Next-Generation Molecular Discovery: From Bottom-Up In Vivo and In Vitro Approaches to In Silico Top-Down Approaches for Therapeutics Neogenesis

Sophie E. Kenny [1], Fiach Antaw [1], Warwick J. Locke [2], Christopher B. Howard [1], Darren Korbie [1,*] and Matt Trau [1,3,*]

[1] Centre for Personalised Nanomedicine, Australian Institute for Bioengineering and Nanotechnology (AIBN), The University of Queensland, Corner of College and Cooper Roads (Bldg 75), Brisbane, QLD 4072, Australia; sophie.kenny@uq.edu.au (S.E.K.); f.antaw@uq.edu.au (F.A.); c.howard2@uq.edu.au (C.B.H.)

[2] Molecular Diagnostic Solutions, Health and Biosecurity, Commonwealth Scientific and Industrial Research Organisation, Building 101, Clunies Ross Street, Canberra, ACT 2601, Australia; warwick.locke@csiro.au

[3] School of Chemistry and Molecular Biosciences, The University of Queensland, Brisbane, QLD 4072, Australia

[*] Correspondence: d.korbie@uq.edu.au (D.K.); m.trau@uq.edu.au (M.T.)

Abstract: Protein and drug engineering comprises a major part of the medical and research industries, and yet approaches to discovering and understanding therapeutic molecular interactions in biological systems rely on trial and error. The general approach to molecular discovery involves screening large libraries of compounds, proteins, or antibodies, or in vivo antibody generation, which could be considered "bottom-up" approaches to therapeutic discovery. In these bottom-up approaches, a minimal amount is known about the therapeutics at the start of the process, but through meticulous and exhaustive laboratory work, the molecule is characterised in detail. In contrast, the advent of "big data" and access to extensive online databases and machine learning technologies offers promising new avenues to understanding molecular interactions. Artificial intelligence (AI) now has the potential to predict protein structure at an unprecedented accuracy using only the genetic sequence. This predictive approach to characterising molecular structure—when accompanied by high-quality experimental data for model training—has the capacity to invert the process of molecular discovery and characterisation. The process has potential to be transformed into a top-down approach, where new molecules can be designed directly based on the structure of a target and the desired function, rather than performing screening of large libraries of molecular variants. This paper will provide a brief evaluation of bottom-up approaches to discovering and characterising biological molecules and will discuss recent advances towards developing top-down approaches and the prospects of this.

Keywords: therapeutics; artificial intelligence; drug development; high-throughput library screening; antibody and peptide discovery; chemical libraries; phage display; protein folding prediction

1. Introduction

Understanding and predicting molecular interactions in living systems and designing new molecules to bind to, destroy, disrupt, or promote biological molecular pathways have long been driving forces behind much of molecular biology research and funding. In recent years, the therapeutics industry has boomed. In 2020, the FDA approved 53 drugs, the second highest approval rate ever [1], and the global pharmaceutical market was estimated to be worth at least USD 1.2 trillion in 2019 [2]. Whilst about 75% of the market is composed of small-molecule drugs, with only 20% biologics and 5% peptides, these markets are expanding, with the monoclonal antibody (mAb) market estimated to reach USD 130–200 billion in 2022 [3]. Interest in the field is massive, and rapidly growing.

The novel coronavirus pandemic has highlighted the urgency for efficient discovery and production of molecular therapeutics. For example, the first drug to be approved across major global regulatory authorities for the antiviral treatment of COVID-19 was remdesivir (see Table 1), where it took approximately one year for other antiviral drugs to receive authorisation. In Australia, it took about 6 months for remdesivir to be approved after the first wave of cases arrived. Additionally, it took over 1.5 years in Australia to see the approval of any other COVID-19 antiviral drug other than remdesivir, where no other antiviral treatment options were available for use until August 2021. In the United States, remdesivir was the only non-emergency use antiviral drug approved by the FDA until the authorisation of baricitinib on 19 November 2020 [4]. Even the use of human plasma—collected from individuals with high titres of anti-SARS-CoV-2 antibodies—was approved (23 August 2020) before baricitinib. The earlier authorisation of remdesivir globally for SARS-CoV-2 treatment in comparison to other antiviral drugs was only possible due to long-term testing on other RNA viruses such as Ebola and other coronaviruses in the years leading up to the pandemic [5,6]. Whilst remdesivir has been shown to be an effective anti-viral agent [7] (although this has been strongly debated [8]), the drug requires IV administration. This means it is not readily given to the general population for mild infections that do not require hospitalisation. In addition, as with most antiviral or antibacterial molecules, there is some possibility for viral resistance to arise [9], which means that having alternative treatments is optimal. The current slow process for therapeutic discovery and the inherent risk in having such a limited suite of anti-viral options prompts the question of how faster pipelines for molecular therapeutics development can be achieved.

