Prediction of cancer driver genes and mutations

the potential of integrative computational frameworks

Nourbakhsh, Mona; Degn, Kristine; Saksager, Astrid Brix; Tiberti, Matteo; Papaleo, Elena

Published in:
Briefings in Bioinformatics

Link to article, DOI:
10.1093/bib/bbad519

Publication date:
2024

Document Version
Publisher's PDF, also known as Version of record

OXFORD

Briefings in Bioinformatics, 2024, **25(2)**, 1–16
https://doi.org/10.1093/bib/bbad519
Review

Prediction of cancer driver genes and mutations: the potential of integrative computational frameworks

Mona Nourbakhsh†, Kristine Degn†, Astrid Saksager, Matteo Tiberti and Elena Papaleo

Corresponding author: Tel./Fax: +4535257500; E-mail: elpap@dtu.dk
†These authors have contributed equally to this work

Abstract

The vast amount of available sequencing data allows the scientific community to explore different genetic alterations that may drive cancer or favor cancer progression. Software developers have proposed a myriad of predictive tools, allowing researchers and clinicians to compare and prioritize driver genes and mutations and their relative pathogenicity. However, there is little consensus on the computational approach or a golden standard for comparison. Hence, benchmarking the different tools depends highly on the input data, indicating that overfitting is still a massive problem. One of the solutions is to limit the scope and usage of specific tools. However, such limitations force researchers to walk on a tightrope between creating and using high-quality tools for a specific purpose and describing the complex alterations driving cancer. While the knowledge of cancer development increases daily, many bioinformatic pipelines rely on single nucleotide variants or alterations in a vacuum without accounting for cellular compartments, mutational burden or disease progression. Even within bioinformatics and computational cancer biology, the research fields work in silos, risking overlooking potential synergies or breakthroughs. Here, we provide an overview of databases and datasets for building or testing predictive cancer driver tools. Furthermore, we introduce predictive tools for driver genes, driver mutations, and the impact of these based on structural analysis. Additionally, we suggest and recommend directions in the field to avoid silo-research, moving towards integrative frameworks.

Keywords: driver genes; driver mutations; protein structural analysis; pathogenicity; predictive tools; computational research; cancer

INTRODUCTION

Cancer is a group of diseases characterized by uncontrolled cell growth and tumor formation due to (epi)genomic alterations, comprised within the hallmarks of cancer [1–3]. (Epi)genomic alterations in cancer occur in so-called driver genes which promote cancer development and progression (Figure 1A), conferring selective growth advantages to cancer cells [2–5]. Those mutations conferring growth advantages to cancer cells are called driver mutations, whereas mutations with no effect on the selective growth advantage of the cell are called passenger mutations [5] (Figure 1B). Driver mutations can impact protein structural stability and function, leading to the gain or loss of function [6–8] (Figure 1C).

To date, many computational tools and frameworks have been presented to predict driver genes, driver mutations, and the structural impact of protein variants. These methodologies are essential for assessing the pathogenic potential of cancer variants and designing the most suitable downstream experiments [9]. Computational methods typically rely on high-quality data, which can be obtained through next-generation sequencing technologies [10]. Thus, researchers should closely examine data quality and justify handling when developing tools [11]. Without robust tools and frameworks to analyze the growing data pool, we cannot infer meaningful biological interpretations, representing a bottleneck. Thus, solid bioinformatics pipelines are pivotal for increasing our knowledge of tumorigenesis and, eventually, drug target discovery.

Here, we provide an overview of relevant datasets and databases and a subset of predictive tools for driver genes, mutations and structural alterations. The aim is to address the advantages and limitations of the tools to give a comprehensive understanding of the status and discuss possible future directions. This review targets developers and users who may benefit from understanding the biological impact of the technical features of the tools. This review differs from other reviews in the

Mona Nourbakhsh is a PhD student in the Cancer Systems Biology group (Department of Health Technology, Technical University of Denmark, DTU, Lyngby, Denmark), and her research focuses on omics data analysis of cancer data and annotation and classification of cancer driver genes.
Kristine Degn is a PhD student in the Cancer Systems Biology group (Department of Health Technology, Technical University of Denmark, DTU, Lyngby, Denmark), and her research focuses on applying and developing computational methodologies to study protein structure alterations caused by cancer mutations.
Astrid Brix Saksager was a research assistant in the Cancer Systems Biology group (Department of Health Technology, Technical University of Denmark, DTU, Lyngby, Denmark) studying driver genes and driver mutations. She is now a PhD student in the Immunoinformatics and Machine Learning group (Department of Health Technology, Technical University of Denmark, DTU, Lyngby, Denmark).
Matteo Tiberti is a staff scientist at the Cancer Structural Biology group (Danish Cancer Institute, Copenhagen, Denmark). Matteo's research focuses on software development for cancer bioinformatics and the study of protein variants.
Elena Papaleo is an Associate Professor and leader of the Cancer Systems Biology group at DTU and group leader of the Cancer Structural Biology group at the Danish Cancer Institute. Her group's research focuses on integrative omics to study cancer drivers and structural methods to study protein–protein interactions and the effects of protein variants. The computational research is integrated by experimental validation with in vitro and cellular assays.
Received: June 9, 2023. **Revised:** November 27, 2023. **Accepted:** December 11, 2023

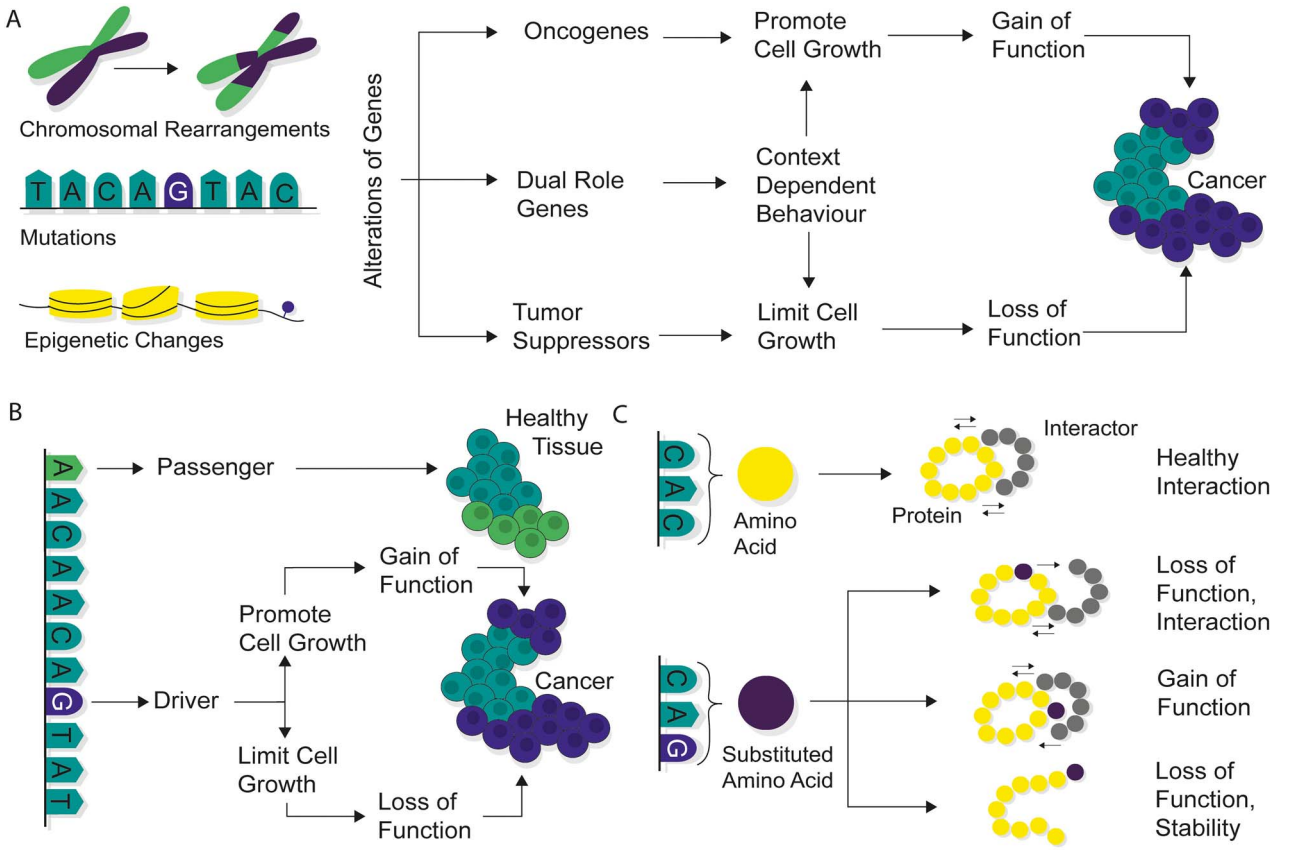

Figure 1. Illustration of the concept of driver genes, driver mutations, and structural impact. (**A**) Cancer involves dynamic changes in the genome caused by alterations such as mutations, epigenetic changes or chromosomal rearrangements. These alterations occur in driver genes that are divided into oncogenes, tumor suppressors, and dual role genes. Oncogenes normally promote cell growth, whereas tumor suppressors normally limit cell growth. The dual role genes exhibit both tumor suppressive and oncogenic behavior depending on the cellular context. The alterations lead to gain of function of oncogenes and loss of function of tumor suppressors which leads to uncontrolled cell growth and cancer. (**B**) Mutations can be categorized as passengers and drivers. Passengers are characterized by the absence of pathogenic impact. This can be due to their placement away from functional sites of the coding region or regulatory elements in the non-coding regions or the nature of the resulting amino acid substitution. (**C**) Genes are translated into proteins, and potential alterations including mutations are thereby also expressed. Understanding how these are expressed can give a mechanistic understanding as to why a particular alteration may give the cells a growth advantage. The top panel illustrates the healthy process of interaction, while the bottom panel illustrates some of the structural expressions of driver mutations.

field by combining the three aspects discussed so far, focusing on integrative frameworks. Other papers concentrate solely on driver genes [12], driver mutations [13, 14], protein-structure-based tools [15, 16] or on one computational approach such as machine learning [17, 18]. Contrarily, we aim to be agnostic towards the computational framework but rather discuss the process of building and validating tools with available data.

DATASETS AND DATABASES TO STUDY CANCER GENES AND MUTATIONS

Ten years ago, the main holdback to developing high-quality tools to differentiate drivers and passengers was the lack of high-quality curated datasets [19]. Since then, several datasets and databases have been developed, based on manual or automated curation (Tables S1–S3). Although manual curation poses the advantages of incorporating expert-based knowledge and critical judgment on a subject, it may suffer from the omission of important discoveries missed or dismissed by the curators. A database built upon literature mining may challenge the speed and ease of future updates of the database, predominantly if the data mining is based on dictionaries that are not standardized [20].

However, even with a perfectly unbiased curated dataset, the data sampling and balance are also potential sources of error in driver classification. This is especially accentuated in the study of cancer variants since the number of passengers is much higher than the one of drivers [21, 22]. To solve this imbalance, two general strategies are used: (i) remove passenger mutations [23] or (ii) increase the driver mutations, either using subsampling or synthetic additions in the training set [24]. Both solutions pose limitations related to loss of information and bias [25]. To overcome such bias, the user should carefully evaluate datasets for their origin, including cancer type, tumor stage, clinical profile and demographic composition, heterogeneity, data balance, data processing and curation method [26].

Pan-cancer initiatives (Table S1) such as The Cancer Genome Atlas [27] and the International Cancer Genome Consortium [28] have accelerated cancer research. In parallel, increased investigation of cancer drivers has led to the creation of datasets annotated with information on specific driver genes (Table S2) and mutations (Table S3). For example, OncoKB [29] is a precision oncology database offering guidance for clinicians and cancer researchers [29].

