	Nanopore specific	SNV, indel	Single-sample somatic and germline SNV caller
qSNP	Heuristic threshold	SNV	Tumor-normal somatic SNV callers
RADIA	Heuristic threshold	SNV	Tumor-normal somatic SNV callers
Seurat	Joint genotype analysis	SNV, indel,sv	Tumor-normal somatic SNV callers
SAMtools	Joint genotype analysis	SNV, indel	Tumor-normal somatic SNV callers
Shimmer	Heuristic threshold	SNV, indel	Tumor-normal somatic SNV callers
SNooPer	Machine learning	SNV, indel	Tumor-normal somatic SNV callers
SNVSniffer	Joint genotype analysis	SNV, indel	Tumor-normal somatic SNV callers
SOAPsnv	Heuristic threshold	SNV	Tumor-normal somatic SNV callers
SomaticSeq	Machine learning (ensemble caller)	SNV	Tumor-normal somatic SNV callers
SomaticSniper	Joint genotype analysis	SNV	Tumor-normal somatic SNV callers
Strelka	Allele frequency analysis	SNV, indel	Tumor-normal somatic SNV callers
Shearwater	Noise level estimation	SNV	Single-sample somatic and germline SNV caller
SiNVICT	Poisson model on read count	SNV, indel	Single-sample somatic and germline SNV caller
SNVer	Allele frequency analysis	SNV, indel	Single-sample somatic and germline SNV caller
SNVMix2	Genotype analysis	SNV	Single-sample somatic and germline SNV caller
SomVarIUS	Noise level estimation	SNV, indel	Single-sample somatic and germline SNV caller
SPLINTER	Noise level estimation	SNV, indel	Single-sample somatic and germline SNV caller
TVC	Ion Torrent specific	SNV, indel, SV	Tumor-normal somatic SNV callers
VarDict	Heuristic threshold	SNV, indel, SV	Tumor-normal somatic SNV callers
VarScan2	Heuristic threshold	SNV, indel	Tumor-normal somatic SNV callers
Virmid	Joint genotype analysis	SNV	Tumor-normal somatic SNV callers

mortality. In addition, prostate and colorectal cancers are the leading causes for incidence of cancer and liver and stomach cancer for cancer-related deaths. In the female population, breast cancer is the most commonly occurring cancer and the primary reason for cancer death followed by colorectal and lung cancer for incidence. Next to these former reasons, cervical cancer ranks fourth for both morbidity and mortality. Over 65% of newly identified cancer morbidity and mortality is caused by top ten cancer types worldwide observed.

4. Complication in Cancer Drug Discovery

From the beginning of human civilization, there has been a long history of drug discovery and development. The discovery and development of drugs is still a time-consuming process, whereby around 10–15 years needed to bring a single effective drug from the laboratory to market. Moreover, it requires huge investments, averaging from US$500 million to $2 billion [43, 44]. The high cost of drug development will probably affect the ability of patients with financial limitations to acquire the treatment. The expenditure to treat cancer in the USA will expect to rise from $124.57 billion in 2010 to $157.77 billion by 2020 [45]. In addition to discovery and development, drug production needs to fulfill satisfactory levels of toxicity, efficacy, and pharmacodynamics and pharmacokinetic profiles of the potential drugs candidate in in vitro and in vivo studies. In addition, preclinical studies were conducted to examine the efficacy and safety of the drug in humans in four different phases. Basically, drug development is hindered by a high rate of failure regarding their toxicity and efficacy profiles. According to the recent reports, even though new drug candidates exhibit high safety profile in Phase I trials, most of the drugs results fail due to poor efficacy in Phase II clinical trials [46]. Compared with other processes of drug discovery, oncology-related therapeutic discovery has the highest failure rate in clinical trials. Recent development in cancer treatment allows for the discovery of target specific drugs. However, only 1 of every 50K to 100K target specific anti-cancer drugs is approved by the US FDA. Furthermore, only 5% of anticancer drugs getting into Phase I clinical trials are often approved [47]. The target-specific anticancer drugs approach failed and it is still being investigated by oncologists to understand the underlying molecular mechanism. From the investigation reports, it is understood that in the development of cancer, more than 500 signaling molecules have been contributed [48]. However, the target-based drug discovery mostly focuses on inhibiting the identified signaling molecules. An investigation has to be made further in examining the drug-gable targets other than the reputed signaling molecules. Most of the drug targets are classified based on the preclinical studies; however, most prefindings are not exactly replicable in the clinical treatment. The number of potential drugs such as olaparib and iniparib showed promising results in preclinical stages. However, these preclinical in vitro and in vivo studies do not exactly consider the human cancer microenvironment [49–51]. In addition, the lack of quality in the pharmacodynamics and pharmacokinetics examination of drugs results in failure. Further poor testing strategies also majorly impact the drug's potential to translate from the preclinical findings to the medical treatment [52].

5. Cancer Drug Resistance

Drug resistance can be attributed to the decrease in the drug potency and efficacy to produce its desired effects. It stands as a big obstruction to treatment of the disease and affects the overall survival of the patient. Notably, local or locoregional, as well as distant tumor metastases leading in the paradox of therapy-induced metastasis (TIM), can result in resistance to anticancer treatments [5, 53, 54]. In a number of cases, tumors such as hepatocellular carcinoma, malignant melanoma, and renal cancer frequently show intrinsic resistance to anticancer drugs even without prior exposure to chemotherapy, resulting in a poor response during the initial stages of the treatment [5]. In some other cases, a chemotherapy agent may initially show its desired outcome. However, it is often followed by a poor response with harmful side effects due to the emergence of acquired drug resistance. So far, radiotherapy and surgery are the possible treatment methods for the removal of cancer cells. More systemic treatments are required to treat metastatic tumors or hematologic malignancies. Current forms of implementing systemic treatment are target-specific chemotherapy, immunotherapy, and antiangiogenic agents [53]. In most cases, drug resistance develops due to acquired and/or intrinsic genetic modulations. Intrinsic resistance may be induced by (a) modification of function and/or expression of the drug target, (b) drug breakdown, (c) changes in the drug carrying mechanism between the cellular membrane, (d) changes in the drug binding efficiency/efficacy with its binding target [54, 55]. Nuclear receptors and ATP-dependent membrane transporters are the primary factors that mediate the intrinsic cellular resistance [56]. Furthermore, cellular metabolic pathway systems, such as ceramide glycosylation, decrease the efficacy of anticancer drugs [57]. In addition, improved DNA damage repair mechanism increases drug resistance by reducing influx, increasing efflux, inhibiting drug accumulation through cell membrane transporters, and inactivating drugs [58, 59]. In reports of recent studies, the primary anticancer drugs had started to show signs of resistance against the known targets such as *TP53* [60]. Moreover, acquired drug resistance induced by environmental and genetic factors that enhance the development of drug resistant tumor cell or induce mutations of genes involved in relevant metabolic pathways [61, 62].