Advances in molecular discovery have traditionally focussed on improving the efficiency and reliability of in vitro and in vivo discovery and synthesis techniques, approaches which can be considered as "bottom-up" techniques. In this review we specifically define "bottom-up" as methods for therapeutic molecular discovery that use animal antibody generation (in vivo), or the screening vast libraries of potential therapeutic molecules against therapeutic targets (in vitro), where molecules with the best biological response are selected (see Figure 1). These approaches are generally slow and arduous and in essence "find what sticks", or "find the needle in the haystack". Moreover, because little is known about the new therapeutic after discovery, there are significant resources spent in molecular characterisation and testing before the therapeutic is ready for clinical trials. There have been many approaches to reduce the time and resources spent in current discovery programs, yet there is not any technique that comprehensively turns this process around through a "top-down" approach, where a specific therapeutic molecule could be built from scratch based directly on the therapeutic target and desired outcome. Ideally, a top-down approach would eliminate the need to perform large library screening campaigns or perform in vivo antibody generation. It would allow for much higher confidence in target-binding specificity through a process of screening "self" molecules and therefore reduce the potential for side-effects in patients. Not only would this reduce risk for patients, but also investors, where financial investment is often one of the bottlenecks for taking a potentially life-saving drug to the market, and therefore deliverable to the patient. In this review we will explore the movement towards the top-down approach for molecular discovery, or rather design, and the potential to fast-track the pipeline for therapeutics from discovery to the clinic.

Table 1. List of major approved SARS-CoV-2 treatments by select regions as of 4th February 2022.
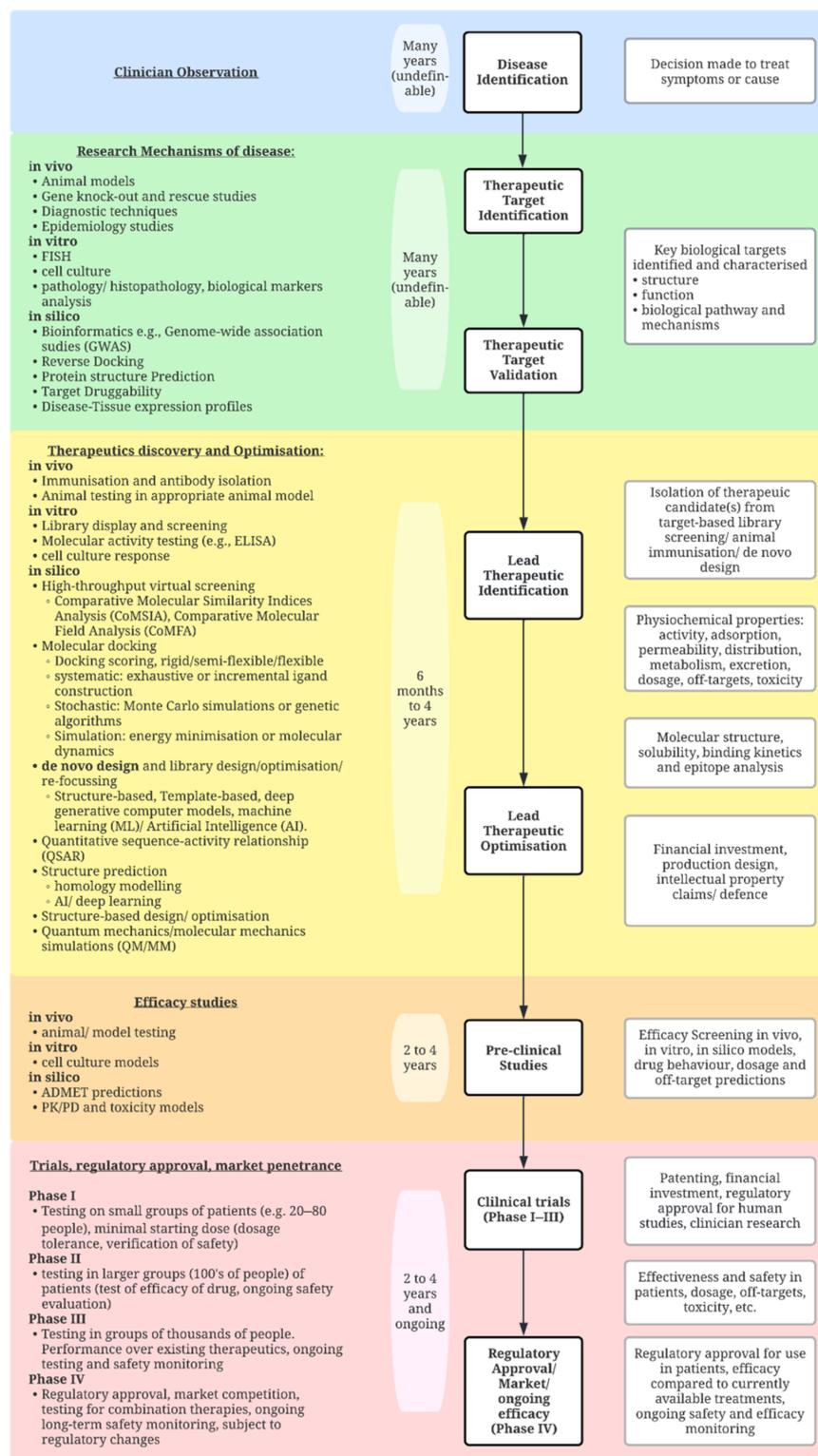
Drug Name	Date of TGA Provisional Approval in Australia (TGA)	Date of Emergency Use Authorisation in United States (FDA)	Date of Authorisation in European Union (EMA)	Date of Approval in Japan (PMDA)	Date of Authorisation in Canada (Health Canada)
Remdesivir	10 July 2020	5 February 2020 (eligible patients)	(conditional) 3 July 2020	7 May 2020	27 July 2020
Sotrovimab	20 August 2021	26 May 2021	17 December 2021	27 September 2021	30 July 2021
Casirivimab and Imdevimab	15 October 2021	21 November 2020 (removed until further notice from 24 January 2022 as of 4 February 2022)	12 November 2021	19 July 2021	9 June 2021
Tocilizumab	1 December 2021	24 June 2021	7 December 2021	-	-
Regdanvimab	6 December 2021	-	12 November 2021	-	-
Molnupiravir	18 January 2022	23 December 2021	-	-	-
Nirmatrelvir and ritonavir	18 January 2022	22 December 2021	28 January 2022	-	17 January 2022
Baricitinib	-	19 November 2020	-	-	-
Bamlanivimab and etesevimab	-	25 February 2021 (removed until further notice from 24 January 2022 as of 4 February 2022)	-	-	20 November 2020 (Bamlanivimab only)
Tixagevimab and cilgavimab	-	8 December 2021	-	-	-