To study proteins using structural data, experimentally solved structures are available in the Protein Data Bank [30]. Furthermore, reliable databases of protein structure models have recently become available, including the AlphaFold2 database [31, 32] and emerging databases such as the ESM metagenomic atlas [33].

Driver gene prediction

Since the detection of the first cancer genes, the field of cancer genomics has exploded. This has led to the discovery of more than several hundred driver genes and continues to be a major goal of cancer research [34]. For example, Bailey *et al.* [35] performed a pan-cancer analysis of 33 cancer types to discover driver genes and mutations using 26 different tools. They found 299 driver genes of which 59 had not previously been reported by six other pan-cancer studies or in the Cancer Gene Census (CGC). Predicting the role of driver genes could contribute to active reversal of disrupted pathways manifested in the hallmarks of cancer [1], which has been accelerated by computational methods in the last decade. We here present a subset of driver gene prediction tools, categorized according to their computational approach (Figure 2).

Interaction network construction

Network-based approaches aim to model the role and impact of each gene in a network [36]. Nodes represent genes and edges represent interactions between them [37, 38]. Additionally, some of these methods employ the concept of influence, where genes with the greatest influence in the network are likely the ones driving carcinogenesis [39–41]. Examples of such tools are iMaxDriver [41], GenHITS [309], KatzDriver [40] and DriverGroup [42], which use interaction and/or gene expression data to construct the network. KatzDriver includes transcription factor-transcription factor and transcription factor-mRNA regulatory interactions and calculates the relative gene level impact in the regulation to measure driver gene status [40]. Two features of DriverGroup [42] worth highlighting are the detection of groups of driver genes and its ability to detect both coding and non-coding groups. Despite the important role of non-protein-coding genes in cancer [43], few tools focus on detecting non-coding driver genes. These tools are challenged by the increased search space of the non-coding genome compared to the coding genome and the low number of known driver genes with non-coding alterations [26], exemplifying data imbalance and ascertainment bias. Other tools integrate networks and/or gene expression data with additional sources of omics data such as mutation, copy number variation and DNA methylation data. For instance, AMARETTO [44], DriverFinder [45], CBNA [37], PRODIGY [46], LNDriver [47], OncoIMPACT [48], iPDG [49] and DriverSubNet [50] fall into this category. AMARETTO integrates multiple data sources to identify potential driver genes and connects these with co-expressed target genes that they control to create regulatory modules [44]. DriverFinder [45] accounts for the influence of gene length on the predictions, a distinctive characteristic compared to most other tools. Similarly to DriverGroup [42], CBNA [37] predicts both coding and non-coding drivers. Notable features of PRODIGY [46] and OncoIMPACT [48] are their ability to predict driver genes at a patient-specific level. Finally, other network-based tools employ the concept of random walks to investigate the relationship between genes in the network. Examples are Driver_IRW [38], Subdyquency [51], MEXCOwalk [52] and RLAG [53]. As gene products seldom exert their effect in isolation, these tools offer an effective way to represent this mechanism. On the other hand, non-coding drivers may be overlooked, and they mainly consider interactions between transcription factors and target genes, excluding other types of interactions relevant as cancer-driving mechanisms [37–41, 45].

Multi-omics data integration

Other tools integrate multi-omics data as well, but do not rely on networks. For example, frDriver [54] integrates functional impact scores with gene expression and mutation data and is built on Bayesian probability and multiple linear regression models. Moreover, a limited number of tools integrate DNA methylation to predict driver genes despite the known fact that DNA methylation aberrations contribute to cancer progression. MethSig [55] focuses on promoter hypermethylation as an inactivating mechanism of tumor suppressors and predicts DNA methylation driver genes through a novel statistical framework. Similarly, iEDGE [56] integrates gene expression profiles with (epi-)DNA alterations and performs differential expression analysis to find cis and trans genes associated with the (epi-)DNA alteration.

An advantage of methods combining various -omics data lies in this data integration. Multiple data types provide complementary sources of information that can provide a more comprehensive picture of the underlying mechanisms, potentially improving performance. On the other hand, the integration of different data types is a challenge. Different datasets may be obtained from different sources, leading to a lack of standardization and unintended confounders. Moreover, it is not guaranteed that all data types required by a given integrative tool will be available to the user, potentially limiting their applicability. Additionally, methylation and copy number variations are largely understudied compared to mutational impacts. Hence, integrating these additional layers can contribute to novel insights within cancer biology.

Machine learning

An example of a supervised machine learning tool for driver gene prediction is DriverML [57] that combines machine learning with a weighted score test. The weights represent the quantification of the functional impacts that different mutation types have on the protein [57]. DriverML is thus a representative of a handful of tools that employ the functional impact of mutations on the protein product. They aim to better predict lowly recurrent mutated genes and genes that are mutated in the later stages of tumorigenesis [58, 59].

Other supervised approaches are based on one-class support vector machines (SVMs), which have the advantage of overcoming class imbalance issues [26]. SVMs are for example used in sysSVM2 [60]. sysSVM2 allows for the prediction of driver genes at the single patient level [60], a feature that is also available in driveR [61]. driveR uses a multi-task learning model to obtain cancer type-specific probabilities of genes being drivers or non-drivers. Multi-task learning is also implemented in MTGCN [62]. This seems to be a promising strategy for cancer type-specific modeling to detect cancer-specific driver genes, which otherwise may not be predicted using pan-cancer models [26].

Neural networks are also represented among the machine learning tools used to identify driver genes, although to a lower extent. Examples include FI-Net [63], DeepDriver [64] and DeepCues [65].

Finally, a relatively underexplored field within driver gene prediction is the intersection between network-based and machine learning approaches, i.e. graph-based machine learning [26], implemented in for example MoProEmbeddings [66].

The performance of the above-mentioned machine learning tools is dependent on (i) the quality of the training data; (ii) the amount of available data for training or validation; (iii) the curation of positive sets of known driver genes that are difficult to define due to context dependency; and (iv) creation of negative sets of non-driver genes that are likewise difficult to define. Additionally, obtaining cancer type-specific known driver genes

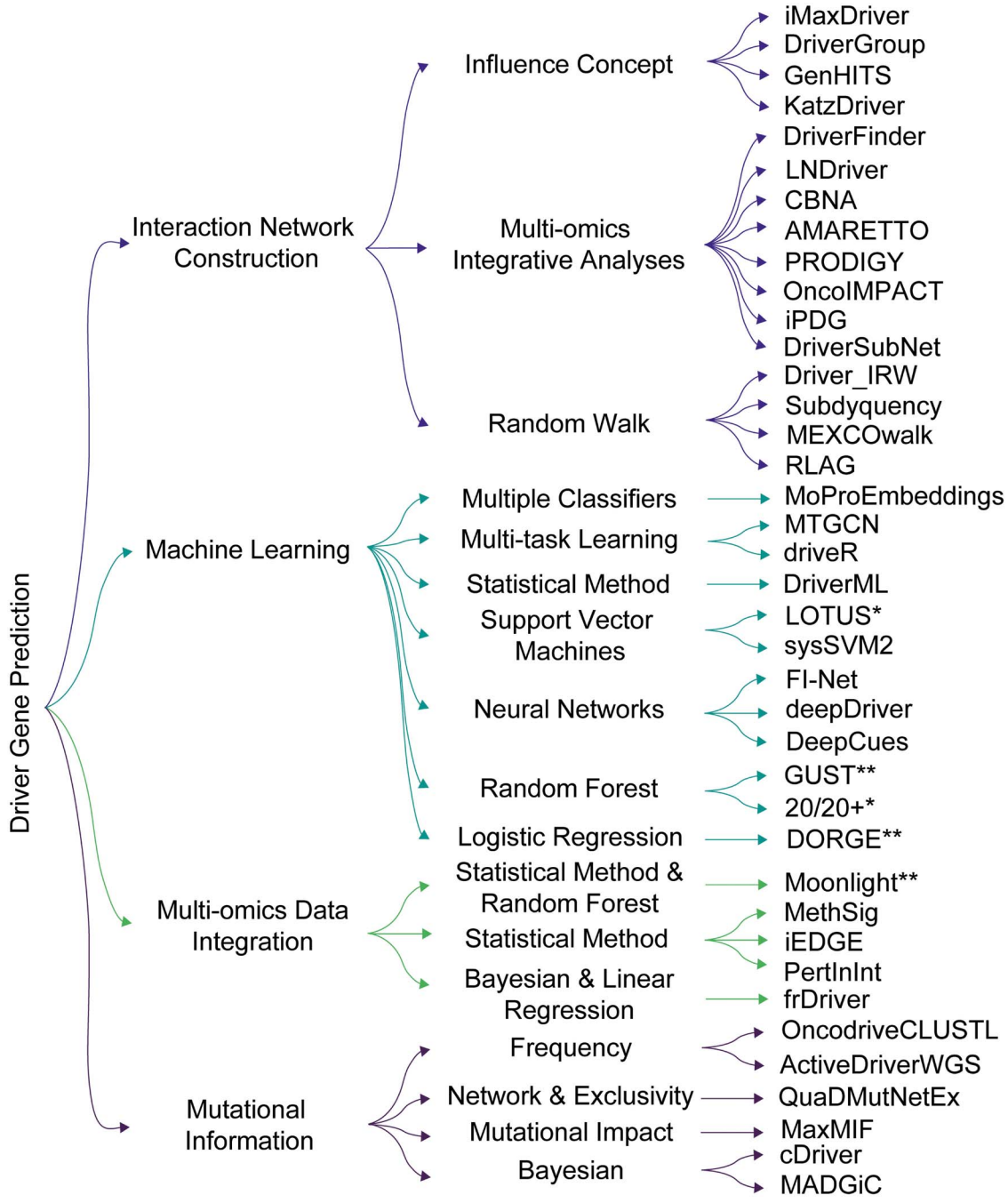

Figure 2. Overview of driver gene prediction tools. All driver gene prediction tools discussed in this review are listed to the right. These tools are divided into four main categories based on the underlying computational architecture. These four categories are subcategorized. **Driver gene prediction tools that predict tumor suppressors, oncogenes and dual role genes. *Driver gene prediction tools that predict tumor suppressors and oncogenes.

is challenging, creating known positive sets too small to yield reliable results. Some cancer types lack the volume of available data to generate a fair training set. One may overcome these issues by applying pan-cancer models [26, 67], yet at the sacrifice of accuracy. For instance, in a comparison among predicted driver genes in different cancer types, markedly different candidate driver genes were found among non-organ-related cancer types [68].

Mutational information

A large number of driver gene prediction tools utilizes and analyzes mutation data to identify driver genes. Many tools compare observed mutation frequencies with a background mutation model to discover driver genes such as OncodriveCLUSTL [69] and ActiveDriverWGS [70], yet an accurate background mutation model is difficult to create due to tumor heterogeneity. These methods are also challenged by driver genes with a low mutation frequency [50, 54, 64, 67, 71–73]. A proposed methodology to overcome such a limitation is QuaDMutNetEx [72] which combines mutual exclusivity and network approaches and is suited for discovering driver genes with low mutation frequency. Another example is MaxMIF [73] which integrates somatic mutation and protein–protein interaction data using a maximal mutational impact function. Alternative approaches include cDriver [74] and MADGiC [71] which are based on Bayesian

frameworks. The advantage of mutation-based tools is their applicability to additional mutation datasets [37]. However, driver genes are prone to diverse types of (epi)genomic alterations that risk being overlooked. Methods that apply scores of functional impacts might be able to mitigate the described problems since they are not based exclusively on the recurrence of the mutations [58].