6. Computational Methods for Variant Classification

In recent days, the genetic mechanism behind human disease can be understood by next-generation sequencing technology approaches such as whole exome sequencing (WES) [63, 64]. Through WES sequencing technology, the genetic variants in the human genome can be detected. So far, several reports have documented that missense variants are the major cause of genetic diseases [65, 66]. However, not all the missense variants are involved in human genetic diseases as only deleterious variants are associated with Mendelian diseases, cancers, and undiagnosed diseases [67]. Identifying all deleterious variants through experimental validation is quite complicated work since it would require large amounts of labor and resources. Hence, computational methods have been developed to address this problem effectively by adopting different approaches like sequence evolutionary, sequence homology, and protein structural similarity [68–87]. Commonly there are three methods of prediction: (i) Sequence conservation methods, which generally note the degree of nucleotide base conservation at a particular position in comparison with the multiple sequence alignments information. (ii) Protein function-prediction methods that calculate the chance of a missense variant creating structural modification that affect protein function. (iii) Ensemble methods that integrate both sequence and structural information to calculate the effect of deleterious variants. In most cases for the missense variant identification tool development, all these methods have been adopted [88–90] and those tools are utilized in our studies [91–94]. VarCards is a database developed with the information on classified human genetic variants [95, 96]. It has integrated the functional consequences of allele frequencies, different computational methods, and other clinical and genetic information associated with all possible coding variants [97]. However, it is still difficult to understand the variance in performance of the computational methods, which differ under different conditions. Different studies have compared the performance of the missense variant prediction computational methods; however, they have not made use of the experimentally evaluated and considered benchmark datasets [98–103]. Particularly, these studies focus on assessing the receiver operating characteristic (ROC) curves. However, other parameters such as accuracy, specificity, sensitivity, and area under the curve (AUC) were not completely evaluated. There might be cases whereby geneticists and clinicians use computational tools to predict the harmful variants among the missense variants during the genetic counseling for known disease causing genes [104]. Hence, it is expected that these tools have to distinguish the pathogenic variants with a high-sensitivity rate [87]. In addition, VEST3 [78], REVEL [85], and M-CAP [87] are some recently developed algorithms that were not completely assessed in the previous studies. However, a recent study compared 23 computational pathogenicity prediction tools such as (i) ten function-prediction methods: fitCons [81], FATHMM [88], LRT [70], Mutation Taster [75], Mutation Assessor [76], PolyPhen2-HVAR [73], PolyPhen2-HDIV [73], SIFT [72], PROVEAN [77], and VEST3 [78]; (ii) four conservation methods: PhastCons [68], phyloP [69], GERP++ [74], and SiPhy [71]; and (iii) nine ensemble methods: DANN [83], CADD [79], Eigen [86], GenoCanyon [82], FATHMMMKL [84], MetaLR [80], M-CAP [87], REVEL [85], and MetaSVM [80]. The pathogenicity prediction scores of the 23 methods can be downloaded from the dbNSFP database v3.3 [105]. These predicted scores have been commonly used in medical genetics to identify the deleterious variant from the benign. Furthermore, prediction scores and other clinical information and genetic information were used alongside the VarCards [97] database. The cutoff values used to identify the deleterious missense variants were observed from ANNOVAR [106], dbNSFP database [105], and the original studies.

7. Artificial Intelligence in Precision Drug Discovery

The National Institute of Health (NIH) highlighted that precision medicine is an emerging strategy for disease prevention and treatment, which considers the individual variation in the gene, lifestyle, and environment [107]. This strategy helps researchers and doctors to prevent and treat the disease more accurately based on the genetic profile of the individuals. To make the strategy more comprehensive, it requires powerful supercomputer facilities and creative algorithms that can independently learn in an unprecedented way from the trained set of data. Artificial intelligence uses the cognitive ability of physicians and biomedical data for further learning to produce results. Artificial intelligence is broadly classified into three categories: artificial general intelligence, artificial narrow intelligence (ANI) and artificial super intelligence [108]. ANI is still in a stage of development and is expected to hit the market in by the next decade. ANI also has the caliber to deeply analyze the data set, find new correlation, draw conclusion, and support physicians. Well-established pharmaceutical companies have started to use the deep learning, super computers, and ANI in precision drug discovery process. Physicians may use the deep learning algorithms in many areas of disease diagnosis and treatment like oncology [109], dermatology [110], cardiology [111], and even in neurodegenerative disorders. However, developing such algorithms is crucial and critical in terms of exploring the knowledge of a physician in synchronizing with the algorithm development. Deep learning aims to identify unique genetic patterns in large genomic data sets and medical records and consequently identify genetic variations/mutations and their association with various diseases. A system of biological approach combined with artificial intelligence can form new algorithms that are able to monitor the changes inside the cell upon genetic modulation in the DNA [112]. Drug development is a highly complicated process that requires a huge amount of time and finances. However, in clinical trials, most of the drugs are rejected due to toxicity and lack of efficacy. Making the process faster and more cost-effective will have a tremendous impact on modern-day health care and how innovations made in drug discovery. Atomwise is the biopharma that uses an artificial intelligence-integrated supercomputing facility to analyze the database's information on small molecular structures. With the AI facility, Atomwise has launched a program to identify medicine to treat the Ebola virus. Through the AI technology, the company has found two better drugs, which are more promising in killing Ebola virus. Without such AI technology, such a drug discovery would take several years, however, with the AI system will doing it in less than one day [113]. Although the use of AI might seem promising in the discovery of drugs, these pharmaceutical companies will need to prove the safety and potential of their method with peer-reviewed research. In continuation of this short summary, the role of artificial intelligence methodologies in genetic variant/mutation identification from genetic data, virtual screening of small molecules, and molecular dynamics simulation programs has been elaborated under the appropriate subheading.

7.1. Artificial Intelligence Methods Applied to Identify Variants/Mutations from Genetic Data.

The aim of predictive models built based on machine learning approaches to draw conclusions from a sample of past observations and to transfer these conclusions to the entire population. Predicted patterns can be in different formats, such as nonlinear, linear, graph, cluster, and tree functions [114–116]. The machine-learning working mechanism is generally classified under four steps: filtering, data preprocessing, feature extraction, and model fitting and model evaluation. Supervised or unsupervised learning approaches are the two methods used in machine learning models. In supervised method to train the model, a known set of genetic information is required (for example, the start and end of the gene, promotors, enhancers, active sites, functional regions, splicing sites, and regulatory regions) in order to set the predictive models. This model is then used to find new genes that are similar to the genes of the training dataset. Supervised methods can only be used if a known training dataset of genetic codes available. Unsupervised methods are used if we are interested in finding the best set of unlabelled sequences that explain the data [117]. Machine learning methodologies have a wide range of application areas, and one of the most important applications is the identification of genetic variants and mutations [114, 118]. The machine learning approach called convolutional neural networks (CNNs) applied to the identification of genetic variants and mutations. The recently developed software's Torracina and Campagne analyzed genomic data to identify genetic variants/mutations and indel's using CNN method. Compared to previous methods [119], CNNs can substantially improve the performance in variant identifications [120]. Recurring variants in the genome content can be efficiently identified by means of this method [120, 121]. In the CNN method, the genetic sequence is analyzed as a 1D window using four channels (A,C,G,T) [122]. Genomic data used in machine learning models are classified under three categories 60% as training data, 30% as model testing data, and 10% as model validation data. Deep Variant is the recent method developed by Popolin et al. [123] for SNPs and indel detection with prediction precision >99% (at 90% recall). Deep sequence is the software used to identify the mutations [124], which also uses latent variables (a model using a decoder and an encoder network to predict the input sequence).

7.2. Applications of Artificial Intelligence in the Identification of Drugs. The virtual screening pipeline has been developed to reduce the cost of high throughput screening and further to increase efficiency and predictability in optimizing the potential small molecule [125, 126]. The strong generalization and learning process and machine-learning methods implementing aspects of AI models have been successfully implemented in different stages of the virtual screening pipeline. Virtual screening can be classified into two types: ligand- and structure-based virtual screening and with the former corresponding to situations wherein structural information from ligand-receptor binding is utilized and the latter to situations with its absence. Knowing the depth of the application of AI methods in virtual screening, we discussed the new findings in structure-based virtual screening driven by such approaches.

Advanced structure-based virtual screening methods have been developed with the help of potential AI algorithms based on nonparametric scoring functions. The correlation between the contributions to protein-ligand binding free energy and the feature vectors is implicitly observed through a data-driven manner from existing experimental data, which should enable the extraction of meaningful nonlinear relationships to obtain generalizing scoring functions [127–129]. The RF-based RF-score [128], SVM-based ID-score [130], and ANN-based NNScore are the AI-based non-predetermined scoring functions that have been developed to identify potential ligands with high accuracy rate. The recent advanced AI-based non-predetermined scoring methods outperform well in comparison with classical approaches in binding affinity predictions that have been discussed in several reviews [131–133].

In order to improve the scoring function performance, most of the AI techniques adopted the five major algorithms, namely, SVM, Bayesian, RF, deep neural network, and feed-forward ANN approaches. Ballester et al. reviewed the importance of machine learning regression algorithms in the enhancement of AI-based non-predetermined scoring functions to provide better binding affinity prediction between protein-ligand complexes. Based on the study, Ballester et al. developed a RF-based software to predict the protein-ligand docking score [134, 135]. Some other RF-based scoring functions such as B2B score [136], SFC score RF [137], and RF-IChem [138] have been developed to calculate the docking scores. On comparing with the above-listed tools, RF-score predictions are outstanding and thus it has been included with the istar platform, which involved large-scale protein-ligand docking [139]. SVM-based automated pipeline has been developed, capitalizing on the known weakness and strength of both ligand- and structure-based virtual screening.