Bottom-up

Molecular screening approach

Top-down

De novo generation approach
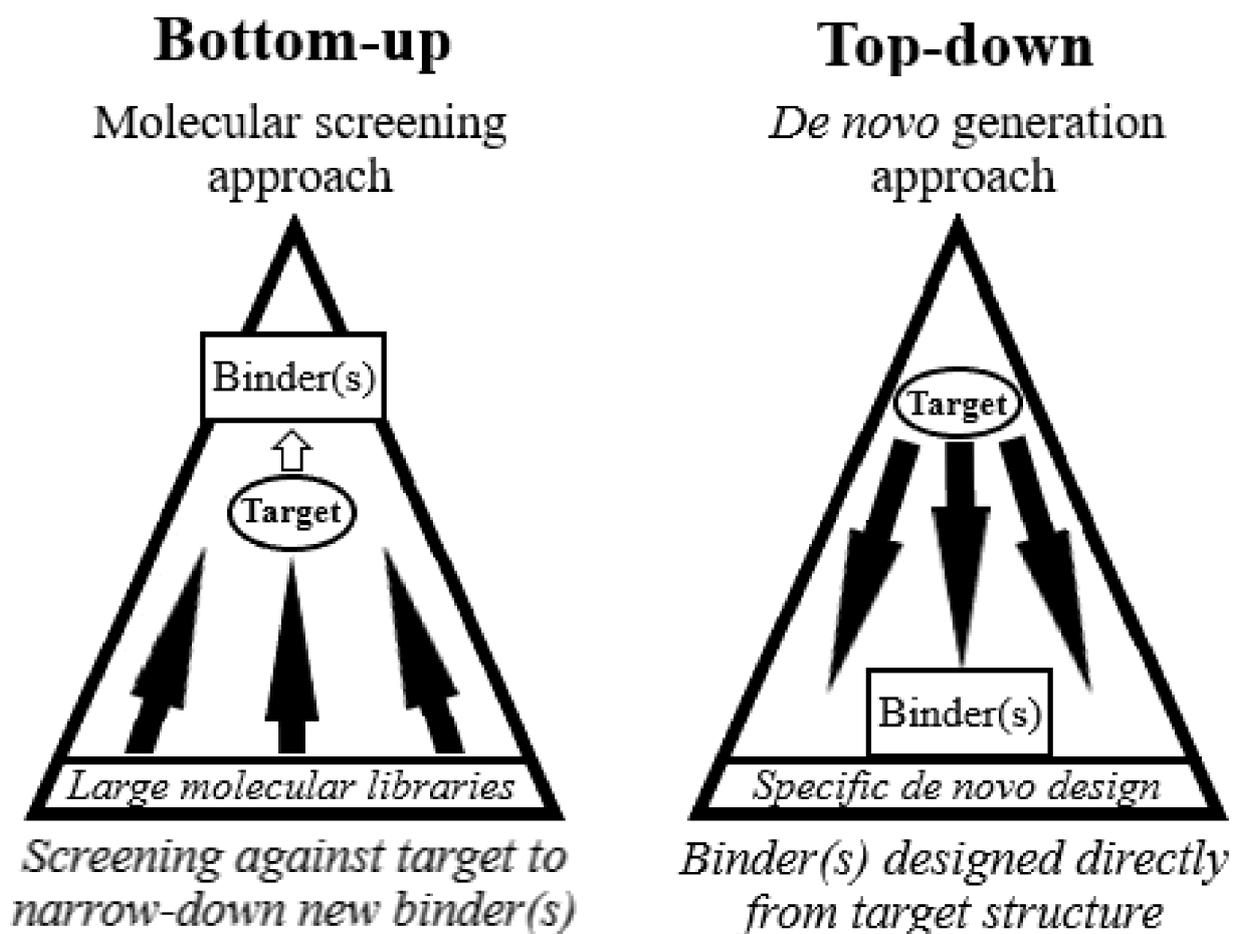
Figure 1. Bottom-up and top-down approaches to ex vivo molecular discovery. Bottom-up approaches generally involve screening broad libraries of candidate molecules against a target of interest and narrowing this down to a single best binder. The top-down approach is to design and generate new binders directly based on the structure of a target molecule.