Prediction of tumor suppressors, oncogenes and dual-role genes

Cancer driver genes are classified into three categories: tumor suppressors (TSGs), oncogenes (OCGs) and dual-role genes, which exhibit both tumor-suppressive and oncogenic behavior depending on the cellular context [5, 75–79]. Many tools offering driver gene classification into TSGs or OCGs are based on machine learning. For instance, GUST [80] and 20/20+ [81] use random forests to predict TSGs, OCGs and passenger genes based on 10 and 24 features, respectively. DORGE [82] uses two elastic net-based logistic regression models (DORGE-OCG and DORGE-TSG), which include 75 features broadly divided into mutational, genomic, epigenetic and phenotypic features. Furthermore, using two separate classifiers allow for the prediction of dual-role genes. LOTUS [67] uses one-class SVM and integrates mutation frequency, functional impact and pathway-based information in terms of protein–protein interaction networks and allows for the prediction of driver genes in both a pan-cancer and a cancer type-specific setting using a multitask learning strategy.

In 2020, we also contributed to this field with Moonlight [83] that predicts OCGs, TSGs and dual-role genes using multi-omics data. In Moonlight, gene expression data and information about biological processes (BPs) are integrated in a primary layer to detect genes defined as oncogenic mediators. The prediction of oncogenic mediators is carried out using either an expert-based approach or a machine learning approach. The expert-based approach utilizes patterns of opposing cancer-related BPs for predictions. For example, if these two processes are apoptosis and cell proliferation, then those differentially expressed genes (DEGs) with a positive effect on apoptosis and a negative effect on cell proliferation are deemed putative TSGs and vice versa for putative OCGs. The machine learning approach predicts the oncogenic mediators using a random forest model. Since using gene expression data alone results in a significant number of false positives, one or multiple layers of mechanistic evidence are required to improve the performance of the method. Such secondary layers can be, for example, DNA methylation, copy number variation, mutation or chromatin accessibility data. If evidence for the deregulation of an oncogenic mediator is provided through this secondary layer, the user can retain those oncogenic mediators as the final set of driver genes. This allows a mechanistic explanation of the activation or inactivation of the oncogenic mediators. We recently automatized the integration of a secondary mutational layer in a new function available in the second version of Moonlight, Moonlight2 [84]. Here, the oncogenic mediators containing at least one driver mutation are retained as the driver genes. Besides the classification of mutations, the new implementation allows for assessment of the effects of mutations on the transcriptional, translational, translational and protein structure/function level, thereby aiding mechanistic explanations of the deregulated driver genes, ultimately illustrating an integrative computational framework (Figure 3). Another example of an integrative framework is PertInInt [85]. PertInInt assesses the enrichment of somatic mutations within genes in functional sites and integrates interaction site information with evolutionary conservation and domain membership information as well as gene-level mutation data.

Only a limited number of predictive tools classify driver genes into TSGs and OCGs. Similarly, few studies have characterized dual-role genes with examples including Shen *et al.* [79] and Datta *et al.* [75]. These three gene classes drive cancer development through different biological mechanisms, and more tools distinguishing these categories, in a context-dependent manner, are needed to increase our understanding of cancer biology.

Summary and recommendations

Numerous driver gene prediction tools have been developed. However, these tools lack consistency in terms of predicted genes and predicted number of genes [63, 68, 81]. While the true number of driver genes in a cancer (sub)type is unknown, discovering an adequate number of driver genes is vital. Tools suffering from under-selection predict too few driver genes, potentially overlooking important genes. On the other hand, tools suffering from over-selection risk a large false positive rate and complicate subsequent experimental studies [57, 63].

Validating driver gene prediction tools is challenging as a gold standard of known driver genes and a universally accepted standard for this procedure does not exist, illustrating a limitation for benchmarking studies [48, 49, 73, 74, 81]. Most studies evaluate the performance of driver predictors by comparing the overlap between the predicted driver genes and driver genes listed in the COSMIC CGC [86] and the Network of Cancer Genes (NCG) [87] databases. Moreover, the original list of known driver genes by Vogelstein *et al.* is often used [5]. However, CGC and the list by Vogelstein *et al.* are embedded in NCG. This can pose a challenge when the datasets of known driver genes are used in both development and validation steps of a tool, e.g., using the CGC for training a model and subsequently validating it on the NCG will lead to overfitting. Method developers should therefore be careful in the study design to ensure that training and validation sets are not overlapping and thereby prevent overfitting during development and validation of driver predictors.

Tokheim *et al.* established an evaluation framework to assess and compare the performance of different methods, circumventing the use of a gold standard. The authors established that a method that can discover many of the driver genes from CGC and those predicted by other methods fulfill two criteria of good performance [81]. This framework was for example used by Parvandeh *et al.* [88]. Combining the results from multiple methods would aid the discovery and evaluation of critical driver genes [68, 81].

It appears that while most studies generally utilize one tool, newer studies are beginning to incorporate two or three approaches [87], a strategy whose popularity likely will increase in the future.

Most driver gene prediction tools are cohort-level methods, meaning they predict driver genes across patient cohorts. However, these methods often fail to identify rare driver genes present in a minority of the patients. Additionally, these predictions are challenging to use in the clinic as they predict driver genes for the whole cohort instead of for individual patients. For these reasons, patient-level tools have recently emerged which are valuable for tailored clinical strategies [46, 60, 61].

DRIVER MUTATIONS PREDICTION

Oncogenesis originates from a few key driver mutations [89, 90], and identification of specific driver mutations could inform

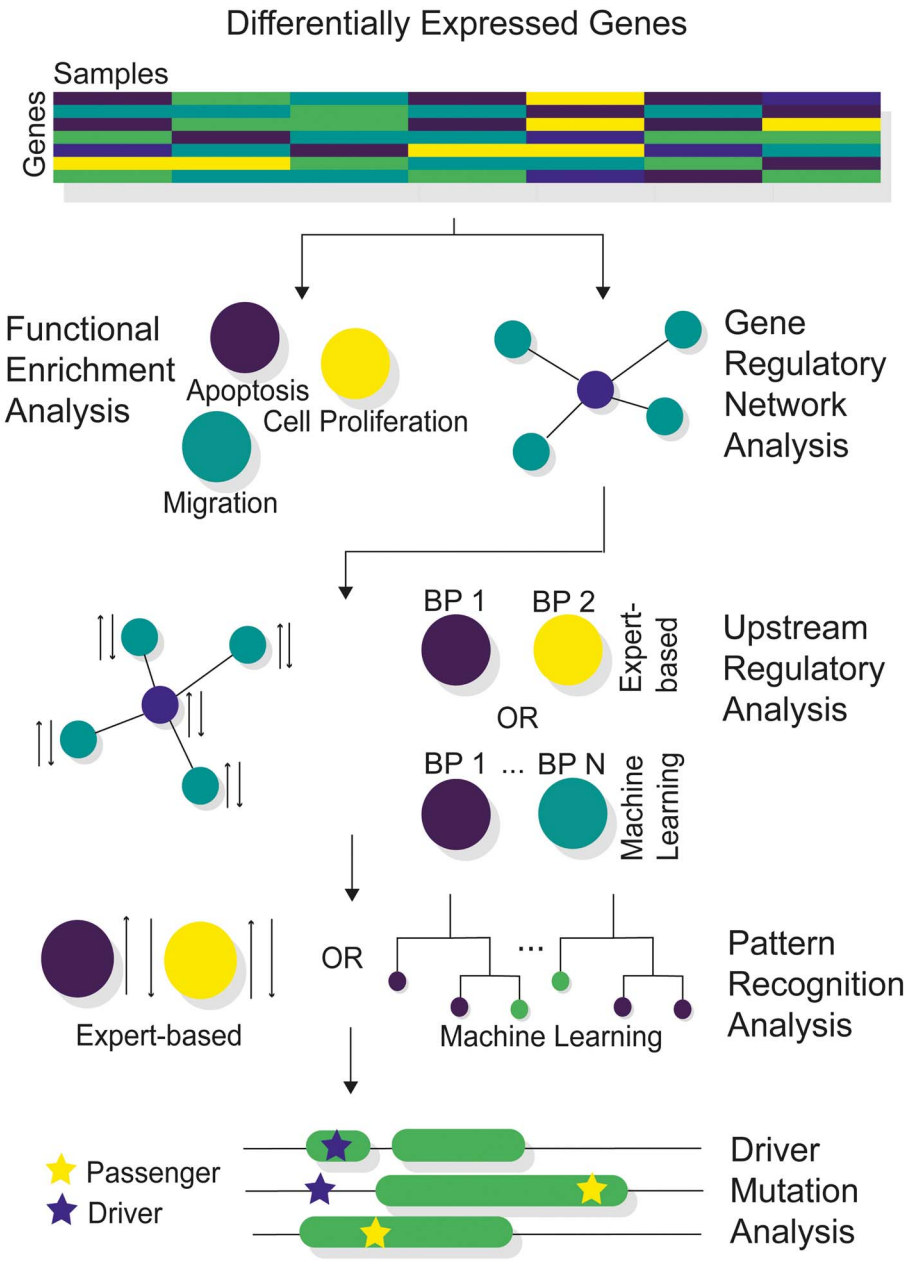

Figure 3. The Moonlight framework for driver gene prediction. Moonlight uses a set of DEGs as input. First, a functional enrichment analysis is carried out to find which of Moonlight's 101 BPs are overrepresented among the DEGs. Then, a gene regulatory network analysis models how the DEGs are connected with each other through mutual information. Following this step, Moonlight diverges into an expert-based and a machine learning approach. In the next step, an upstream regulatory analysis, the expert-based approach examines the effect of DEGs on user-selected BPs whereas the machine learning approach examines this on all of Moonlight's BPs. Subsequently, putative tumor suppressors and oncogenes collectively called oncogenic mediators are predicted through a pattern recognition analysis using either patterns (the expert-based approach) or a random forest classifier (the machine learning approach). Finally, a driver mutation analysis analyzes mutations in the cancer patient cohort and categorizes these into drivers and passengers. Those oncogenic mediators containing at least one driver mutation are retained as driver genes.

further studies of disruptive gene function [91, 92]. A plethora of tools aims to predict driver mutations where some are termed variant effect predictors (VEPs). Even tools developed to discover pathogenic mutations across diseases are routinely used to identify cancer driver mutations. The first published tools in the field relied on frequency measurements, identifying mutations that appear significantly more than in a background model [71]. While these frequency measurements indicate an evolutionary association, the approach lacks sensitivity toward healthy human variation [93] and misclassifies 26–38% of known pathogenic mutations [94]. To overcome this, most current tools rely on a combination of genomic features, evolutionary features, physicochemical properties and protein domains [95, 96]. To understand and discuss these tools, we here present a subset as categorized based on the computational approach (Figure 4).