For instance, from a pool of 18 million compounds to predict the novel c-Met tyrosine kinase inhibitors, Xie et al. [140] designed and used combined docking and SVM-based method. PROFILER is the automated workflow designed by Meslamani et al. [141] to identify the perfect targets having the highest probability of binding with bioactive compounds. PROFILER integrates with two structure-based approaches (protein-ligand-based pharmacophore searching and docking) and four ligand-based approaches (support vector regression affinity prediction, SVM binary classification, three-dimensional similarity search, and nearest neighbor affinity interpolation). In structure-based virtual screening, RF-score have been applied and performed well in identifying the targets. RF-Score-VS is the enhanced (DUD-E) scoring function that was trained on the full directory of useful decoy data sets (a set of 102 targets was docked with 15,426 active and 893,897 inactive ligands) [142].

The integration of AI techniques with structure-based virtual screening methods is the promising idea in the prediction of likely potential ligands. The AI technology has been adopted to improve the postprocessing process after the structure-based virtual screening process by reconsidering the scoring process calculated with docking algorithms using machine-learning models, with or without a consensus scoring. For example, AutoDock Vina can be incorporated with RF-Score-VS-enhanced method to get better performance in the virtual screening. The integration of advanced machine learning algorithms and automated ligand screening can help bring down the number of false positive and false negative predictions. Future work in this area is expected to consider physicochemical properties and structural information of the target protein.

7.3. Enhanced Molecular Dynamics Simulations with Artificial Intelligence. Computational chemistry is a potential technology to explore biochemical and structural behaviours of interest in a wide range of environments. Molecular dynamics simulations combined with multiscale molecular or quantum mechanics methods to measure the atomic level movement of a biomolecular system have been predominantly used to understand the behavior of molecules in recent studies [143–145]. However, it is too difficult to analyse the movement of large groups of atom in a stretch, and it requires powerful computational facilities. Integration of AI technology and computational chemistry can complete the high volume of simulation in an efficient way [146–148]. An established example is the construction of neural network potentials for high-dimensional systems with the Behler–Parinello symmetry function to asses thousands of atoms [149–151]. Many scientifically intensified problems have been explored recently such as solvation for Schrodinger equation [152], machine-learned density functional development [153–156], classification of chemical trajectory data, predictions of the molecular properties prediction of the excited state electrons [157, 158], many-body expansions [159], classification of chemical trajectory data [160], high-throughput virtual screening to identify novel materials [161–166], heterogeneous catalysts [167], and band gap prediction [168, 169].

Many advancements have been made in this field, such as introduction of reweighting correction to calculate the output at an estimated level of theory with high precision (for example: quantum chemistry methods) based on the output predicted at an inexpensive baseline theory level (for example: semi-empirical quantum chemistry), which has been examined for the estimation of thermochemical properties of active molecules [170] and more recently in the calculation of free energy changes during chemical reactions [171]. Even though it is a challenging task to combine AI algorithms and computational chemistry to explore the chemical datasets in order to identify the potential drug candidates in high magnitude of time, the molecular mechanics/quantum mechanics inspired artificial intelligence developers will likely be widely used to speed up the process while keeping quantum mechanical precision. This technical combination truly supporting AI approaches become a live technique in drug discovery.

8. Summary and Outlook

New targeted drugs for cancer treatment have to be developed to overcome cellular chemotherapy resistance and in addition must have the potential to inhibit “hub” genes. The primary role of those identified drugs is to achieve the highest therapeutic effect by eliminating tumor cells, with less adverse effects. Understanding the underlying mechanisms of the patient’s responses to cancer drugs and the unravelling of their genetic code would lead to the identification of new precision therapies that may improve the patient’s overall health and quality of life. Classical methods employed in the discovery of drugs are time- and cost-consuming. In response, computational biology has the efficiency to identify the precision drugs quickly. Current computational tools and software have an impact on the different phases of the drug discovery process. A number of studies have been performed by utilizing different computational approaches to identify the precision drugs that are suitable to particular genetic variant/s [91–94]. The methodology combined with the collection of genetic variants, prediction of pathogenicity using various computational tools, modeling the protein three-dimensional structure with particular variant/s, molecular docking of standard drug with variant/mutant structures, virtual screening to identify the specific drug, and performing molecular dynamics simulation allow for a better understanding of the efficacy of the drug (Figure 1). However, one limitation of the adopted methodology was that all the steps have been performed manually. It is necessary to bring radical change in the current computational methodology in order to identify precision drugs. We have shown in this review how artificial intelligence and computational biology approaches can be integrated to identify and discover cancer precision medicines.

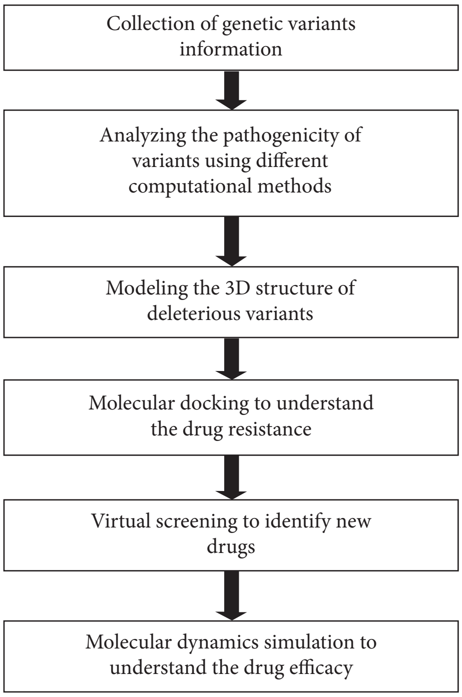
Collection of genetic variants information
Analyzing the pathogenicity of variants using different computational methods
Modeling the 3D structure of deleterious variants
Molecular docking to understand the drug resistance
Virtual screening to identify new drugs
Molecular dynamics simulation to understand the drug efficacy

Figure 1: Computational pipeline to analyze the variants and to identify the precision drugs.

Artificial intelligence integrated with computational biology has the potential to change the way drugs are designed and discovered. This approach was initially implemented at the Chapel Hill Eshelman School of Pharmacy at the University of North Carolina. The system is known as Reinforcement Learning for Structural Evolution, and it is well known by its acronym ReLeaSE. It is the computer software involving a set of algorithms incorporated with two neural networks programs, which can be considered to fulfill both roles of a student and a teacher. The teacher knows the linguistic rules and the syntax, which underlies the vocabulary of about 1.7 million known biologically active small molecules. Having been trained by the teacher, the student will understand the process over time and eventually become adept at finding the potential molecules that could be considered for developing new drugs. AI also positively influences precision medicine. The traditional drug discovery process of analyzing small data sets focused on a particular disease is offset by AI technology, which can rationally discover and optimize effective combinations of chemotherapies based on big datasets. The AI systems are built based on the experimental results and does not involve mechanistic hypotheses or any predictive models. Further, artificial intelligence technology can be applied in various ways such as to identify biomarkers, develop better diagnoses, and identify novel drugs. However, one important application of