2. Bottom-Up Approaches

Molecular discovery may be broadly grouped into in vivo, in vitro and in silico approaches. Most current approaches to molecular discovery may be considered bottom-up, particularly for in vivo or in vitro approaches, as generally little is known about the conformation of the final therapeutic before screening or antibody generation takes place. Whilst these techniques have not yet succeeded in flipping discovery towards top-down approaches, the field has produced a wealth of innovation and advancements that have enabled significant leaps in the field of therapeutic discovery. The therapeutics discovery pipeline has been outlined in Figure 2, with examples of some of the prominent approaches to each stage of the pipeline. Some of the most recent innovations that contribute to this pipeline will be outlined in the following sections.

Clinician Observation

Many years (undefinable)

Disease Identification

Decision made to treat symptoms or cause

Research Mechanisms of disease:
in vivo
• Animal models
• Gene knock-out and rescue studies
• Diagnostic techniques
• Epidemiology studies
in vitro
• FISH
• cell culture
• pathology/ histopathology, biological markers analysis
in silico
• Bioinformatics e.g., Genome-wide association sudies (GWAS)
• Reverse Docking
• Protein structure Prediction
• Target Druggability
• Disease-Tissue expression profiles

Therapeutic Target Identification

Many years (undefinable)

Therapeutic Target Validation

Key biological targets identified and characterised
• structure
• function
• biological pathway and mechanisms

Therapeutics discovery and Optimisation:
in vivo
• Immunisation and antibody isolation
• Animal testing in appropriate animal model
in vitro
• Library display and screening
• Molecular activity testing (e.g., ELISA)
• cell culture response
in silico
• High-throughput virtual screening
  ◦ Comparative Molecular Similarity Indices Analysis (CoMSIA), Comparative Molecular Field Analysis (CoMFA)
• Molecular docking
  ◦ Docking scoring, rigid/semi-flexible/flexible
  ◦ systematic: exhaustive or incremental igand construction
  ◦ Stochastic: Monte Carlo simulations or genetic algorithms
  ◦ Simulation: energy minimisation or molecular dynamics
• de novo design and library design/optimisation/ re-focussing
  ◦ Structure-based, Template-based, deep generative computer models, machine learning (ML)/ Artificial Intelligence (AI).
• Quantitative sequence-activity relationship (QSAR)
• Structure prediction
  ◦ homology modelling
  ◦ AI/ deep learning
• Structure-based design/ optimisation
• Quantum mechanics/molecular mechanics simulations (QM/MM)

Lead Therapeutic Identification

6 months to 4 years

Isolation of therapeuic candidate(s) from target-based library screening/ animal immunisation/ de novo design

Physiochemical properties: activity, adsorption, permeability, distribution, metabolism, excretion, dosage, off-targets, toxicity

Molecular structure, solubility, binding kinetics and epitope analysis

Lead Therapeutic Optimisation

Financial investment, production design, intellectual property claims/ defence

Efficacy studies
in vivo
• animal/ model testing
in vitro
• cell culture models
in silico
• ADMET predictions
• PK/PD and toxicity models

Pre-clinical Studies

2 to 4 years

Efficacy Screening in vivo, in vitro, in silico models, drug behaviour, dosage and off-target predictions

Trials, regulatory approval, market penetrance

Phase I
• Testing on small groups of patients (e.g. 20–80 people), minimal starting dose (dosage tolerance, verification of safety)
Phase II
• testing in larger groups (100's of people) of patients (test of efficacy of drug, ongoing safety evaluation)
Phase III
• Testing in groups of thousands of people. Performance over existing therapeutics, ongoing testing and safety monitoring
Phase IV
• Regulatory approval, market competition, testing for combination therapies, ongoing long-term safety monitoring, subject to regulatory changes

Clinical trials (Phase I–III)

2 to 4 years and ongoing

Patenting, financial investment, regulatory approval for human studies, clinician research

Effectiveness and safety in patients, dosage, off-targets, toxicity, etc.

Regulatory Approval/ Market/ ongoing efficacy (Phase IV)

Regulatory approval for use in patients, efficacy compared to currently available treatments, ongoing safety and efficacy monitoring

Figure 2. The current journey of therapeutics from discovery to the clinic (top to bottom), with prominent examples given of in vivo, in vitro and in silico approaches. A broad approximation of time taken for each stage of development is provided. (Created with lucidchart.com).

2.1. In Vivo Techniques

In vivo techniques have typically relied on the animal and human natural adaptive immune response, where antibodies are produced upon immunisation with a desired antigen [10–12] (see Figure 3). Many types of animals can be used, although mice, rats,