Frequency measurements and evolutionary conservation

Some methods are primarily based on frequency measurements and are often considered functional impact scores based on conservation. The aim is to measure the frequency of a mutation compared to a background model. For example, Protein Variation

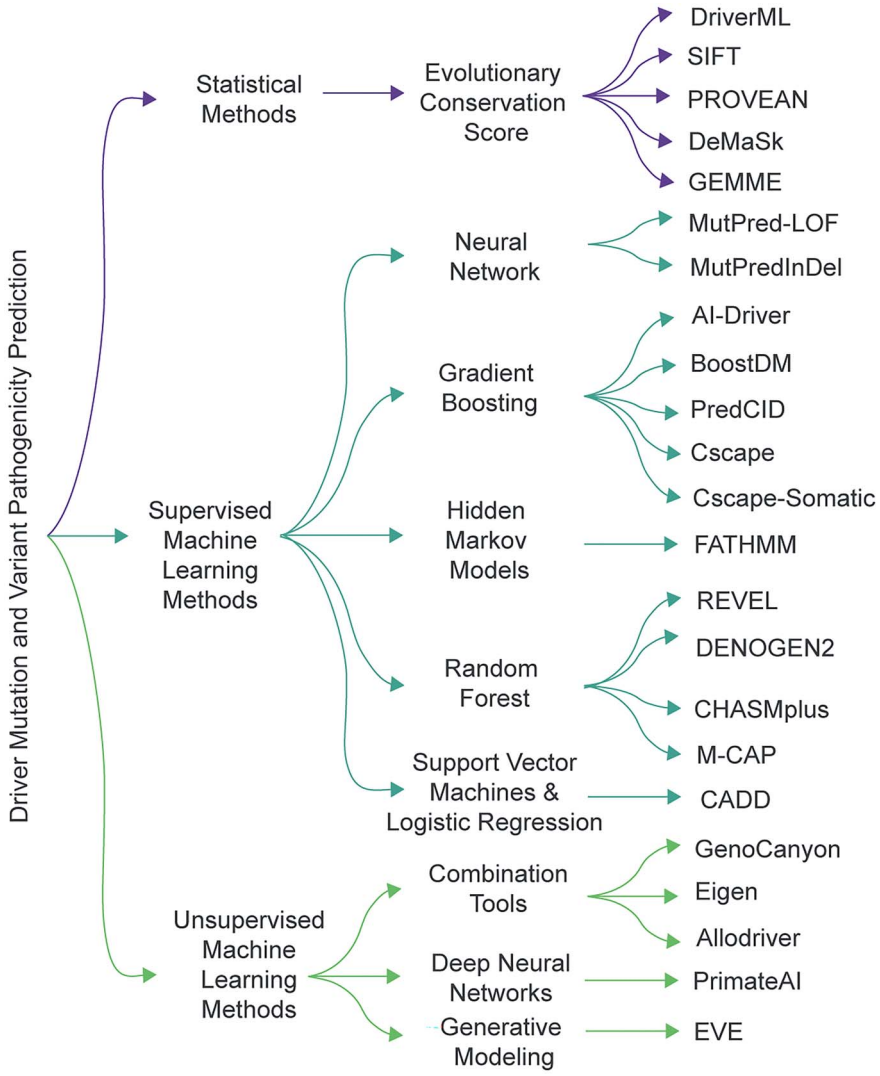

Figure 4. Overview of driver mutation and variant pathogenicity prediction tools. All driver mutation and variant pathogenicity prediction tools discussed in this review are listed to the right. These tools are divided into three main categories based on the underlying computational architecture. These three categories are subcategorized.

Effect Analyzer [97] predicts the functional effect of alterations using pairwise alignment. Sorting Tolerant From Intolerant [98, 99] predicts if any amino acid substitution affects protein function based on an evolutionary conservation score using multiple sequence alignment. Such statistical tools are limited by the annotation methods and conservation metrics that vary between them and seldom account for allele specificity or functional information [100]. The performance of the tools is evaluated by comparison to known pathogenic mutations, which may introduce ascertainment bias due to a limited number of available annotations. Each tool utilizes its own scoring measure, which is not compatible with other scores or any physical measurement [100]. Furthermore, the accuracy of the background model for driver mutations can be a limitation due to tumor heterogeneity and the recurrence of mutations, which are often not included in background mutation rates [101]. DriverML relies on frequency measurements and further applies the scores to create mutational clusters that account for different mutation types [57]. Alternative approaches to the study of pathogenic mutations relying on evolutionary information are GEMME [102] and DeMaSk [103]. GEMME is based on evolutionary-informed conservation where the quantification of the impact of variants considers global similarities and is not limited to a single amino acid change. DeMaSk predicts the variant impact using the fitness impact that is estimated based on a linear model of data from deep mutational scans and an asymmetrical amino acid substitution matrix.

Machine learning

Both supervised and unsupervised machine learning models have been developed for driver mutation prediction. In terms of supervised methods, a range of tree-based methods exists, showing how a combination of evolutionarily based tools is beneficial and that these can be combined with other types of descriptors. A well-known random forest model is REVEL [104], a method for predicting the pathogenicity of missense variants. REVEL is routinely used as the benchmarking standard for new driver mutation prediction tools. In a comparison among different tools, REVEL featured consistently better performance, illustrating how there may be an optimum in the number of features or ensemble methodologies [105]. Another random forest is M-CAP [94] that classifies clinical pathogenic mutations utilizing a broader range of input features.

CHASMplus [21] aims at scoring the oncogenic impact of a driver missense mutation by cancer type. The tool is built on many

features of which the most impactful feature is HotMaps [106], indicating how structural information could be utilized to find and assess driver mutations. Another example of utilizing structural information is DEOGEN2 [107], aiming to predict deleterious single nucleotide variants.

Within the category of supervised machine learning tools, one should also mention the FATHMM [108] family based on hidden Markov models and CADD based on SVMs and logistic regression [109].

Another approach to supervised machine learning is gradient boosting. The advantage of boosting algorithms is the transparency of the calculations and their resilience to overfitting. However, some disadvantages include heightened sensitivity to outliers and historical issues with upscaling. For example, AI-Driver predicts the driver status of somatic missense mutations [92]. BoostDM considers gene-tumor combinations, and in particular, cancer mutations available for those specific combinations, which constitute a driver mutation set [24].

Other approaches include PredCID, which classifies driver frameshift indels [110], and CScape-somatic, which discriminates between somatic driver and passenger mutations [95, 111].

One way to overcome the upscaling issues is to use neural networks implemented in for example the MutPred framework. MutPred-LOF [112] predicts pathogenic and tolerated loss-of-function variants (frameshift and stop codons) while MutPredInDel [113] covers pathogenic and tolerated non-frameshifting insertion/deletion variants. The main advantages of neural networks are their computational affordability and ability to find connections flexibly. However, they are a black box that may challenge the interpretation of the output, especially when the labeled training data may not be fully annotated from a biological standpoint.

In summary, all the supervised machine learning methods listed here combine pre-existing tools and occasionally physicochemical and structural information to achieve better classification performance. The limitation of using supervised learning to classify mutations as drivers or passengers is the fact that the tools are limited by the data on which they are built—a good performing supervised learning model requires a balanced dataset and considerations of the class imbalance problem.

A comparatively smaller range of tools based on unsupervised machine learning methods exists. This lack of abundance is connected to the assumption that differentiating driver and passenger mutations is a binary classification problem. When introducing unsupervised machine learning, the answer is unlikely to be binary. For instance, when considering clustering algorithms, there is no guarantee that all or most mutations within the cluster are actually driver mutations. A classic unsupervised machine learning approach is used in PrimateAI [114] that identifies pathogenic missense mutations. PrimateAI approaches the driver mutation question by creating an artificial neural network that uses sequences and bases all calculations on sequence homology to other species.

One non-neural network example is Eigen [115]. Here, an eigenvector weighted score is calculated for the identification and annotation of disease variants.

A different approach has been taken by Allodriver [116] relying entirely on structural data. Allodriver aims to identify and prioritize driver mutations based on known allosteric and orthosteric sites derived from three-dimensional structures. The model is constructed as a combination of random forest and feed-forward neural network models on an oversampled dataset of driver mutations.

Another example of a combined tool is GenoCanyon [117] that relies on the posterior probability of conserved regions and biochemical annotations to annotate each position in the human genome.

Lastly, EVE [118] relies on protein sequence and its evolution to estimate protein variant pathogenicity. EVE is an unsupervised generative model which relies on encoding multiple sequence analysis, allowing the computation of an evolutionary index that is used as input for the Gaussian mixture model separating benign and pathogenic variants.

Unsupervised methods, in general, rely on more basic biological annotations such as structure, which could potentially remove some biases that may have been included in previous tools for the prediction of driver mutations.

Summary and recommendations

Even though a universally recognized benchmarking dataset for driver mutation prediction methods is not available, some benchmarking efforts have been conducted. One benchmarking study was conducted by Livesey *et al.* who compared the performance of 46 predictors towards a dataset of deep mutational scanning data from 31 experiments [119]. The studied mutations were not associated with any single disease. They found the predictive performance to vary considerably between tools. They suggest that these differences stem from known limitations in predictive models: (i) re-use of training data for assessing performance for example due to database overlap and (ii) ascertainment bias, performing well on heavily studied genes. Another benchmarking study was done by Chen *et al.* [120] who aimed to compare the performance of different algorithms on five datasets. They found that tools specifically designed to deal with cancer performed better than disease-agnostic tools. However, it is also evident that the performance of the tools changed significantly depending on the selected dataset. This illustrates some fundamental limitations of these tools, which are connected to either: (i) a limited amount of known and annotated driver mutations, which restricts training set size, or (ii) the scope of the tool. The challenge is especially pronounced for rare drivers [12]. Rather than designing a comprehensive tool to predict any driver mutation, it may be preferable to design a set of tools that may predict specific effects of these mutations such as regulatory impact or protein loss- or gain-of-function. Rogers *et al.* suggest that the fundamental idea of dividing mutations into drivers and passengers is a constraint as it would be more prudent to ask why a mutation is a driver rather than if [17]. Another constraint is the annotation of identified driver mutations. A mutation could be annotated as a driver; however, this could be the result of insufficient or incomplete data if the said mutation is only a driver in one context but a passenger in another. Tools trained on this annotation may introduce underlying biases. A solution could be achieved by stratifying the prediction not only down to a cancer type, but at a cancer subtype level and including information regarding the genes in which the mutation dwells and the associated cellular pathways and regulatory networks [120].

Future directions within this field should focus on cancer subtype-specific resources, including information regarding the placement of the mutation in a gene context, or applying new types of co-evolutionary analysis and understanding the molecular mechanism related to the mutations as applied in other contexts handling mutations, e.g., protein engineering [121–124]. A possibility to solve some of these issues is to move towards structure-based frameworks to understand cancer mutations.

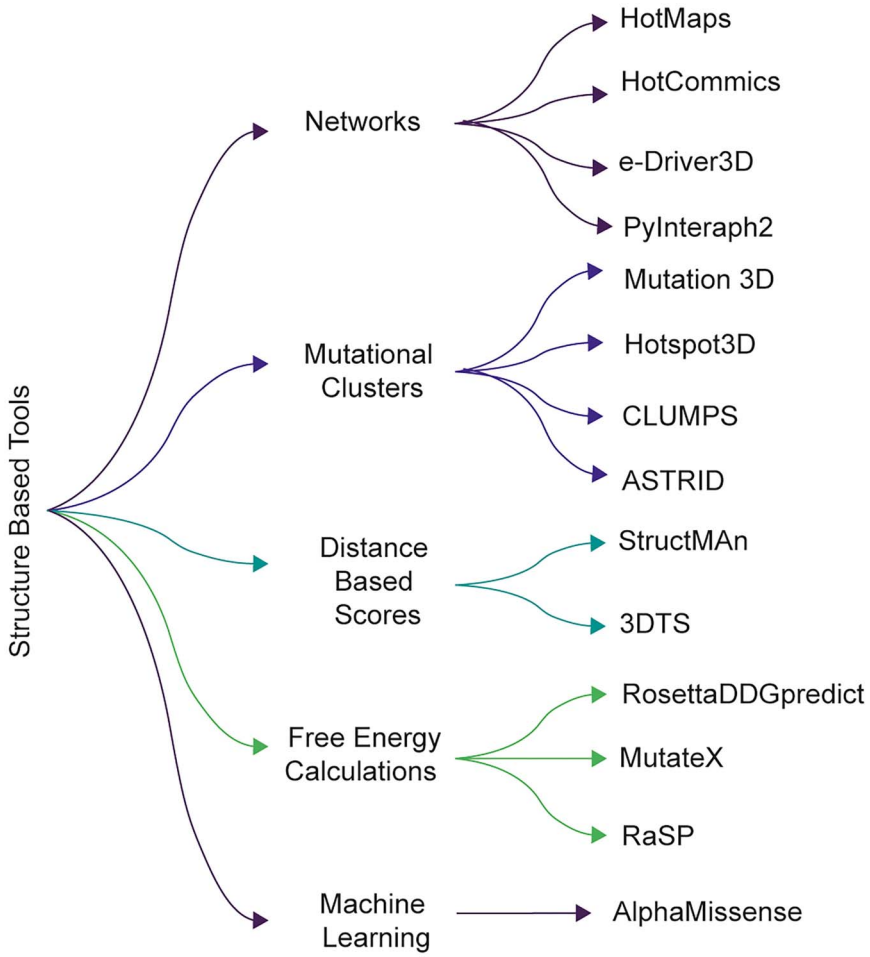

Figure 5. Overview of structural analysis tools. All structural analysis tools discussed in this review are listed to the right. These tools are divided into five categories based on the underlying computational architecture.