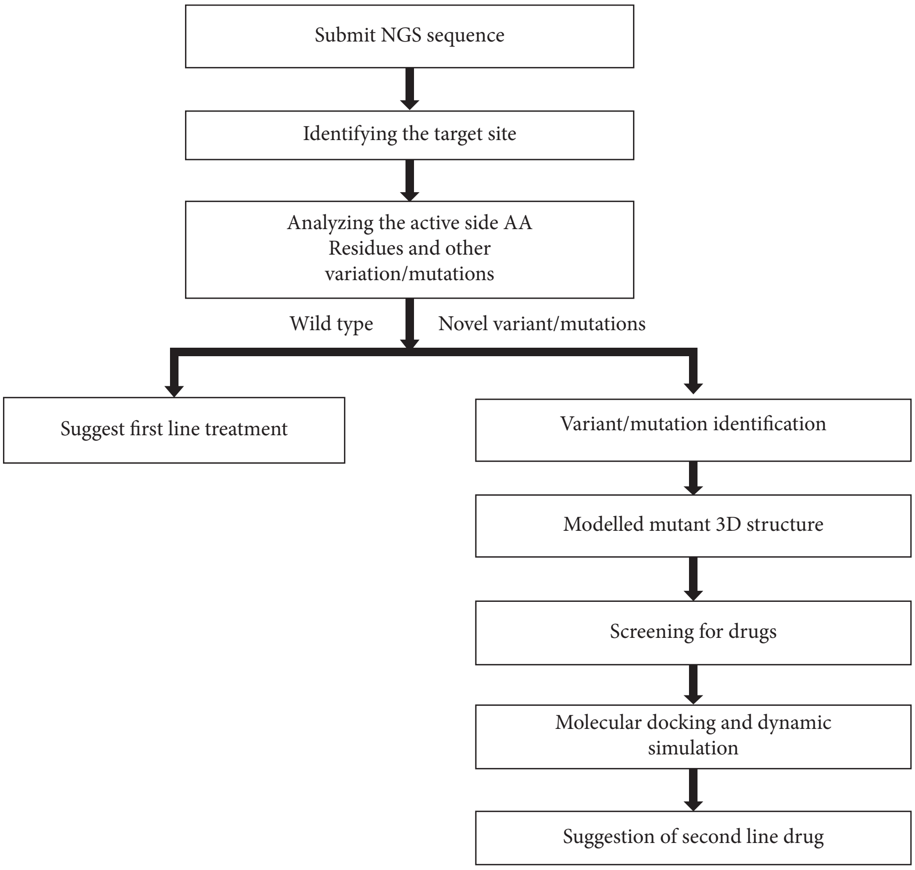
Submit NGS sequence
Identifying the target site
Analyzing the active side AA Residues and other variation/mutations
Wild type
Novel variant/mutations
Suggest first line treatment
Variant/mutation identification
Modelled mutant 3D structure
Screening for drugs
Molecular docking and dynamic simulation
Suggestion of second line drug

Figure 2: Suggested pipeline for cancer precision drug discovery.

artificial intelligence lies in finding target-based precision drugs. As we can see, artificial intelligence has acquired a key role in shaping the future of the health sector. An automated integrated system, involving the analysis of genetic variants by deep/machine learning methods, molecular modeling, high throughput structure-based virtual screening, molecular docking, and molecular dynamics simulation methods, will enable rapid and accurate identification of precision drugs (Figure 2). Developing an AI-based system will indeed be beneficial in the drug discovery process and in the discovery of cancer precision medicine.

Conflicts of Interest

The authors declared no conflicts of interest.

Authors' Contributions

NN and HYY were involved in designing the experiments. Acquisition, analysis, and interpretation of the data were performed by NN, HYY, EKYY, NQKL, BK, and AMAS. All the authors approved the manuscript.

Acknowledgments

The authors take this opportunity to thank the Nanyang Technological University for providing the facilities and for encouragement to carry out this work.

References

[1] C. Massard, S. Michiels, C. Ferté et al., "High-throughput genomics and clinical outcome in hard-to-treat advanced cancers: results of the MOSCATO 01 trial," *Cancer Discovery*, vol. 7, no. 6, pp. 586–595, 2017.

[2] F. Meric-Bernstam and G. B. Mills, "Overcoming implementation challenges of personalized cancer therapy," *Nature Reviews Clinical Oncology*, vol. 9, no. 9, pp. 542–548, 2012.

[3] F. Bray, J. Ferlay, I. Soerjomataram, R. L. Siegel, L. A. Torre, and A. Jemal, "Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA: A Cancer Journal for Clinicians*, vol. 68, no. 6, pp. 394–424, 2018.

[4] M. M. Jemal, J. Ludwig, D. Xia, and G. Szakacs, "Defeating drug resistance in cancer," *Discovery Medicine*, vol. 69, pp. 18–23, 2006.

[5] M. M. Gottesman, "Mechanisms of cancer drug resistance," *Annual Review of Medicine*, vol. 53, no. 1, pp. 615–627, 2002.

[6] World Health Organization, *Global Health Observatory*, World Health Organization, Geneva, Switzerland, 2018, http://who.int/gho/database/en/.

[7] F. Sanger, S. Nicklen, and A. R. Coulson, "DNA sequencing with chain-terminating inhibitors," *Proceedings of the National Academy of Sciences*, vol. 74, no. 12, pp. 5463–5467, 1977.

[8] M. C. J. Maiden, J. A. Bygraves, E. Feil et al., "Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms,"

Proceedings of the National Academy of Sciences, vol. 95, no. 6, pp. 3140–3145, 1998.

[9] R. Fleischmann, M. Adams, O. White et al., "Whole-genome random sequencing and assembly of Haemophilus influenzae Rd," *Science*, vol. 269, no. 5223, pp. 496–512, 1995.

[10] E. S. Lander, "Initial impact of the sequencing of the human genome," *Nature*, vol. 470, no. 7333, pp. 187–197, 2011.

[11] S. Tripp and M. Grueber, *Economic Impact of the Human Genome Project*, Battelle Memorial Institute, Columbus, OH, USA, 2011, http://www.battelle.org.

[12] L. A. KingF. Nogareda et al., "Outbreak of Shiga toxin-producing *Escherichia coli* O104 : H4 associated with organic fenugreek sprouts, France, June 2011," *Clinical Infectious Diseases*, vol. 54, no. 11, pp. 1588–1594, 2012.

[13] A. Mellmann, D. Harmsen, C. A. Cummings, E. B. Zentz, S. R. Leopold, and A. Rico, "Prospective genomic characterization of the German enterohemorrhagic *Escherichia coli* O104 : H4 outbreak by rapid next generation sequencing technology," *PloS One*, vol. 6, Article ID e22751, , 2011.

[14] C. Nadon, I. Van Walle, I. Chinen, J. Campos, E. Trees, and B. Gilpin, "Pulse Net International vision for the implementation of whole genome sequencing for global foodborne disease surveillance," *Eurosurveillance*, vol. 22, no. 23, 2017.

[15] M. Struelens, "Rapid microbial NGS and bioinformatics: translation into practice. Hamburg: June 9–11," *ECDC Roadmap for integration of molecular and genomic typing into European level surveillance*, Stockholm, Sweden, 2016.

[16] M. Margulies, M. Egholm, W. E. Altman, S. Attiya, J. S. Bader, and L. A. Bemben, "Genome sequencing in microfabricated high-density picolitre reactors," *Nature*, vol. 437, pp. 376–80, 2005.

[17] H. P. J. Buermans and J. T. den Dunnen, "Next generation sequencing technology: advances and applications," *Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease*, vol. 1842, no. 10, pp. 1932–1941, 2014.

[18] E. L. van Dijk, H. Auger, Y. Jaszczyszyn, and C. Thermes, "Ten years of next-generation sequencing technology," *Trends in Genetics*, vol. 30, no. 9, pp. 418–426, 2014.

[19] J. Rothberg and J. Myers, "Semiconductor sequencing for life," *Journal of Biomolecular Techniques*, vol. 22, pp. S41–S2, 2011.

[20] R. K. Patel and M. Jain, "NGS QC toolkit: a toolkit for quality control of next generation sequencing data," *PLoS One*, vol. 7, no. 2, Article ID e30619, 2012.

[21] M. Martin, "Cutadapt removes adapter sequences from high-throughput sequencing reads," *EMBnet.Journal*, vol. 17, no. 10, 2011.

[22] H. Li and R. Durbin, "Fast and accurate short read alignment with burrows-wheeler transform," *Bioinformatics*, vol. 25, no. 14, pp. 1754–1760, 2009.

[23] A. Dobin, C. A. Davis, F. Schlesinger et al., "STAR: ultrafast universal RNA-seq aligner," *Bioinformatics*, vol. 29, no. 1, pp. 15–21, 2013.

[24] C. Trapnell, L. Pachter, and S. L. Salzberg, "TopHat: discovering splice junctions with RNA-Seq," *Bioinformatics*, vol. 251, pp. 105–11, 2009.

[25] A. McKenna, M. Hanna, E. Banks et al., "The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data," *Genome Research*, vol. 20, no. 9, pp. 1297–1303, 2010.

[26] M. A. DePristo, E. Banks, R. Poplin et al., "A framework for variation discovery and genotyping using next-generation DNA sequencing data," *Nature Genetics*, vol. 43, no. 5, pp. 491–498, 2011.