rabbits, guinea pigs, hamsters, chickens, and larger animals such as goats, horses (particularly for antivenoms), camelids and sheep are common. Antibody variation, such as the combinatorial diversification of V(D)J recombination, somatic hypermutation and class switching [10,13,14], enables a high immunoglobulin repertoire, which in humans can reach 10^{11} different molecular combinations [15]. Humanised rodent-produced mAbs were the first mAbs to be approved for therapeutics [10], where production required the immunisation of mice with an antigen of interest, followed by fusion of spleen cells with myeloma cells to form immortalised hybridomas [16]. Hybridomas are then screened and selected based on the desired affinity and specificity of the produced antibody.
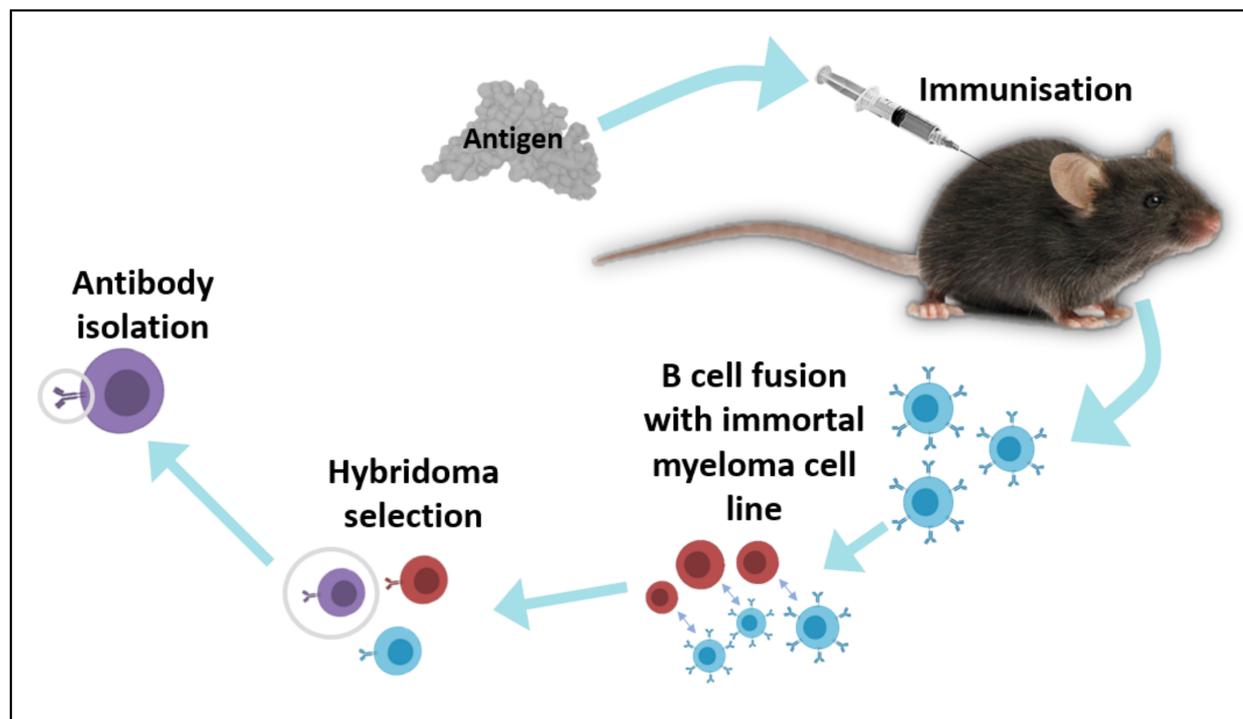
Figure 3. Traditional process of in vivo antibody isolation. First, animals are immunised with an antigen of interest. After an immune response has been mounted, B cells are isolated from the spleen of the animal and fused with an immortal myeloma cell line. Hybridomas are selected and screened for their activity against the antigen of interest. Target-binding clones are retained and used to obtain purified antibodies.

There are some limitations to in vivo techniques, but new innovations have sought to overcome these. First, mAbs produced in non-human models traditionally require humanisation in order to minimise undesirable responses from the human immune system [17]. However, the engineering of transgenic mice now allows the production of fully humanised antibodies directly from the animal [10,18,19], as well as techniques for the grafting of the complementarity determining regions of murine antibodies into human antibody frameworks [17]. Recently, the development of camelid nanobody or VHH (variable heavy chain domains of heavy chain antibodies)-producing mice enabled the production of enhanced avidity, neutralising binders for SARS-CoV-2 [20]. It was suggested that these could reduce likelihood of immune evasion by binding to conserved regions of the virus that are usually inaccessible to larger normal antibodies. There have also been human donor approaches to antibody generation. For example, B cells isolated from patients naturally infected with SARS-CoV-2 produced virus-neutralising activity in Syrian hamsters [21]. This enabled estimation of the most common patient antibody response in terms of targeted sites on the virus, particularly ACE2 binding site (competition studies), and isolation and characterisation of neutralising antibodies. However, likewise to antibodies produced in

animal models, significant testing in animals is generally still required for human-derived antibodies before progression to clinical trials would be possible.

While traditional in vivo methods benefit from the tremendous natural diversity of the animal immune response, they are not as high throughput as library screening techniques. It can take a significant amount of time (around 6 weeks [22]) for immunised animals to illicit an immune response and produce antigen-targeting B cells before target-binding properties can be screened. In vivo techniques are also generally expensive, requiring experienced laboratory technicians, specialised laboratory facilities, ethics approval and mouse sacrifice. Although, as antigen properties and conformation are affected by the conditions of the solution they reside in, they will be closest to their natural state when presented to the immune system in in vivo. This means that in vivo techniques allow for the generation of antibodies that will bind to antigens in the real biological setting, marking them as an attractive method for therapeutics discovery, despite time and handling issues.