STRUCTURE-BASED TOOLS TO STUDY THE EFFECTS OF CANCER MUTATIONS

One way to study alterations driving cancer in the coding regions is to model these in the protein structure. For mutations with unknown consequences, this can be useful to predict their effect and understand what changes they impart to the protein structure. For mutations with known effects, structural studies help derive a mechanistic explanation of their consequence. This is achieved by identifying patterns in the structural changes in terms of stability disruption and changes in conformation at functional sites [125]. Functional sites are often difficult to identify from the sequence alone, and mutations far from each other in the protein sequence may be adjacent in the structure. After folding, seemingly distant substitutions can impact functionality by employing orthosteric and even allosteric effects [116]. We here present a subset of the current analysis tools as categorized based on the computational approach (Figure 5).

Protein structure networks

Protein structures can be modeled as a network. The advantage of describing a protein structure as a network is a great simplification of the protein structure and the possibility to use graph-theoretical methods. HotMaps aims to identify missense mutational hotspots considering the three-dimensional structure [106]. Other methods identify positions in the protein structure particularly enriched with cancer mutations and identify spatially close groups of such residues as components in a residue interaction graph. An example in this category is FASMIC that was developed on a combination of experimental structures and homology models [126]. HotCommics is developed for somatic cancer missense mutations and was developed on experimentally solved structures [127]. The e-Driver3D method, contrarily, is cancer-specific and is based on protein–protein interaction networks and aims to analyze the mutation distribution of specific interaction interfaces [128]. The potential of this approach was consolidated by Cheng *et al.* who applied protein–protein networks to cancer mutations to shed light on the high mutational burden in protein–protein interfaces by use of networks, which they experimentally validated [125]. PyInteraph2 applies graph theory to ensembles of structures with a broader scope [129]. This tool has been used to study mutational effects, design variants for proteins, study biomolecular complexes or understand the effect of post-translational modifications (PTMs). Here, the creation of the network relies on intra- and intermolecular interactions between residues or side-chain contacts. The nodes are the residues, and the edges are the non-bonded interactions weighted based on the occurrence of the interaction in the ensemble.

Future steps in structural graph networks may include graph attention neural networks [130]. Graph attention neural networks capture predictive features weighted by co-evolution. This aligns with the driver mutation predictive development and deployment of sophisticated machine learning approaches.

Mutational cluster identification

Several structural tools aim to identify and assess mutation clusters. Clusters are often defined based on a specific measure of spatial distance between residues in the structure, yet there is no consensus on how such a measure is defined. For example, Mutation 3D aims to identify clusters of somatic missense mutations, based on alpha-carbon linkage, to find driver genes [131]. Two tools aiming to identify mutational clusters and interpret their functional role are Hotspot3D [132] and CLUMPS [133] that both rely on somatic cancer mutations and protein structures with high sequence identity to the chosen target. In Hotspot3D, the clusters are found based on the minimum distance between atoms in pairs of residues. The functional impact is subsequently annotated using ClinVar data [134]. Hotspot3D predictions have been experimentally validated, and the tool has been utilized within other applications, e.g. FASMIC [126]. In CLUMPS, clusters are found based on proximity as the clusters are calculated based on centroids as pairwise Euclidean distance. The identification of cancer mutations relies on testing if a cluster is more mutated than expected by chance. Furthermore, it is a possibility that one residue is present in multiple clusters. The method was also experimentally validated [135]. Finally, ASTRID evaluates the three-dimensional spatial patterns of human germline and somatic variation, which is not necessarily cancer-specific. ASTRID relies on both neutral germline variants, disease-causing germline variants, and recurrent somatic variants. This approach quantifies spatial distributions of protein-coding mutations to create clusters [136].

The application of cluster-based methodologies identifying hotspots can be a valuable tool to differentiate drivers and passengers. The identification of mutational clusters by means of clustering algorithms holds great promise due to their efficiency in identifying alterations at functional sites. However, the methodologies have two main limitations: (i) they rely on a background model whose shortcomings have previously been described, and (ii) they apply various distance measurements and thresholds, making comparisons difficult.

Scores based on analysis of three-dimensional structures

The impact of mutations on the protein product of a gene can also be evaluated using methods that account for the three-dimensional (3D) structure of the protein. One example is StructMAn [137], which scores the change of conformation in terms of distance to a ligand. The aim is to classify nonsynonymous single nucleotide variants as either Quasi-WT (no apparent change of interactions), quasi-null (complete loss of interactions) or edgetic (specific loss of some interactions) based on set score thresholds. An entirely different approach to a score is used in 3DTS (three-dimensional tolerance score) which scores missense mutations to describe functional constraints [138]. 3DTS takes variants with mutations within 5 Å from a feature defined in Uniprot and assesses the probability that the three-dimensional site is intolerant to a missense mutation. The main limitation of the distance-based scores is the reliance on single amino acid substitutions to drive a change in conformation resulting in a functional impact. These scores, however, may create a more comprehensive picture if incorporated with the mutational clusters. With the advent of AlphaFold [31] different approaches to using these are also incorporated, notably with AlphaMissense [139], a tool using machine learning to score missense mutations, illustrating that we may see a new generation of structure-based VEPs.

Free energy calculations

An alternative to the tools described above is structural evaluations of alterations using free energy calculations [7], a task that has been recently streamlined thanks to high throughput workflow for mutational scans in silico such as MutateX [140] and RosettaDDGPrediction [8]. The overall idea is to estimate the change of energy upon one or more mutations to assess the functional impact in terms of stability and interaction with protein partners. They are commonly applied to understand the impact of structural alterations. The impact is expressed as changes in Gibbs free energy, and the interpretation of these values depends on, for example, the expected accuracy of the prediction and whether the tool assesses stability changes or changes in binding energy. The advantage of these tools is that the changes in Gibbs' free energy are physical measurements that can be compared to real-world experimental data rather than an arbitrary score. These methods still suffer from limitations, partially because of the limited conformational sampling they can carry out, because of their intrinsic bias due to their unbalanced training set and of their sensitivity concerning the used structure. These free energy calculations were built based on experimentally found structures but also apply to homology models [141] and de-novo models such as Alphafold2 [142]. Another possible scenario is to use deep learning models to predict free energy changes. An example is RaSP [143], a protein stability prediction tool capable of conducting saturation mutagenesis. RaSP is created as a deep learning counterpart to Rosetta predictions with similar performance.

Summary and recommendations

Much like the identification and assessment of driver genes and driver mutations, the area of structural assessments is fast developing. Particularly with the advent of Alphafold2, the limited available structures may be a constraint of the past. This may pave the way for a range of new tools employing different machine learning algorithms and even more sophisticated clustering. To ease the barrier of entry to structural studies, further development may include options to visualize and analyze missense variants in a protein sequence and structural space for a set of variants found in the general population including protein–protein interactions, PTMs, and functional features as seen in MISCAST [144]. Improved availability of structures will foster new ideas and applications from other computational science fields. To further strengthen the structural framework aiming to funnel -omics data into a tangible protein outcome, novel tools could be inspired by the driver mutation ensemble methodology. The aim of such an ensemble could be to describe the structural alterations resulting from the mutational clusters rather than individual assessments. This could lead to understanding the collective impact on function and binding, both using energy changes and distance scoring as well as considering the relative impacts of the mutations in the clusters such as compensation or synergistic effects. Creating an annotation and classification system for variants may lower the barrier of entry in interpreting protein structure studies as well as help prioritize experimental validation [9]. Yet, one limitation that should be addressed in future tools is the current modeling of proteins in a static conformation. Both the experimentally solved structures and the predicted structures rely on a single conformation of a protein.

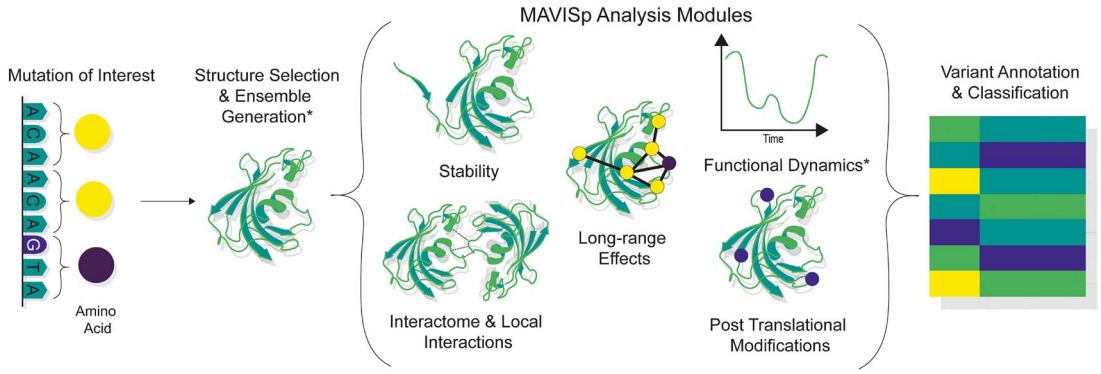

Figure 6. MAVISp framework for structural analysis. MAVISp is a module-dependent framework to study structural alterations. The curators of MAVISp take mutations as an input and study these using the applicable modules starting with the structure selection and potentially ensemble generation. The mutations are modeled in the structure and analyzed for their impact on stability, local interactions, long-range effects, PTMs and, if an ensemble was generated, functional dynamics. All of these analyses provide information for variant assessment. *Modules are part of the ensemble-mode of MAVISp.

A protein's functionality is most likely dependent on its interaction with ligands and macromolecules, rendering them dynamic entities. One way to mitigate this limitation is to use molecular dynamic simulations to generate a representative ensemble of protein conformations in solution [9]. During the last couple of years, our group has been developing the first steps towards a more generalized structure-based framework to assess cancer variants on proteins involved in cancer hallmarks [9, 145–148]. These studies created the foundation for a Multi-layered assessment of Variants by Structure for proteins framework, MAVISp [149]. MAVISp is an integrative module-based framework building a reproducible protocol to systematically study structural alterations. MAVISp creates an end-to-end framework that can be applied to a single three-dimensional structure and its complexes or an ensemble of structures. The framework initially identifies known cancer mutations via COSMIC [150] and cBioPortal [151] but can also be supplied on a specific set of mutations from a particular research project [152]. MAVISp then identifies possible structures and known interactors of the protein within the structure selection and interactome modules. The modules are stability, estimating the mutational impact on protein stability compared to the wildtype measured in changes in Gibbs free energy, local interactions, estimating the mutational impact on the interaction, and long-range effect, by estimating the allosteric free energy resulting from the perturbation of any residue and finally PTMs. The point is to gain a thorough and rounded understanding of a protein. Additionally, MAVISp can handle data from structural ensembles and use them for the mentioned modules and add functional dynamics to the toolkit overcoming the limitations of using only a representative structure for a protein of interest (Figure 6).

DISCUSSION

The field of predictive tools in cancer genomics has made significant progress, but several challenges still need to be overcome. One question is whether we can trust the predictions made by the tools. While it is possible to predict pathogenic mutations, the mechanism of action behind genetic alterations is not fully understood [153, 154]. Another limitation is the trade-off between annotating every possible mutation and scoping the tools so narrowly that only one context is investigated. Computational approaches serve as a vital starting point to elucidate the underlying mechanisms of cancer biology. These methods greatly reduce the number of genes and mutations to be tested experimentally and thereby decrease the experimental load, cost, and time associated with such tasks. Despite the importance of prior computational studies, wet lab validations should always follow to confirm the findings.