[27] H. Li, "A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data," *Bioinformatics*, vol. 27, no. 21, pp. 2987–2993, 2011.

[28] D. C. Koboldt, Q. Zhang, D. E. Larson et al., "Varscan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing," *Genome Research*, vol. 22, no. 3, pp. 568–576, 2012.

[29] F. Xu, W. Wang, P. Wang, M. J. Li, C. Sham Pak, and J. Wang, "A fast and accurate SNP detection algorithm for next-generation sequencing data," *Nature Communications*, vol. 3, p. 1258, 2012.

[30] J. Qi, F. Zhao, A. Buboltz, and S. C. Schuster, "inGAP: an integrated next-generation genome analysis pipeline," *Bioinformatics*, vol. 26, no. 1, pp. 127–129, 2009.

[31] H. Li, J. Ruan, and R. Durbin, "Mapping short DNA sequencing reads and calling variants using mapping quality scores," *Genome Research*, vol. 18, pp. 851–858, 2008.

[32] H. Xu, J. DiCarlo, R. Satya, Q. Peng, and Y. Wang, "Comparison of somatic mutation calling methods in amplicon and whole exome sequence data," *BMC Genomics*, vol. 15, no. 1, p. 244, 2014.

[33] S. Sandmann, A. O. De Graaf, M. Karimi, B. A. Van Der Reijden, E. Hellstrom-Lindberg, and J. H. Jansen, "Evaluating variant calling tools for non-matched next generation sequencing data," *Science Reports*, vol. 74, pp. 31–69, 2017.

[34] V. Bansal, "A statistical method for the detection of variants from next-generation resequencing of DNA pools," *Bioinformatics*, vol. 26, no. 12, pp. i318–i324, 2010.

[35] A. R. Omran, "The epidemiologic transition: a theory of the epidemiology of population change," *The Milbank Memorial Fund Quarterly*, vol. 49, no. 4, pp. 509–538, 1971.

[36] O. Gersten and J. R. Wilmoth, "The cancer transition in Japan since 1951," *Demographic Research*, vol. 7, pp. 271–306, 2002.

[37] F. Bray, "Transitions in human development and the global cancer burden," in *World Cancer Report 2014*, B. W. Stewart and C. P. Wild, Eds., pp. 42–55, IARC Press, Lyon, France, 2014.

[38] M. Maule and F. Merletti, "Cancer transition and priorities for cancer control," *The Lancet Oncology*, vol. 13, no. 8, pp. 745-746, 2012.

[39] J. Ferlay, M. Colombet, I. Soerjomataram et al., "Global and Regional Estimates of the Incidence and Mortality for 38 Cancers," *GLOBOCAN 2018*, International Agency for Research on Cancer/World Health Organization, Lyon, France, 2018.

[40] D. M. Parkin, F. Bray, J. Ferlay, and P. Pisani, "Global cancer statistics, 2002," *CA: A Cancer Journal for Clinicians*, vol. 55, no. 2, pp. 74–108, 2005.

[41] A. Jemal, F. Bray, M. M. Center, J. Ferlay, E. Ward, and D. Forman, "Global cancer statistics," *CA: A Cancer Journal for Clinicians*, vol. 61, no. 2, pp. 69–90, 2011.

[42] L. A. Torre, F. Bray, R. L. Siegel, J. Ferlay, J. Lortet-Tieulent, and A. Jemal, "Global cancer statistics, 2012," *CA: A Cancer Journal for Clinicians*, vol. 65, no. 2, pp. 87–108, 2015.

[43] C. P. Adams and V. V. Brantner, "Estimating the cost of new drug development: is it really $802 million?," *Health Affairs*, vol. 25, no. 2, pp. 420–428, 2006.

[44] J. A. Di Masi, R. W. Hansen, and H. G. Grabowski, "The price of innovation: new estimates of drug development costs," *Journal of Health Economics*, vol. 22, pp. 15–185, 2003.

[45] A. B. Mariotto, K. Robin Yabroff, Y. Shao, E. J. Feuer, and M. L. Brown, "Projections of the cost of cancer care in the United States: 2010–2020," *JNCI Journal of the National Cancer Institute*, vol. 103, no. 2, pp. 117–128, 2011.

[46] G. A. Petsko, "When failure should be the option," *BMC Biology*, vol. 8, no. 1, p. 61, 2010.

[47] I. Kola and J. Landis, "Can the pharmaceutical industry reduce attrition rates?," *Nature Reviews Drug Discovery*, vol. 3, no. 8, pp. 711–716, 2004.

[48] S. C. Gupta, J. H. Kim, S. Prasad, and B. B. Aggarwal, "Regulation of survival, proliferation, invasion, angiogenesis, and metastasis of tumor cells through modulation of inflammatory pathways by nutraceuticals," *Cancer and Metastasis Reviews*, vol. 29, no. 3, pp. 405–434, 2010.

[49] H. Ledford, "Drug candidates derailed in case of mistaken identity," *Nature*, vol. 28, p. 483, 2012.

[50] B. B. Aggarwal, D. Danda, S. Gupta, and P. Gehlot, "Models for prevention and treatment of cancer: problems vs promises," *Biochemical Pharmacology*, vol. 78, no. 9, pp. 1083–1094, 2009.

[51] G. Francia and R. S. Kerbel, "Raising the bar for cancer therapy models," *Nature Biotechnology*, vol. 28, no. 6, pp. 561-562, 2010.

[52] C. G. Begley and L. M. Ellis, "Raise standards for preclinical cancer research," *Nature*, vol. 483, no. 7391, pp. 531–533, 2012.

[53] M. Clynes, *Multiple Drug Resistance in Cancer 2: Molecular, Cellular and Clinical Aspects*, Kluwer Academic Publishers, Dodrecht, Netherlands, 1998.

[54] J. M. L. Ebos, "Prodding the beast: assessing the Impact of treatment-induced metastasis," *Cancer Research*, vol. 75, no. 17, pp. 3427–3435, 2015.

[55] M. M. Gottesman, J. Ludwig, D. Xia, and G. Szakács, "Defeating drug resistance in cancer," *Discovery Medicine*, vol. 6, pp. 18–23, 2006.

[56] K. S. Sherlach and P. D. Roepe, "Drug resistance associated membrane proteins," *Frontiers in Physiology*, vol. 5, p. 108, 2014.

[57] B. Mansoori, A. Mohammadi, S. Davudian, S. Shirjang, and B. Baradaran, "The different mechanisms of cancer drug resistance: a brief review," *Advanced Pharmaceutical Bulletin*, vol. 7, no. 3, pp. 339–348, 2017.

[58] T. W. Synold, I. Dussault, and B. M. Forman, "The orphan nuclear receptor SXR coordinately regulates drug metabolism and efflux," *Nature Medicine*, vol. 7, no. 5, pp. 584–590, 2001.

[59] Y.-Y. Liu, T.-Y. Han, A. E. Giuliano, and M. C. Cabot, "Ceramide glycosylation potentiates cellular multidrug resistance," *The FASEB Journal*, vol. 15, no. 3, pp. 719–730, 2001.

[60] G. Housman, S. Byler, S. Heerboth et al., "Drug resistance in cancer: an overview," *Cancers*, vol. 6, no. 3, pp. 1769–1792, 2014.

[61] A. Sarkar and B. Schumacher, "DNA repair mechanisms in cancer development and therapy," *Frontiers in Genetics*, vol. 6, p. 157, 2015.

[62] S. W. Lowe, H. E. Ruley, T. Jacks, and D. E. Housman, "p53-dependent apoptosis modulates the cytotoxicity of anticancer agents," *Cell*, vol. 74, no. 6, pp. 957–967, 1993.

[63] B. Rabbani, M. Tekin, and N. Mahdieh, "The promise of whole-exome sequencing in medical genetics," *Journal of Human Genetics*, vol. 59, no. 1, pp. 5–15, 2014.

[64] S. Goodwin, J. D. McPherson, and W. R. McCombie, "Coming of age: ten years of next-generation sequencing technologies," *Nature Reviews Genetics*, vol. 17, no. 6, pp. 333–351, 2016.