2.2. In Vitro Techniques

In vitro approaches were initially aimed towards widening the range of possible therapeutic molecule types and making discovery more easily scalable than in vivo techniques. Techniques are generally geared towards mimicking the natural immune response in vitro and technological advances broadly involve widening the sampling of chemical space through high-throughput wet lab technologies. In addition to now having higher throughput than in vivo approaches, in vitro techniques can allow for screening against toxic compounds that may alternatively have been difficult or cumbersome to produce neutralising antibodies for in vivo approaches. For example, antivenoms can be produced in vitro, rather than using the traditional approach of horse immunisation. One approach used phage-displayed camelid nanobodies to produce neutralising activity against *Bothrops jararacussu* snake venom [23]. High costs of animal maintenance and difficulties in producing antivenom (stability of antivenom, obtaining enough venom from snakes) [24] make in vitro approaches attractive. However, in vitro discovery is still limited by the time taken for wet-lab experimental procedures to be carried out and by molecular library size, and all techniques still fall into the "bottom-up" approach classification.

Most in vitro techniques rely on the production of hugely diverse libraries of chemicals, peptides, antibodies, or antibody fragments, with the aim of isolating unique variants from the combinatory pool that will bind to a desired antigen. This means that larger library diversities are more advantageous as the likelihood of identifying a suitable binder scale proportionally. Libraries can be either labelled (linked to a descriptive tag such as DNA) or unlabelled. Labelled library types include well-established phage display [25,26], bacteria [27,28], yeast [29–31], mammalian cell display [32], and cell-free techniques such as ribosome [33], mRNA display [34,35] and DNA-linked chemical libraries [36]. Generally, most biologically generated libraries such as antibody or protein libraries are labelled, whereas unlabelled libraries are composed of individually stored molecules, for example high-throughput chemistry (HTC) [37,38] chemical libraries.

Initial molecular discovery techniques did not involve the use of labelled libraries, and instead relied on the synthesis and individual storage of unique compounds. A limitation in the size of available molecular libraries was, for a significant period, one of the major bottlenecks to new molecule discovery [39]. The advent of HTC techniques has now allowed for the creation of unlabelled chemical libraries of up to 3 million individually stored compounds. To screen these libraries in vitro, each molecule must be examined individually in separate reactions against targets to assess activity. Whilst this process is lengthy and inefficient if performed by hand, the speed of this has been significantly increased by high-throughput screening (HTS). This technology is the use of highly efficient automated robotic platforms to screen each library compound one-by-one against a target in high multiplicity multi-well plates [40–42]. HTS is essentially a brute-force solution to the lengthy wet lab procedure, but is still used frequently, particularly for chemical library screening.

Developed by George Smith in 1985, phage display has been one of the driving technologies for the discovery of new antibodies and peptides through labelled molecular library screening [26]. It is one of the original efficient methods for linking a molecular phenotype with its descriptive genotype, which allows for the identification and characterisations of a molecule isolated from a library of molecules during library screening. Phage display is the cloning of a desired variable DNA sequence, either of a peptide [25] or antibody fragment [26,43,44], into the positive-sense ssDNA genome of filamentous bacteriophage, where it is expressed as a protein fusion to one of the M13 bacteriophage coat proteins (generally pIII). These genetically linked molecular libraries can be divided into naïve, immune, semi-synthetic and fully synthetic libraries, depending on the origin of the displayed molecules. Non-immunised and immunised animals provide the source for naïve and immune libraries for antibody or antibody fragment display, respectively, whereas semi-synthetic libraries are sequences modified from animal or human donors. Fully synthetic libraries are those that display completely synthetically constructed complementary determining regions (CDRs) on known antibody frameworks, or synthetically constructed peptides, allowing for the generation of libraries with huge, randomised diversity [45,46]. These genotype-linked molecules can be screened through affinity enrichment, in a process called biopanning (see Figure 4), where molecules that bind to a target of interest are isolated.
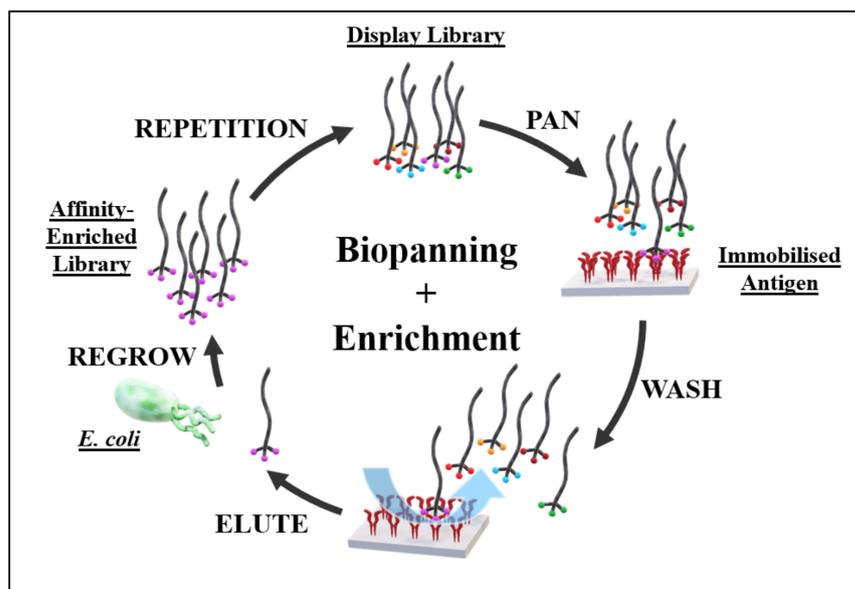
Figure 4. Traditional library biopanning and enrichment against a target. Broadly, the genetically encoded library is incubated against a surface-bound antigen—the biological target of choice. Unbound phage are washed away whilst target-bound phage are reserved. Target-bound phage are then eluted from the immobilised antigen and regrown in *E. coli*. The now affinity-enriched phage cohort is then fed into the next cycle of biopanning.