Not all tools are continuously maintained which can prevent steady use. This illustrates another challenge—the need to update the tool based on biological knowledge and computational resources available. Since many tools integrate existing biological knowledge from various databases, as seen for example with training sets in machine learning methods, these tools may benefit from routine revisions and updates. While such knowledge is continuously updated, failing to incorporate novel findings and data dynamically in the tools may potentially decrease their performance and longevity. Moreover, the tools require maintenance in terms of software. Most of the tools are available as command-line programs and Python and R packages. These platforms are also regularly updated, and, consequently, the user may experience problems when installing and applying the tools due to incompatibility between the user's programming environment and the software requirements of the tool under which it was built. Besides maintenance, solid documentation is an important factor for successful usage. We compiled an overview of all computational tools discussed here which includes availability of source code, documentation, and test data, coding language, dependencies, type of input data, compatible operating systems and whether they were installable and runnable in November 2023 based on available documentation. It is evident that not all tools are regularly updated, offer accessible source code, or demonstrate well-structured documentation which hinders installation and usage of the tools (Table S4, Document S1). Indeed, 14 out of the 74 described tools did not supply source code or the code was inaccessible. Of the tools with source code, 82.5% were installable; however, in only 43.6% of the cases, the documentation was sufficient to run the tool on provided test data (Document S1).

Collectively, the predictive tools presented here allow for analyses of the structural impact of the predicted driver alterations. Hence, we propose studying these fields collectively rather than individually, with each field serving as the input to the next, thereby promoting integration and avoiding research silos. Such a workflow would include initial driver gene prediction, subsequent driver mutation prediction in the driver genes, and finally, a structural assessment of the impact of mutations (Figure 7). This could in the future enable a transition from omics analyses to drug discovery, repurposing and development illustrating

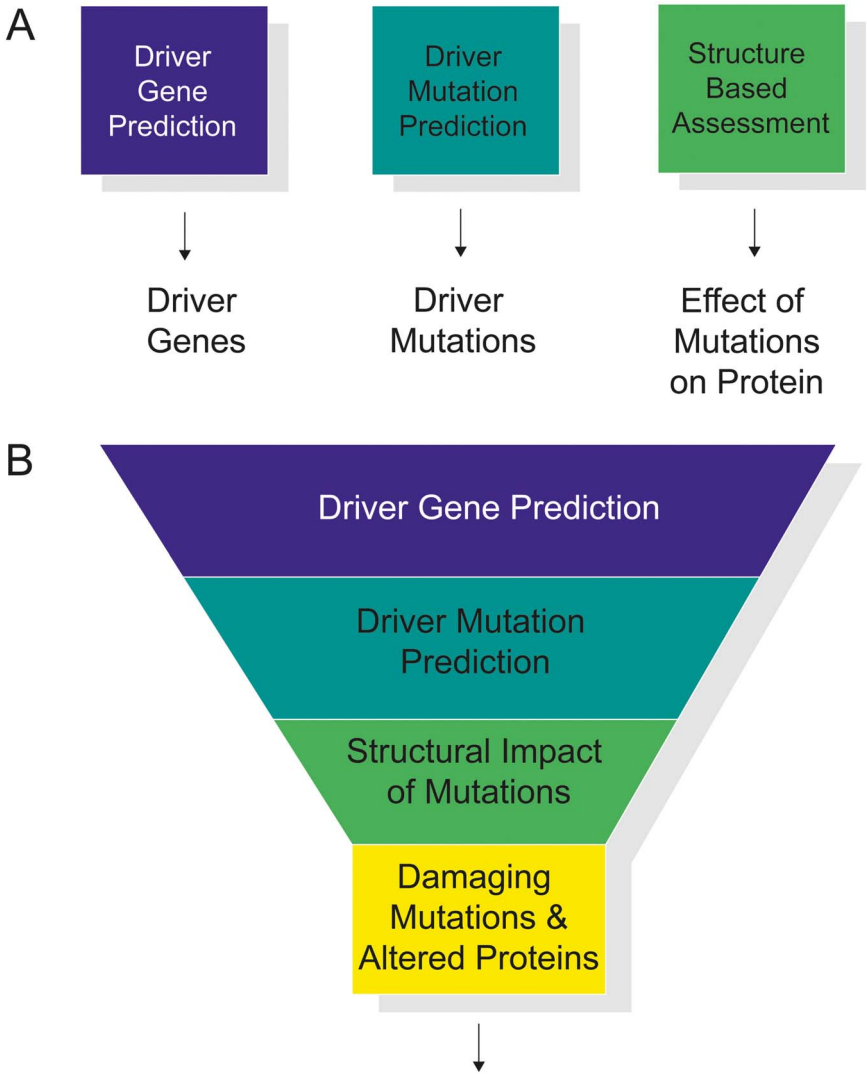

Figure 7. Current and suggested future workflow. (**A**) Current research within driver gene prediction, driver mutation prediction and structural assessment of mutations is characterized by tools within these fields working in silos. Such an approach risks overlooking potential synergies. (**B**) To overcome some of the challenges by this approach, a suggested protocol for the future development of these fields is illustrated. This protocol includes the collective study of the driver gene prediction, driver mutation and structural assessment of mutation fields, with each field serving as the input to the next, creating a funnel approach. Such a novel workflow would include initial driver gene prediction, selection of driver genes, driver mutation prediction in the predicted driver genes, and, finally, choice of mutations to be studied structurally to assess the impact of these mutations on protein function and stability. This will result in a set of damaged proteins for further exploration.

the profound clinical implications of the discovery of cancer drivers and their alterations. Integrative bioinformatic analyses are a crucial link between the extensive data generated by next-generation sequencing technologies and clinical decision-making. The discovery of molecular alterations has led to targeted therapies, improving patient survival [155]. For example, is imatinib—a kinase inhibitor targeting *BCR-ABL* translocations—routinely used for treating chronic myeloid leukemia [156].

Key Points

- We critically reviewed a range of computational prediction methods, discussing the reliance on the quality of the used dataset, the underlying algorithm and the output of the tools, aiming to shed light on potential biases.
- While aiming to answer some of the same biological questions, the research fields of cancer driver gene prediction, driver mutation prediction and structure-based analyses of proteins currently work in silos.
- The discovery of driver mutations in driver genes and the effect of mutations on proteins in cancer allows for targeting disrupted pathways and thereby reversal of the cancer hallmarks. In this review, we highlight the potential of integrating different fields aiming to gain a more profound understanding of cancer alterations.

SUPPLEMENTARY DATA

Supplementary data are available online at http://bib.oxfordjournals.org/.

FUNDING

EP group has been supported by Hartmanns Fond (R241-A33877), Leo Foundation (LF17006), Carlsberg Foundation Distinguished Fellowship (CF18-0314), NovoNordisk Fonden Bioscience and Basic Biomedicine (NNF20OC0065262). EP group is also part of the Center of Excellence in Autophagy, Recycling and Disease (CARD), funded by the Danish National Research Foundation (DNRF-125).

CONTRIBUTIONS

Conceived and designed the review contents: K.D., M.N., E.P., Visualization: K.D., M.N., Wrote the review: M.N., K.D., M.T., A.S., E.P.

DATA AVAILABILITY

No new data were generated or analyzed in support of this research.

REFERENCES

1. Hanahan D. Hallmarks of cancer: new dimensions. *Cancer Discov* 2022;**12**:31–46.
2. Hanahan D, Weinberg RA. The hallmarks of cancer review evolve progressively from normalcy via a series of pre. *Cell* 2000;**100**:57–70.
3. Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. *Cell* 2011;**144**:646–74.
4. Hahn WC, Weinberg RA. Modelling the molecular circuitry of cancer. *Nat Rev Cancer* 2002;**2**:331–41.
5. Vogelstein B, Papadopoulos N, Velculescu VE, *et al.* Cancer genome landscapes. *Science* 1979;**339**(339):1546–58.
6. Gerasimavicius L, Liu X, Marsh JA. Identification of pathogenic missense mutations using protein stability predictors. *Sci Rep* 202010:1 2020;**10**:1–10.
7. Stein A, Fowler DM, Hartmann-Petersen R, Lindorff-Larsen K. Biophysical and mechanistic models for disease-causing protein variants. *Trends Biochem Sci* 2019;**44**:575–88.
8. Sora V, Otamendi Laspiur A, Degn K, *et al.* RosettaDDGPrediction for high-throughput mutational scans: from stability to binding. *Protein Sci* 2023;**32**:e4527.
9. Fas BA, Maiani E, Sora V, *et al.* The conformational and mutational landscape of the ubiquitin-like marker for autophagosome formation in cancer. *Autophagy* 2021;**17**:2818–41.
10. Morash M, Mitchell H, Beltran H, *et al.* The role of next-generation sequencing in precision medicine: a review of outcomes in oncology. *J Pers Med* 2018;**8**(3):30.
11. Hardwick SA, Deveson IW, Mercer TR. Reference standards for next-generation sequencing. *Nat Rev Genet* 2017;**18**:473–84.
12. Ostroverkhova D, Przytycka TM, Panchenko AR. Cancer driver mutations: predictions and reality. *Trends Mol Med* 2023;**29**:554–66.
13. Ata SK, Wu M, Fang Y, *et al.* Recent advances in network-based methods for disease gene prediction. *Brief Bioinform* 2021;**22**:1–15.
14. Zhao F, Zheng L, Goncearenco A, *et al.* Computational approaches to prioritize cancer driver missense mutations. *Int J Mol Sci* 2018;**19**:2113.
15. David A, Sternberg MJE. Protein structure-based evaluation of missense variants: resources, challenges and future directions. *Curr Opin Struct Biol* 2023;**80**:102600, 1–8.
16. Paiva V D A, Gomes I D S, Monteiro CR, *et al.* Protein structural bioinformatics: an overview. *Comput Biol Med* 2022;**147**:105695.
17. Rogers MF, Gaunt TR, Campbell C. Prediction of driver variants in the cancer genome via machine learning methodologies. *Brief Bioinform* 2021;**22**:bbaa250.
18. Shea A, Bartz J, Zhang L, Dong X. Predicting mutational function using machine learning. *Mutat Res/Rev Mutat Res* 2023;**791**:108457, 1–7.
19. Gonzalez-Perez A, Deu-Pons J, Lopez-Bigas N. Improving the prediction of the functional impact of cancer mutations by baseline tolerance transformation. *Genome Med* 2012;**4**:89.
20. Learned K, Durbin A, Currie R, *et al.* Barriers to accessing public cancer genomic data. *Sci Data* 2019;**6**:98.
21. Tokheim C, Karchin R. CHASMplus reveals the scope of somatic missense mutations driving human cancers. *Cell Syst* 2019;**9**:9–23.e8.
22. Rentzsch P, Schubach M, Shendure J, Kircher M. CADD-Splice—improving genome-wide variant effect prediction using deep learning-derived splice scores. *Genome Med* 2021;**13**:31.
23. Shihab HA, Rogers MF, Gough J, *et al.* An integrative approach to predicting the functional effects of non-coding and coding sequence variation. *Bioinformatics* 2015;**31**:1536–43.
24. Muiños F, Martínez-Jiménez F, Pich O, *et al.* In silico saturation mutagenesis of cancer genes. *Nature* 2021;**596**:428–32.
25. Lawrence MS, Stojanov P, Mermel CH, *et al.* Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature* 2014;**505**:495–501.
26. Andrades R, Recamonde-Mendoza M. Machine learning methods for prediction of cancer driver genes: a survey paper. *Brief Bioinform* 2022;**23**:1–19.
27. Ganini C, Amelio I, Bertolo R, *et al.* Global mapping of cancers: the cancer genome atlas and beyond. *Mol Oncol* 2021;**15**:2823–40.
28. Zhang J, Bajari R, Andric D, *et al.* The International Cancer Genome Consortium data portal. *Nat Biotechnol* 2019;**37**:367–9.
29. Chakravarty D, Gao J, Phillips S, *et al.* OncoKB: a precision oncology Knowledge Base. *JCO Precis Oncol* 2017;**2017**:1–16.
30. Berman HM, Westbrook J, Feng Z, *et al.* The Protein Data Bank. *Nucleic Acids Res* 2000;**28**:235–42.
31. Jumper J, Evans R, Pritzel A, *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* 2021;**596**:583–9.
32. Varadi M, Anyango S, Deshpande M, *et al.* AlphaFold protein structure database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res* 2022;**50**:D439–44.
33. Lin Z, Akin H, Rao R, *et al.* Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* 1979;**379**(379):1123–30.
34. Martínez-Jiménez F, Muiños F, Sentís I, *et al.* A compendium of mutational cancer driver genes. *Nat Rev Cancer* 2020;**20**:555–72.
35. Bailey MH, Tokheim C, Porta-Pardo E, *et al.* Comprehensive characterization of cancer driver genes and mutations. *Cell* 2018;**173**:371–385.e18.
36. Zhang P, Itan Y. Biological network approaches and applications in rare disease studies. *Genes (Basel)* 2019;**10**:797.
37. Pham VVH, Liu L, Bracken CP, *et al.* CBNA: a control theory based method for identifying coding and non-coding cancer drivers. *PLoS Comput Biol* 2019;**15**:e1007538, 1–23.
38. Wei PJ, Wu FX, Xia J, *et al.* Prioritizing cancer genes based on an improved Random Walk method. *Front Genet* 2020;**11**:11.
39. Akhavan-Safar M, Teimourpour B, Kargari M. GenHITS: a network science approach to driver gene detection in human regulatory network using gene's influence evaluation. *J Biomed Inform* 2021;**114**:103661.