[65] M. Lek, K. J. Karczewski, E. V. Minikel et al., "Analysis of protein-coding genetic variation in 60,706 humans," *Nature*, vol. 536, no. 7616, pp. 285–291, 2016.

[66] K. M. Boycott, M. R. Vanstone, D. E. Bulman, and A. E. MacKenzie, "Rare-disease genetics in the era of next-generation sequencing: discovery to translation," *Nature Reviews Genetics*, vol. 14, no. 10, pp. 681–691, 2013.

[67] D. G. MacArthur, T. A. Manolio, D. P. Dimmock et al., "Guidelines for investigating causality of sequence variants in human disease," *Nature*, vol. 508, no. 7497, pp. 469–476, 2014.

[68] A. Siepel, G. Bejerano, J. S. Pedersen, A. S. Hinrichs, M. Hou, and K. Rosenbloom, "Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes," *Genome Research*, vol. 15, no. 8, pp. 1034–1050, 2005.

[69] A. Siepel, K. S. Pollard, and D. Haussler, "New methods for detecting lineage-specific selection," in *RECOMB 2006. LNCS (LNBI)*, A. Apostolico, C. Guerra, S. Istrail, P. A. Pevzner, and M. Waterman, Eds., vol. 3909, pp. 90–205, Springer, Heidelberg, Berlin, Germany, 2006.

[70] S. Chun and J. C. Fay, "Identification of deleterious mutations within three human genomes," *Genome Research*, vol. 19, no. 9, pp. 1553–1561, 2009.

[71] M. Garber, M. Guttman, M. Clamp, M. C. Zody, N. Friedman, and X. Xie, "Identifying novel constrained elements by exploiting biased substitution patterns," *Bioinformatics*, vol. 25, no. 12, pp. i54–i62, 2009.

[72] P. Kumar, S. Henikoff, and P. C. Ng, "Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm," *Nature Protocols*, vol. 4, no. 7, pp. 1073–1081, 2009.

[73] I. A. Adzhubei, S. Schmidt, L. Peshkin et al., "A method and server for predicting damaging missense mutations," *Nature Methods*, vol. 7, no. 4, pp. 248-249, 2010.

[74] E. V. Kondrashov, D. L. Goode, M. Sirota, G. M. Cooper, A. Sidow, and S. Batzoglou, "Identifying a high fraction of the human genome to be under selective constraint using GERP++," *PLoS Computational Biology*, vol. 6, no. 12, Article ID e1001025, 2010.

[75] J. M. Schwarz, C. Rödelsperger, M. Schuelke, and D. Seelow, "MutationTaster evaluates disease-causing potential of sequence alterations," *Nature Methods*, vol. 7, no. 8, pp. 575-576, 2010.

[76] B. Reva, Y. Antipin, and C. Sander, "Predicting the functional impact of protein mutations: application to cancer genomics," *Nucleic Acids Research*, vol. 39, no. 17, p. e118, 2011.

[77] Y. Choi, G. E. Sims, S. Murphy, J. R. Miller, and A. P. Chan, "Predicting the functional effect of amino acid substitutions and indels," *PLoS One*, vol. 7, no. 10, Article ID e46688, 2012.

[78] H. Carter, C. Douville, P. D. Stenson, D. N. Cooper, and R. Karchin, "Identifying Mendelian disease genes with the variant effect-scoring tool," *BMC Genomic*, vol. 14, no. S3, 2013.

[79] M. Kircher, D. M. Witten, P. Jain, B. J. O'Roak, G. M. Cooper, and J. Shendure, "A general framework for estimating the relative pathogenicity of human genetic variants," *Nature Genetics*, vol. 46, no. 3, pp. 310–315, 2014.

[80] C. Dong, P. Wei, X. Jian et al., "Comparison and integration of deleteriousness prediction methods for nonsynonymous SNVs in whole exome sequencing studies," *Human Molecular Genetics*, vol. 24, no. 8, pp. 2125–2137, 2015.

[81] B. Liu, M. J. Hubisz, I. Gronau, and A. Siepel, "A method for calculating probabilities of fitness consequences for point mutations across the human genome," *Nature Genetics*, vol. 47, no. 3, pp. 276–283, 2015.

[82] Q. Lu, Y. Hu, J. Sun, Y. Cheng, K. H. Cheung, and H. Zhao, "A statistical framework to predict functional non-coding regions in the human genome through integrated analysis of annotation data," *Science Reports*, vol. 5, no. 1, p. 10576, 2015.

[83] D. Quang, Y. Chen, and X. Xie, "DANN: a deep learning approach for annotating the pathogenicity of genetic variants," *Bioinformatics*, vol. 31, no. 5, pp. 761–763, 2015.

[84] H. A. Shihab, M. F. Rogers, J. Gough et al., "An integrative approach to predicting the functional effects of non-coding and coding sequence variation," *Bioinformatics*, vol. 31, no. 10, pp. 1536–1543, 2015.

[85] N. M. Gaunt, J. H. Rothstein, V. Pejaver et al., "REVEL: an ensemble method for predicting the pathogenicity of rare missense variants," *The American Journal of Human Genetics*, vol. 99, no. 4, pp. 877–885, 2016.

[86] I. Ionita-Laza, K. McCallum, B. Xu, and J. D. Buxbaum, "A spectral approach integrating functional genomic annotations for coding and noncoding variants," *Nature Genetics*, vol. 48, no. 2, pp. 214–220, 2016.

[87] K. A. Jagadeesh, A. M. Wenger, M. J. Berger et al., "M-CAP eliminates a majority of variants of uncertain significance in clinical exomes at high sensitivity," *Nature Genetics*, vol. 48, no. 12, pp. 1581–1586, 2016.

[88] H. A. Bernstein, J. Gough, D. N. Cooper et al., "Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden Markov models," *Human Mutation*, vol. 34, no. 1, pp. 57–65, 2013.

[89] L. G. Day and R. C. Green, "Diagnostic clinical genome and exome sequencing," *The New England Journal of Medicine*, vol. 370, pp. 2418–2425, 2013.

[90] F. Cheng, J. Zhao, and Z. Zhao, "Advances in computational approaches for prioritizing driver mutations and significantly mutated genes in cancer genomes," *Briefings in Bioinformatics*, vol. 17, no. 4, pp. 642–656, 2016.

[91] N. Nagasundaram, H. Zhu, J. Liu et al., "Analysing the effect of mutation on protein function and discovering potential inhibitors of CDK4: molecular modelling and dynamics studies," *PLoS One*, vol. 7, no. 10, Article ID e0133969, 2015.

[92] N. Nagasundaram, H. Zhu, J. Liu et al., "Mechanism of artemisinin resistance for malaria PfATP6 L263 mutations and discovering potential antimalarials: an integrated computational approach," *Science Reports*, vol. 29, no. 6, Article ID 30106, 2016.

[93] N. Nagasundaram, C. R. Wilson Alphonse, P. V. Samuel Gnana, and R. K. Rajaretinam, "Molecular dynamics validation of crizotinib resistance to ALK mutations (L1196M and G1269A) and identification of specific inhibitors," *Journal of Cellular Biochemistry*, vol. 118, no. 10, pp. 3462–3471, 2017.

[94] N. Nagasundaram, K. Y. Edward, N. Q. Khanh Le, and H.-Y. Yeh, "In silico screening of sugar alcohol compounds to inhibit viral matrix protein VP40 of Ebola virus," *Molecular Biology Reports*, vol. 46, no. 3, pp. 3315–3324, 2019.

[95] K. A. Johansen Taber, B. D. Dickinson, and M. Wilson, "The promise and challenges of next-generation genome sequencing for clinical care," *JAMA Internal Medicine*, vol. 174, no. 2, pp. 275–280, 2014.

[96] C. F. Wright, D. R. FitzPatrick, and H. V. Firth, "Paediatric genomics: diagnosing rare disease in children," *Nature Reviews Genetics*, vol. 19, no. 5, pp. 253–268, 2018.

[97] J. Li, L. Shi, K. Zhang et al., "VarCards: an integrated genetic and clinical database for coding variants in the human genome," *Nucleic Acids Research*, vol. 46, no. D1, pp. D1039–D1048, 2017.

[98] J. Thusberg, A. Olatubosun, and M. Vihinen, "Performance of mutation pathogenicity prediction methods on missense variants," *Human Mutation*, vol. 32, pp. 58–368, 2011.