There is a range of different labelled approaches (see Table 2) that are used for different purposes. DNA-encoded chemical libraries [36] have enabled expansion of the screening repertoire for up to 10^8 molecules, where compounds are chemically conjugated to DNA sequences that indicate their chemistry. Cell-free display techniques such as mRNA display allow for the production of libraries, with some of the largest degrees of genetic variation—reportedly above 10^{12} variants—because they can employ PCR-based approaches for generation of library variation [47]. Some techniques such as the protocol designed by Jones et al. (2016) have also performed phage display on cell-presented antigens (such as membrane-bound receptors) [48]. In this system, panning occurs on cells in solution, where the target is presented in its native state, including post-translational modifications.

Table 2. Prominent types of labelled library display.

Display Modality	Library Molecule Types	Maximum Library Size (Unique Clones)
DNA-displayed chemical library [36]	Single pharmacore (DNA-recorded synthesis), Dual-pharmacore.	10^{11} [49]
Phage Display (pIII coat protein most common fusion) [25,26]	Peptides, ScFv, Fab, sdAb/nanobodies.	10^{11} [50]
Yeast Display (Aga1p + Aga2p most common) [51]	Peptides, ScFv, Fab, sdAb/nanobodies, whole antibodies	10^9 [52]
Bacterial Display (Ipp + ompA, PAL, AhaA and intimin β-domains, APEx-NlpA or g3p, MAD-TRAP) [53,54]	Peptides, ScFv, Fab, sdAb/nanobodies, whole antibodies	10^{10} [54]
Mammalian cell display (fusion to transmembrane domain such as H-2Kk or PDGF receptor) and secretion (LoxP site inclusion on membrane anchor domain) [55,56]	Peptides, ScFv, Fab, sdAb/nanobodies, whole antibodies	10^9 [57]
mRNA display/cDNA display [34]	Peptides, ScFv, Fab, sdAb/nanobodies	10^{15} [58]
Ribosome Display [33]	Peptides, ScFv, Fab, sdAb/nanobodies	10^{15} [58]

One notable drawback of cell-free in vitro display methods is that the size and complexity of the molecules that can be expressed are usually limited. For example, phage display is limited by attenuation of phage infectivity and inefficient incorporation of large proteins into the major pIII phage coat protein [59]. Normally the pIII coat protein mediates phage infection of bacterial cells by attachment to the F-pilus [60], and this means that there is limit to the size of the molecule that can cloned into the pIII protein (about 50 residues) before phage infectivity is affected [59]. It also means that it is possible for library diversity to be affected through biopanning by the variable ability of each phage clone to re-infect the bacterial host. Other cell-free techniques such as mRNA display may eliminate this cell infection fitness bias. However, mRNA display is typically only useful for small peptide screening, as there is a reduced display efficiency for proteins larger than about 110 residues [61].

In contrast, cellular techniques such as bacteria, yeast, and mammalian surface display have the advantage of allowing for the display of larger molecules and do not require the post-panning bacterial infection step that may cause infection efficiency bias in enriched phage libraries. However, due to the necessity for cell transformation, cell display library sizes are smaller, reaching about a maximum of about 10^9 variants [62]. Both cellular and acellular in vitro approaches are generally higher throughput than in vivo methods, because they do not require the lengthy process of mouse sacrifice and spleen cell recovery for myeloid cell fusion. However, because in vitro techniques screen against targets outside the conditions present in normal immune response, antibodies generated against desired antigens may have altered pharmacokinetics and have more likelihood of causing unintended responses.

It is possible that future peptide library panning campaigns will not require a genotype–phenotype linkage. Recently, a nanopore sequencing technique has allowed for the direct sequencing of a 26-amino-acid peptide sequence [63]. This novel technique uses azide click chemistry to link a DNA strand to the C-terminus of the protein, where the DNA is then pulled through a bacterial pore using a DNA helicase. As the nanopore system detects charge variation across the pore to detect molecular signatures, it may be possible that it can also sequence protein phosphorylation or glycosylation. The technology has been used to distinguish unphosphorylated, monophosphorylated and diphosphorylated protein states [64]. Technologies such as this could be the next generation of template-free library screening for peptides.