40. Akhavan-Safar M, Teimourpour B. KatzDriver: a network based method to cancer causal genes discovery in gene regulatory network. *Biosystems* 2021;**201**:104326.
41. Rahimi M, Teimourpour B, Marashi SA. Cancer driver gene discovery in transcriptional regulatory networks using influence maximization approach. *Comput Biol Med* 2019;**114**: 103362.
42. Pham VVH, Liu L, Bracken CP, *et al.* DriverGroup: a novel method for identifying driver gene groups. *Bioinformatics* 2020;**36**:I583–91.
43. Elliott K, Larsson E. Non-coding driver mutations in human cancer. *Nat Rev Cancer* 2021;**21**:500–9.
44. Champion M, Brennan K, Croonenborghs T, *et al.* Module analysis captures pancancer genetically and epigenetically deregulated cancer driver genes for smoking and antiviral response. *EBioMedicine* 2018;**27**:156–66.
45. Wei PJ, Zhang D, Li HT, *et al.* DriverFinder: a gene length-based network method to identify cancer driver genes. *Complexity* 2017;**2017**:1–10.
46. Dinstag G, Shamir R. PRODIGY: personalized prioritization of driver genes. *Bioinformatics* 2020;**36**:1831–9.
47. Wei PJ, Zhang D, Xia J, Zheng CH. LNDriver: identifying driver genes by integrating mutation and expression data based on gene-gene interaction network. *BMC Bioinformatics* 2016;**17**: 221–30.
48. Bertrand D, Chng KR, Sherbaf FG, *et al.* Patient-specific driver gene prediction and risk assessment through integrated network analysis of cancer omics profiles. *Nucleic Acids Res* 2015;**43**:e44, 1–13.
49. Zhang W, Wang SL. A novel method for identifying the potential cancer driver genes based on molecular data integration. *Biochem Genet* 2020;**58**:16–39.
50. Zhang D, Bin Y. DriverSubNet: a novel algorithm for identifying cancer driver genes by subnetwork enrichment analysis. *Front Genet* 2021;**11**:11.
51. Song J, Peng W, Wang F. A random walk-based method to identify driver genes by integrating the subcellular localization and variation frequency into bipartite graph. *BMC Bioinform* 2019;**20**:238.
52. Ahmed R, Baali I, Erten C, *et al.* MEXCOwalk: mutual exclusion and coverage based random walk to identify cancer modules. *Bioinformatics* 2020;**36**:872–9.
53. Peng W, Yi S, Dai W, Wang J. Identifying and ranking potential cancer drivers using representation learning on attributed network. *Methods* 2021;**192**:13–24.
54. Lu X, Wang X, Ding L, *et al.* FrDriver: a functional region driver identification for protein sequence. *IEEE/ACM Trans Comput Biol Bioinform* 2021;**18**:1773–83.
55. Pan H, Renaud L, Chaligne R, *et al.* Discovery of candidate DNA methylation cancer driver genes. *Cancer Discov* 2021;**11**: 2266–81.
56. Li A, Chapuy B, Varelas X, *et al.* Identification of candidate cancer drivers by integrative Epi-DNA and Gene Expression (iEDGE) data analysis. *Sci Rep* 2019;**9**, 1–12.
57. Han Y, Yang J, Qian X, *et al.* DriverML: a machine learning algorithm for identifying driver genes in cancer sequencing studies. *Nucleic Acids Res* 2019;**47**:e45, 1–12.
58. Guo WF, Zhang SW, Zeng T, *et al.* Network control principles for identifying personalized driver genes in cancer. *Brief Bioinform* 2020;**21**:1641–62.
59. Gonzalez-Perez A, Lopez-Bigas N. Functional impact bias reveals cancer drivers. *Nucleic Acids Res* 2012;**40**:e169, 1–10.
60. Nulsen J, Misetic H, Yau C, Ciccarelli FD. Pan-cancer detection of driver genes at the single-patient resolution. *Genome Med* 2021;**13**:1–14.
61. Ülgen E, Sezerman OU. driveR: a novel method for prioritizing cancer driver genes using somatic genomics data. *BMC Bioinform* 2021;**22**:263.
62. Peng W, Tang Q, Dai W, Chen T. Improving cancer driver gene identification using multi-task learning on graph convolutional network. *Brief Bioinform* 2022;**23**:1–12.
63. Gu H, Xu X, Qin P, Wang J. FI-net: identification of cancer driver genes by using functional impact prediction neural network. *Front Genet* 2020;**11**:11.
64. Luo P, Ding Y, Lei X, Wu FX. DeepDriver: predicting cancer driver genes based on somatic mutations using deep convolutional neural networks. *Front Genet* 2019;**10**:10.
65. Zeng Z, Mao C, Vo A, *et al.* Deep learning for cancer type classification and driver gene identification. *BMC Bioinform* 2021;**22**:491.
66. Gumpinger AC, Lage K, Horn H, Borgwardt K. Prediction of cancer driver genes through network-based moment propagation of mutation scores. *Bioinformatics* 2020;**36**:I508–15.
67. Collier O, Stoven V, Vert JP. LOTUS: a single- and multitask machine learning algorithm for the prediction of cancer driver genes. *PLoS Comput Biol* 2019;**15**:e1007381, 1–27.
68. Shi X, Teng H, Shi L, *et al.* Comprehensive evaluation of computational methods for predicting cancer driver genes. *Brief Bioinform* 2022;**23**:1–14.
69. Arnedo-Pac C, Mularoni L, Muiños F, *et al.* OncodriveCLUSTL: a sequence-based clustering method to identify cancer drivers. *Bioinformatics* 2019;**35**:4788–90.
70. Zhu H, Uusküla-Reimand L, Isaev K, *et al.* Candidate cancer driver mutations in distal regulatory elements and long-range chromatin interaction networks. *Mol Cell* 2020;**77**:1307–1321.e10.
71. Korthauer KD, Kendziorski C. MADGiC: a model-based approach for identifying driver genes in cancer. *Bioinformatics* 2015;**31**:1526–35.
72. Bokhari Y, Alhareeri A, Arodz T. QuaDMutNetEx: a method for detecting cancer driver genes with low mutation frequency. *BMC Bioinform* 2020;**21**:122.
73. Hou Y, Gao B, Li G, Su Z. MaxMIF: a new method for identifying cancer driver genes through effective data integration. *Adv Sci* 2018;**5**:1–9.
74. Zapata L, Susak H, Drechsel O, *et al.* Signatures of positive selection reveal a universal role of chromatin modifiers as cancer driver genes. *Sci Rep* 2017;**7**:13124.
75. Datta N, Chakraborty S, Basu M, Ghosh MK. Tumor suppressors having oncogenic functions: the double agents. *Cell* 2020;**10**: 1–26.
76. Stepanenko AA, Vassetzky YS, Kavsan VM. Antagonistic functional duality of cancer genes. *Gene* 2013;**529**:199–207.
77. Croce CM. Oncogenes and cancer. *N Engl J Med* 2008;**358**:502–11.
78. Wang LH, Wu CF, Rajasekaran N, Shin YK. Loss of tumor suppressor gene function in human cancer: an overview. *Cell Physiol Biochem* 2018;**51**:2647–93.
79. Shen L, Shi Q, Wang W. Double agents: genes with both oncogenic and tumor-suppressor functions. *Oncogenesis* 2018;**7**:25.
80. Chandrashekar P, Ahmadinejad N, Wang J, *et al.* Somatic selection distinguishes oncogenes and tumor suppressor genes. *Bioinformatics* 2020;**36**:1712–7.
81. Tokheim CJ, Papadopoulos N, Kinzler KW, *et al.* Evaluating the evaluation of cancer driver genes. *Proc Natl Acad Sci USA* 2016;**113**:14330–5.