[99] D. G. Grimm, C.-A. Azencott, F. Aicheler et al., "The evaluation of tools used to predict the impact of missense variants is hindered by two types of circularity," *Human Mutation*, vol. 36, no. 5, pp. 513–523, 2015.

[100] P. Wei, X. Liu, and Y.-X. Fu, "Incorporating predicted functions of nonsynonymous variants into gene-based analysis of exome sequencing data: a comparative study," *BMC Proceedings*, vol. 5, no. S9, p. S20, 2011.

[101] F. Gnad, A. Baucom, K. Mukhyala, G. Manning, and Z. Zhang, "Assessment of computational methods for predicting the effects of missense mutations in human cancers," *BMC Genomics*, vol. 14, no. S7, 2013.

[102] C. Rodrigues, A. Santos-Silva, E. Costa, and E. Bronze-Da-Rocha, "Performance of in silico tools for the evaluation of UGT1A1 missense variants," *Human Mutation*, vol. 36, no. 12, pp. 1215–1225, 2015.

[103] E. König, J. Rainer, and F. S. Domingues, "Computational assessment of feature combinations for pathogenic variant prediction," *Molecular Genetics & Genomic Medicine*, vol. 4, no. 4, pp. 431–446, 2016.

[104] S. Richards, N. Aziz, S. Bale et al., "Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology," *Genetics in Medicine*, vol. 17, no. 5, pp. 405–423, 2015.

[105] X. Liu, C. Wu, C. Li, and E. Boerwinkle, "dbNSFP v3.0: a one-stop database of functional predictions and annotations for human nonsynonymous and splice-site SNVs," *Human Mutation*, vol. 37, no. 3, pp. 235–241, 2016.

[106] K. Wang, M. Li, and H. Hakonarson, "ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data," *Nucleic Acids Research*, vol. 38, no. 16, p. e164, 2010.

[107] F. Collins, *Precision Medicine Initiative*, National Institutes of Health, Bethesda, MD, USA, 2015, https://www.nih.gov/precision-medicine-initiative-cohort-program111.

[108] N. Bostrom, "Superintelligence: paths, dangers, strategies. Superintelligence: paths, dangers, strategies," 2014, http://ovidsp.112.

[109] D. Wang, A. Khosla, and R. Gargeya, "Deep learning for identifying metastatic breast cancer," 2016, http://arxiv.org/abs/1606.05718.

[110] A. Esteva, B. Kuprel, R. A. Novoa et al., "Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, vol. 542, no. 7639, pp. 115–118, 2017.

[111] G. Luo, G. Sun, K. Wang et al., "A novel left ventricular volumes prediction method based on deep learning network in cardiac MRI," *Computing in Cardiology*, vol. 2–5, 2010.

[112] Deep Genomics, 2017, https://www.deepgenomics.com/.

[113] Atomwise finds first evidence towards new Ebola treatments, 2017, http://www.atomwise.com/atomwise-finds-first-evidence-towards-newebola-treatments.

[114] M. W. Libbrecht and W. S. Noble, "Machine learning applications in genetics and genomics," *Nature Reviews Genetics*, vol. 16, no. 6, pp. 321–332, 2015.

[115] I. Sutskever, O. Vinyals, and Q. V. Le, *Advances in Neural Information Processing Systems*, MIT Press, Cambridge, MA, USA, 2014.

[116] T. Wasson and A. J. Hartemink, "An ensemble model of competitive multi-factor binding of the genome," *Genome Research*, vol. 19, pp. 2102–2112, 2009.

[117] K. Y. Yip, C. Cheng, and M. Gerstein, "Machine learning and genome annotation: a match meant to be?," *Genome Biology*, vol. 14, no. 5, p. 205, 2013.

[118] J. Zhou and O. G. Troyanskaya, "Predicting effects of noncoding variants with deep learning-based sequence model," *Nature Methods*, vol. 12, no. 10, pp. 931–934, 2015.

[119] R. Torracinta and F. Campagne, "Training genotype callers with neural networks," bioRxiv, 097469, 2016.

[120] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proceedings of the 25th International Conference on Neural Information Processing Systems*, pp. 1097–1105, Lake Tahoe, NV, USA, 2012.

[121] J. Lanchantin, Z. Lin, and Y. Qi, "Deep motif: visualizing genomic sequence classifications," 2016, http://arxiv.org/abs/1605,01133.

[122] H. Zeng, M. D. Edwards, G. Liu, and D. K. Gifford, "Convolutional neural network architectures for predicting DNA–protein binding," *Bioinformatics*, vol. 32, pp. 121–127, 2016.

[123] R. Poplin, D. Newburger, J. Dijamco et al., "Creating a universal SNP and small indel variant caller with deep neural networks," bioRxiv, 2018.

[124] J. Schreiber, M. Libbrecht, J. Bilmes, and W. Noble, "Nucleotide sequence and dnasei sensitivity are predictive of 3d chromatin architecture," bioRxiv, 103614, 2017.

[125] G. Schneider, "Virtual screening: an endless staircase?," *Nature Reviews Drug Discovery*, vol. 9, no. 4, pp. 273–276, 2010.

[126] T. Scior, A. Bender, G. Tresadern et al., "Recognizing Pitfalls in virtual screening: a critical review," *Journal of Chemical Information and Modeling*, vol. 52, no. 4, pp. 867–881, 2012.

[127] Q. U. Ain, A. Aleksandrova, F. D. Roessler, and P. J. Ballester, "Machine-learning scoring functions to improve structure-based binding affinity prediction and virtual screening," *Wiley Interdisciplinary Reviews: Computational Molecular Science*, vol. 5, no. 6, pp. 405–424, 2015.

[128] P. J. Ballester and J. B. O. Mitchell, "A machine learning approach to predicting protein-ligand binding affinity with applications to molecular docking," *Bioinformatics*, vol. 26, no. 9, pp. 1169–1175, 2010.

[129] S. L. Kinnings, N. Liu, P. J. Tonge, R. M. Jackson, L. Xie, and P. E. Bourne, "A machine learning-based method to improve docking scoring functions and its application to drug repurposing," *Journal of Chemical Information and Modeling*, vol. 51, no. 2, pp. 408–419, 2011.

[130] G.-B. Li, L.-L. Yang, W.-J. Wang, L.-L. Li, and S.-Y. Yang, "ID-Score: a new empirical scoring function based on a comprehensive set of descriptors related to protein-ligand interactions," *Journal of Chemical Information and Modeling*, vol. 53, no. 3, pp. 592–600, 2013.

[131] T. Cheng, Q. Li, Z. Zhou, Y. Wang, and S. H. Bryant, "Structure-based virtual screening for drug discovery: a problem-centric review," *The AAPS Journal*, vol. 14, no. 1, pp. 133–141, 2012.

[132] S.-Y. Huang, S. Z. Grinter, and X. Zou, "Scoring functions and their evaluation methods for protein-ligand docking: recent advances and future directions," *Physical Chemistry Chemical Physics*, vol. 12, no. 40, pp. 12899–12908, 2010.

[133] D.-L. Ma, D. S.-H. Chan, and C.-H. Leung, "Drug repositioning by structure-based virtual screening," *Chemical Society Reviews*, vol. 42, no. 5, pp. 2130–2141, 2013.

[134] P. J. Ballester, A. Schreyer, and T. L. Blundell, "Does a more precise chemical description of protein-ligand complexes lead to more accurate prediction of binding affinity?," *Journal of Chemical Information and Modeling*, vol. 54, no. 3, pp. 944–955, 2014.

[135] H. Li, K.-S. Leung, M.-H. Wong, and P. J. Ballester, "Improving AutoDock vina using random forest: the growing accuracy of binding affinity prediction by the effective exploitation of larger data sets," *Molecular Informatics*, vol. 34, no. 2-3, pp. 115–126, 2015.

[136] Q. Liu, C. K. Kwoh, and J. Li, "Binding affinity prediction for protein-ligand complexes based on β contacts and B factor," *Journal of Chemical Information and Modeling*, vol. 53, no. 11, pp. 3076–3085, 2013.