82. Lyu J, Li JJ, Su J, *et al.* DORGE: discovery of oncogenes and tumoR suppressor genes using genetic and epigenetic features. *Sci Adv* 2020;**6**:1–17.
83. Colaprico A, Olsen C, Bailey MH, *et al.* Interpreting pathways to discover cancer driver genes with Moonlight. *Nat Commun* 2020;**11**:69.
84. Nourbakhsh M, Saksager A, Tom N, *et al.* A workflow to study mechanistic indicators for driver gene prediction with Moonlight. *Brief Bioinform* 2023;**24**:1–13.
85. Kobren SN, Chazelle B, Singh M. PertInInt: an integrative, analytical approach to rapidly uncover cancer driver genes with perturbed interactions and functionalities. *Cell Syst* 2020;**11**:63–74.e7.
86. Sondka Z, Bamford S, Cole CG, *et al.* The COSMIC cancer gene census: describing genetic dysfunction across all human cancers. *Nat Rev Cancer* 2018;**18**:696–705.
87. Repana D, Nulsen J, Dressler L, *et al.* The network of cancer genes (NCG): a comprehensive catalogue of known and candidate cancer genes from cancer sequencing screens. *Genome Biol* 2019;**20**:1.
88. Parvandeh S, Donehower LA, Katsonis P, *et al.* EPIMUTESTR: a nearest neighbor machine learning approach to predict cancer driver genes from the evolutionary action of coding variants. *Nucleic Acids Res* 2022;**50**:e70–0.
89. Darbyshire M, du Toit Z, Rogers MF, *et al.* Estimating the frequency of single point driver mutations across common solid tumours. *Sci Rep* 2019;**9**:13452.
90. Martincorena I, Raine KM, Gerstung M, *et al.* Universal patterns of selection in cancer and somatic tissues. *Cell* 2017;**171**:1029–1041.e21.
91. Dong C, Wei P, Jian X, *et al.* Comparison and integration of deleteriousness prediction methods for nonsynonymous SNVs in whole exome sequencing studies. *Hum Mol Genet* 2015;**24**:2125–37.
92. Wang H, Wang T, Zhao X, *et al.* AI-Driver: an ensemble method for identifying driver mutations in personal cancer genomes. *NAR Genom Bioinform* 2020;**2**:lqaa084.
93. Davydov EV, Goode DL, Sirota M, *et al.* Identifying a high fraction of the human genome to be under selective constraint using GERP++. *PLoS Comput Biol* 2010;**6**:e1001025, 1–13.
94. Jagadeesh KA, Wenger AM, Berger MJ, *et al.* M-CAP eliminates a majority of variants of uncertain significance in clinical exomes at high sensitivity. *Nat Genet* 2016;**48**:1581–6.
95. Rogers MF, Gaunt TR, Campbell C. CScape-somatic: distinguishing driver and passenger point mutations in the cancer genome. *Bioinformatics* 2020;**36**:3637–44.
96. Mao Y, Chen H, Liang H, *et al.* CanDrA: cancer-specific driver missense mutation annotation with optimized features. *PloS One* 2013;**8**:e77945, 1–8.
97. Choi Y, Chan AP. PROVEAN web server: a tool to predict the functional effect of amino acid substitutions and indels. *Bioinformatics* 2015;**31**:2745–7.
98. Kumar P, Henikoff S, Ng PC. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat Protoc* 2009;**4**:1073–81.
99. Vaser R, Adusumalli S, Leng SN, *et al.* SIFT missense predictions for genomes. *Nat Protoc* 2016;**11**:1–9.
100. Kircher M, Witten DM, Jain P, *et al.* A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet* 2014;**46**:310–5.
101. Brown A-L, Li M, Goncearenco A, Panchenko AR. Finding driver mutations in cancer: elucidating the role of background mutational processes. *PLoS Comput Biol* 2019;**15**:e1006981, 1–25.
102. Laine E, Karami Y, Carbone A. GEMME: a simple and fast global Epistatic model predicting mutational effects. *Mol Biol Evol* 2019;**36**:2604–19.
103. Munro D, Singh M. DeMaSk: a deep mutational scanning substitution matrix and its use for variant impact prediction. *Bioinformatics* 2020;**36**:5322–9.
104. Ioannidis NM, Rothstein JH, Pejaver V, *et al.* REVEL: an ensemble method for predicting the pathogenicity of rare missense variants. *Am J Hum Genet* 2016;**99**:877–85.
105. Li J, Zhao T, Zhang Y, *et al.* Performance evaluation of pathogenicity-computation methods for missense variants. *Nucleic Acids Res* 2018;**46**:7793–804.
106. Tokheim C, Bhattacharya R, Niknafs N, *et al.* Exome-scale discovery of hotspot mutation regions in human cancer using 3D protein structure. *Cancer Res* 2016;**76**:3719–31.
107. Raimondi D, Tanyalcin I, FertCrossed JSD, *et al.* DEOGEN2: prediction and interactive visualization of single amino acid variant deleteriousness in human proteins. *Nucleic Acids Res* 2017;**45**:W201–6.
108. Rogers MF, Shihab HA, Mort M, *et al.* FATHMM-XF: accurate prediction of pathogenic point mutations via extended features. *Bioinformatics* 2018;**34**:511–3.
109. Rentzsch P, Witten D, Cooper GM, *et al.* CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Res* 2019;**47**:D886–94.
110. Yue Z, Chu X, Xia J. PredCID: prediction of driver frameshift indels in human cancer. *Brief Bioinform* 2021;**22**:1–9.
111. Rogers MF, Shihab HA, Gaunt TR, Campbell C. CScape: a tool for predicting oncogenic single-point mutations in the cancer genome. *Sci Rep* 2017;**7**:1–10.
112. Pagel KA, Pejaver V, Lin GN, *et al.* When loss-of-function is loss of function: assessing mutational signatures and impact of loss-of-function genetic variants. *Bioinformatics* 2017;**33**:i389–98.
113. Pagel KA, Antaki D, Lian A, *et al.* Pathogenicity and functional impact of non-frameshifting insertion/deletion variation in the human genome. *PLoS Comput Biol* 2019;**15**:e1007112, 1–21.
114. Sundaram L, Gao H, Padigepati SR, *et al.* Predicting the clinical impact of human mutation with deep neural networks. *Nat Genet* 2018;**50**:1161–70.
115. Ionita-Laza I, Mccallum K, Xu B, *et al.* A spectral approach integrating functional genomic annotations for coding and noncoding variants. *Nat Genet* 2016;**48**:214–20.
116. Song K, Li Q, Gao W, *et al.* AlloDriver: a method for the identification and analysis of cancer driver targets. *Nucleic Acids Res* 2019;**47**:W315–21.
117. Lu Q, Hu Y, Sun J, *et al.* A statistical framework to predict functional non-coding regions in the human genome through integrated analysis of annotation data. *Sci Rep* 2015;**5**:10576.
118. Frazer J, Notin P, Dias M, *et al.* Disease variant prediction with deep generative models of evolutionary data. *Nature* 2021;**599**:91–5.
119. Livesey BJ, Marsh JA. Using deep mutational scanning to benchmark variant effect predictors and identify disease mutations. *Mol Syst Biol* 2020;**16**:e9380, 1–12.
120. Chen H, Li J, Wang Y, *et al.* Comprehensive assessment of computational algorithms in predicting cancer driver mutations. *Genome Biol* 2020;**21**:43.
121. Jiang L, Guo F, Tang J, *et al.* SBSA: an online service for somatic binding sequence annotation. *Nucleic Acids Res* 2022;**50**:e4.
122. Luo Y, Jiang G, Yu T, *et al.* ECNet is an evolutionary context-integrated deep learning framework for protein engineering. *Nat Commun* 2021;**12**:5743.

123. Li M, Kang L, Xiong Y, *et al.* SESNet: sequence-structure feature-integrated deep learning method for data-efficient protein engineering. *J Chem* 2023;**15**:12.
124. Kim E, Novak LC, Lin C, *et al.* Dynamic rewiring of biological activity across genotype and lineage revealed by context-dependent functional interactions. *Genome Biol* 2022;**23**:140.
125. Cheng F, Zhao J, Wang Y, *et al.* Comprehensive characterization of protein–protein interactions perturbed by disease mutations. *Nat Genet* 2021;**53**:342–53.
126. Ng PKS, Li J, Jeong KJ, *et al.* Systematic functional annotation of somatic mutations in cancer. *Cancer Cell* 2018;**33**:450–462.e10.
127. Kumar S, Clarke D, Gerstein MB. Leveraging protein dynamics to identify cancer mutational hotspots using 3D structures. *Proc Natl Acad Sci USA* 2019;**116**:18962–70.
128. Porta-Pardo E, Garcia-Alonso L, Hrabe T, *et al.* A Pan-cancer catalogue of cancer driver protein interaction interfaces. *PLoS Comput Biol* 2015;**11**:e1004518, 1–18.
129. Sora V, Tiberti M, Beltrame L, *et al.* PyInteraph2 and PyInKnife2 to Analyze Networks in Protein Structural Ensembles. *J Chem Inf Model* 2023;**63**(14):4237–45.
130. Zhang H, Xu MS, Fan X, *et al.* Predicting functional effect of missense variants using graph attention neural networks. *Nat Mach Intell* 2022;**4**:1017–28.
131. Meyer MJ, Lapcevic R, Romero AE, *et al.* mutation3D: cancer gene prediction through atomic clustering of coding variants in the structural proteome. *Hum Mutat* 2016;**37**:447–56.
132. Niu B, Scott AD, Sengupta S, *et al.* Protein-structure-guided discovery of functional mutations across 19 cancer types. *Nat Genet* 2016;**48**:827–37.
133. Kamburov A, Lawrence MS, Polak P, *et al.* Comprehensive assessment of cancer missense mutation clustering in protein structures. *Proc Natl Acad Sci* 2015;**112**:E5486–95.
134. Landrum MJ, Lee JM, Riley GR, *et al.* ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res* 2014;**42**:D980–5.
135. Gao J, Chang MT, Johnsen HC, *et al.* 3D clusters of somatic mutations in cancer reveal numerous rare mutations as functional targets. *Genome Med* 2017;**9**:4.
136. Sivley RM, Dou X, Meiler J, *et al.* Comprehensive analysis of constraint on the spatial distribution of missense variants in human protein structures. *Am J Hum Genet* 2018;**102**:415–26.
137. Gress A, Ramensky V, Büch J, *et al.* StructMAn: annotation of single-nucleotide polymorphisms in the structural context. *Nucleic Acids Res* 2016;**44**:W463–8.
138. Hicks M, Bartha I, di Iulio J, *et al.* Functional characterization of 3D protein structures informed by human genetic diversity. *Proc Natl Acad Sci* 2019;**116**:8960–5.
139. Cheng J, Novati G, Pan J, *et al.* Accurate proteome-wide missense variant effect prediction with AlphaMissense. *Science* 19792023;**381**:eadg7492.
140. Tiberti M, Terkelsen T, Degn K, *et al.* MutateX: an automated pipeline for in silico saturation mutagenesis of protein structures and structural ensembles. *Brief Bioinform* 2022;**23**:bbac074.
141. Valanciute A, Nygaard L, Zschach H, *et al.* Accurate protein stability predictions from homology models. *Comput Struct Biotechnol J* 2023;**21**:66–73.
142. Akdel M, Pires DE V, Pardo EP, *et al.* A structural biology community assessment of AlphaFold2 applications. *Nat Struct Mol Biol* 2022;**29**:1056–67.
143. Blaabjerg LM, Kassem MM, Good LL, *et al.* Rapid protein stability prediction using deep learning representations. *Elife* 2023;**12**:e82593, 1–19.
144. Iqbal S, Hoksza D, Pérez-Palma E, *et al.* MISCAST: MIssense variant to protein structure analysis web suite. *Nucleic Acids Res* 2021;**48**:W132–9.
145. Nygaard M, Terkelsen T, Olsen AV, *et al.* The mutational landscape of the oncogenic MZF1 SCAN domain in cancer. *Front Mol Biosci* 2016;**3**:1–18.
146. Kumar M, Papaleo E. A pan-cancer assessment of alterations of the kinase domain of ULK1, an upstream regulator of autophagy. *Sci Rep* 2020;**10**:14874.
147. Kønig SM, Rissler V, Terkelsen T, *et al.* Alterations of the interactome of Bcl-2 proteins in breast cancer at the transcriptional, mutational and structural level. *PLoS Comput Biol* 2019;**15**:e1007485, 1–28.
148. Degn K, Beltrame L, Dahl Hede F, *et al.* Cancer-related mutations with local or long-range effects on an allosteric loop of p53. *J Mol Biol* 2022;**434**:167663, 1–33.
149. Arnaudi M, Beltrame L, Degn K, *et al.* MAVISp: Multi-layered Assessment of VarIants by Structure for proteins. *bioRxiv* 2022; 1–12.
150. Tate JG, Bamford S, Jubb HC, *et al.* COSMIC: the catalogue of somatic mutations in cancer. *Nucleic Acids Res* 2019;**47**:D941–7.
151. Gao J, Aksoy BA, Dogrusoz U, *et al.* Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Sci Signal* 2013;**6**:pl1.
152. Tiberti M, Di Leo L, Vistesen MV, *et al.* The Cancermuts software package for the prioritization of missense cancer variants: a case study of AMBRA1 in melanoma. *Cell Death Dis* 2022;**13**: 872.
153. Høie MH, Cagiada M, Beck Frederiksen AH, *et al.* Predicting and interpreting large-scale mutagenesis data using analyses of protein stability and conservation. *Cell Rep* 2022;**38**:110207, 1–16.
154. Cagiada M, Johansson KE, Valanciute A, *et al.* Understanding the origins of loss of protein function by analyzing the effects of thousands of variants on activity and abundance. *Mol Biol Evol* 2021;**38**:3235–46.
155. Zhou Z, Li M. Targeted therapies for cancer. *BMC Med* 2022;**20**:90.
156. Rossari F, Minutolo F, Orciuolo E. Past, present, and future of Bcr-Abl inhibitors: from chemical development to clinical efficacy. *J Hematol Oncol* 2018;**11**:84.