[137] D. Zilian and C. A. Sotriffer, "SFCscoreRF: a random forest-based scoring function for improved affinity prediction of protein-ligand complexes," *Journal of Chemical Information and Modeling*, vol. 53, no. 8, pp. 1923–1933, 2013.

[138] J. Gabel, J. Desaphy, and D. Rognan, "Beware of machine learning-based scoring functions-on the danger of developing black boxes," *Journal of Chemical Information and Modeling*, vol. 54, no. 10, pp. 2807–2815, 2014.

[139] H. Li, K. S. Leung, P. J. Ballester, and M. H. Wong, "Istar: a web platform for large-scale protein-ligand docking," *PLoS One*, vol. 9, Article ID e85678, 2014.

[140] Q.-Q. Xie, L. Zhong, Y.-L. Pan et al., "Combined SVM-based and docking-based virtual screening for retrieving novel inhibitors of c-met," *European Journal of Medicinal Chemistry*, vol. 46, no. 9, pp. 3675–3680, 2011.

[141] J. Meslamani, R. Bhajun, F. Martz, and D. Rognan, "Computational profiling of bioactive compounds using a target-dependent composite workflow," *Journal of Chemical Information and Modeling*, vol. 53, no. 9, pp. 2322–2333, 2013.

[142] M. Wojcikowski, P. J. Ballester, and P. Siedlecki, "Performance of machine-learning scoring functions in structure-based virtual screening," *Science Reports*, vol. 7, p. 46710, 2017.

[143] H. M. Senn and W. Thiel, "QM/MM studies of enzymes," *Current Opinion in Chemical Biology*, vol. 11, no. 2, pp. 182–187, 2007.

[144] G. D. M. Seabra, R. C. Walker, M. Elstner, D. A. Case, and A. E. Roitberg, "Implementation of the SCC-DFTB method for hybrid QM/MM simulations within the amber molecular dynamics package," *The Journal of Physical Chemistry A*, vol. 111, no. 26, pp. 5655–5664, 2007.

[145] H. M. Senn and W. Thiel, "QM/MM methods for biomolecular systems," *Angewandte Chemie International Edition*, vol. 48, no. 7, pp. 1198–1229, 2009.

[146] L. Li, J. C. Snyder, I. M. Pelaschier et al., "Understanding machine-learned density functionals," *International Journal of Quantum Chemistry*, vol. 116, no. 11, pp. 819–833, 2016.

[147] M. Rupp, "Machine learning for quantum mechanics in a nutshell," *International Journal of Quantum Chemistry*, vol. 115, no. 16, pp. 1058–1073, 2015.

[148] J. Behler, "Representing potential energy surfaces by high-dimensional neural network potentials," *Journal of Physics: Condensed Matter*, vol. 26, no. 18, Article ID 183001, 2014.

[149] J. Behler, “Perspective: machine learning potentials for atomistic simulations,” *The Journal of Chemical Physics*, vol. 145, Article ID 170901, 2016.

[150] J. Behler, “Constructing high-dimensional neural network potentials: a tutorial review,” *International Journal of Quantum Chemistry*, vol. 115, no. 16, pp. 1032–1050, 2015.

[151] J. Behler, “First principles neural network potentials for reactive simulations of large molecular and condensed systems,” *Angewandte Chemie International Edition*, vol. 56, no. 42, pp. 12828–12840, 2017.

[152] K. Mills, M. Spanner, and I. Tamblyn, “Deep learning and the Schroodinger equation,” *Physical Review A*, vol. 96, Article ID 042113, 2017.

[153] J. C. Snyder, M. Rupp, K. Hansen, K. R. Muller, and K. Burke, “Finding density functionals with machine learning,” *Physical Review Letters*, vol. 108, Article ID 253002, 2012.

[154] F. Brockherde, L. Vogt, L. Li, M. E. Tuckerman, K. Burke, and K. R. Muller, “Bypassing the Kohn-Sham equations with machine learning,” *Nature Communications*, vol. 8, no. 1, p. 872, 2017.

[155] K. Yao and J. Parkhill, “Kinetic energy of hydrocarbons as a function of electron density and convolutional neural networks,” *Journal of Chemical Theory and Computation*, vol. 12, no. 3, pp. 1139–1147, 2016.

[156] J. C. Snyder, M. Rupp, K. Hansen, L. Blooston, K. R. Muller, and K. Burke, “Orbital-free bond breaking via machine learning,” *The Journal of Chemical Physics*, vol. 139, no. 22, Article ID 224104, 2013.

[157] F. Liu, L. Du, D. Zhang, and J. Gao, “Direct learning hidden excited state interaction patterns from ab initio dynamics and its implication as alternative molecular mechanism models,” *Science Reports*, vol. 7, p. 8737, 2017.

[158] F. Häse, S. Valleau, E. Pyzer-Knapp, and A. Aspuru-Guzik, “Machine learning exciton dynamics,” *Chemical Science*, vol. 7, no. 8, pp. 5139–5147, 2016.

[159] K. Yao, J. E. Herr, and J. Parkhill, “The many-body expansion combined with neural networks,” *The Journal of Chemical Physics*, vol. 146, no. 1, Article ID 014106, 2017.

[160] B. K. Carpenter, G. S. Ezra, S. C. Farantos, Z. C. Kramer, and S. Wiggins, “Empirical classification of trajectory data: an opportunity for the use of machine learning in molecular dynamics,” *The Journal of Physical Chemistry B*, vol. 122, no. 13, pp. 3230–3241, 2018.

[161] A. Mannodi-Kanakkithodi, G. Pilania, T. D. Huan, T. Lookman, and R. Ramprasad, “Machine learning strategy for accelerated design of polymer dielectrics,” *Science Reports*, vol. 6, no. 1, Article ID 20952, 2016.

[162] T. D. Huan, A. Mannodi-Kanakkithodi, and R. Ramprasad, “Accelerated materials property predictions and design using motif- based fingerprints,” *Physical Review B*, vol. 92, no. 1, Article ID 014106, 2015.

[163] G. Pilania, C. Wang, X. Jiang, S. Rajasekaran, and R. Ramprasad, “Accelerating materials property predictions using machine learning,” *Science Reports*, vol. 3, no. 1, p. 2810, 2013.

[164] J. Lee, A. Seko, K. Shitara, K. Nakayama, and I. Tanaka, “Prediction model of band gap for inorganic compounds by combination of density functional theory calculations and machine learning techniques,” *Physical Review B*, vol. 93, no. 11, p. 115104, 2016.

[165] E. O. Pyzer-Knapp, K. Li, and A. Aspuru-Guzik, “Learning from the Harvard Clean Energy Project: the use of neural networks to accelerate materials discovery,” *Advanced Functional Materials*, vol. 25, no. 41, pp. 6495–6502, 2015.

[166] R. Gómez-Bombarelli, J. Aguilera-Iparraguirre, T. D. Hirzel et al., “Design of efficient molecular organic light-emitting diodes by a high-throughput virtual screening and experimental approach,” *Nature Materials*, vol. 15, no. 10, pp. 1120–1127, 2016.

[167] X. Ma, Z. Li, L. E. K. Achenie, and H. Xin, “Machine-learning-augmented chemisorption model for CO_2 electroreduction catalyst screening,” *The Journal of Physical Chemistry Letters*, vol. 6, no. 18, pp. 3528–3533, 2015.

[168] G. Pilania, J. E. Gubernatis, and T. Lookman, “Multi-fidelity machine learning models for accurate bandgap predictions of solids,” *Computational Materials Science*, vol. 129, pp. 156–163, 2017.

[169] G. Pilania, A. Mannodi-Kanakkithodi, B. Uberuaga, R. Ramprasad, J. Gubernatis, and T. Lookman, “Machine learning bandgaps of double perovskites,” *Science Reports*, vol. 6, no. 1, Article ID 19375, 2016.

[170] R. Ramakrishnan, P. O. Dral, M. Rupp, and O. A. von Lilienfeld, “Big data meets quantum chemistry approximations: the Δ-machine learning approach,” *Journal of Chemical Theory and Computation*, vol. 11, no. 5, pp. 2087–2096, 2015.

[171] L. Shen, J. Wu, and W. Yang, “Multiscale quantum mechanics/molecular mechanics simulations with neural networks,” *Journal of Chemical Theory and Computation*, vol. 12, no. 10, pp. 4934–4946, 2016.