Approaches to improve the efficiency of labelled molecular library screening are becoming increasingly inventive. Microfluidics platforms are one of the most prevalent tactics

to improve speed and efficiency of biopanning. Generally, these platforms are designed to reduce sample volume, create controllable fluid dynamics (uniform shear stress and velocity gradient), improve panning efficiency, and allow for automation of the biopanning process. It is important to be able to control fluid dynamics in panning systems, as mechanical force variation from shear stress induced in fluid flow affects peptide and antibody affinity for targets significantly [65–68]. Therefore, microfluidic channel fluid dynamics for library panning is an extensively studied field. Various modes of microfluidic panning exist from surface-adhered protein targets to microfluidic panning and screening using whole cells. Biopanning of phage-displayed proteins against surface-adhered cell cultures in microfluidics platforms has been shown to have a 600-fold greater level of enrichment when compared to conventional cell suspension-based biopanning [69]. Single-cell screening approaches such as Celli*GO* [70] use droplet microfluidics to perform high-throughput screening of immunised mouse-secreted IgG activity against soluble or membrane-bound antigens on a cell-by-cell basis. Other novel approaches include the use of alternating electrical current electrohydrodynamic (AC-EHD) flow, applied to disproportionally spaced gold electrodes, to improve microfluidic mixing at the target-phage library interface [71]. This phage library screening technique termed "PhageXpress" has been reported to require only a single biopanning step with AC-EHD flow, in combination with next-generation nanopore sequencing to achieve sufficient physical enrichment for isolation of target-binding phage clones. AC-EHD flow, used also in on-chip diagnostics [72–78], is proposed to reduce the "non-slip" property of fluid as it approaches a surface—a phenomenon described by Poiseuille's law [79]—which can perturb microfluidic technologies. On-chip application of electrical potentials has also been used to produce digital microfluidics, where for example, electrowetting on dielectrics has been used to automate all steps—including counter selection—in multiple rounds of phage biopanning to produce a competent target-binding peptide [80].

2.3. In Silico Techniques

Bottom-up in silico (computational) approaches have grown from complementing and augmenting in vivo and in vitro experimental techniques for molecular discovery. Techniques either search the theoretical chemical space to produce a list of candidate molecules that can be verified for function with wet-lab experiments or search already wet-lab-characterised molecules to find suitable structural templates for novel molecular design. More recently, computational methods have been used to begin to design molecules de novo, using data available online compiled from years of wet-lab experimental work. This is a movement towards more top-down molecular discovery approaches and will be described in more detail in the latter part of this review. In this section, some of the prominent bottom-up molecular discovery techniques will be briefly discussed.

It has become common for chemical libraries to undergo screening for activity against therapeutic targets using in silico approaches, as it is infeasible for in vitro synthesised libraries to represent the full extent of the possible drug-like molecules [81]. The accessible chemical space has been estimated to include at least 10^{60} [82] observable molecules, which is currently an impractical number of molecules to attempt to screen with wet-lab approaches [83]. In silico approaches enable a wider chemical space to be explored, where databases of online virtual libraries can computationally model chemical-target docking, enabling the screening of vast libraries of potential chemotherapeutics prior to synthesis for experimental validation. Recently, it has been reported that virtual libraries of up to 11 billion compounds can be screened [84]. Generally, the target structure is converted into a representation that can be used by the docking software. The binding pocket structure is then optimised using control calculations with data from known interactions. After screening and docking the virtual compound library, the top hits are selected for experimental validation [81]. Virtual screening enables potential therapeutic molecules to be narrowed down to a smaller, more experimentally viable number of compounds to be synthesised and screened. For example, the ZINC free online database [85] currently contains over

230 million compounds that can be readily synthesised, which are also available on the database in 3D enumerated formats that can be virtually screened and docked against structure-solved targets. Quantitative structure–activity relationships (QSAR) or similar methods are frequently used for optimising lead compounds after screening is performed, as they help elucidate the relationship between molecular structure and activity and can help to enhance this. These models seek to find a mathematical relationship between structure of chemicals and their various properties using regression and classification techniques [86,87]. Molecular docking programs also allow for predictive capability for interactions of known drugs with other targets, which can be a useful drug repurposing tool or off-target toxicity prediction tool.

Most current in silico techniques for discovery of biologically generated molecules (peptides and antibodies) are designed to enhance and compliment wet-lab experimental data by building on and enhancing known structures. For example, computational modelling using data from BLAST searches of custom online databases and PDB has been used to assist prediction of the in vitro developability of antibodies [88] or to model antibody variable fragments (Fv) to predict protein–protein and protein–antibody interactions [89,90]. Deep generative computational models—for example, long short-term memory networks (LSTM)—have been used to produce improved binders from a phage-displayed single-chain variable antibody fragment (scfv) panning campaign. Saka et al. (2021) [91] used an LSTM network to produce antibodies that were reported to have over 100-fold improved affinity compared to binders that were affinity matured purely with experimental work. Input data were generated from next generation sequencing of enriched phage sequences obtained from multiple biopanning rounds against Kynurenine protein. Codons were translated into amino acids and then encoded into a format suitable for data processing. The LSTM model was then trained from this enriched sequence data. New sequence generation of proposed improved binders was then performed, where for each new sequence, the negative logarithm of likelihood (assumed to correlate with binding affinity) was calculated to select new high-affinity binders. Other studies include the use of statistical analysis [92] or machine learning (ML) approaches [93] to enhance enzyme activity, where enzymes containing genetic mutations or variations were screened based on sequence for altered enzymatic activity. The resulting sequencing and activity data enabled the synthesis of new enzymes that had greater activity than to the original enzyme.

Whilst using the available molecular docking programs enables a far higher-throughput approach to screening the vast chemical space, these approaches are not yet able to comprehensively flip therapeutic discovery to a top-down approach. Although chemical compound structures are more easily computationally predicted than proteins, molecular docking still requires the input of a high-resolution image of the target structure (preferably ligand-bound) [81]. Protein Data Bank (PDB) [94], one of the main online databases for protein sequence–structure information, currently contains 184,700 released protein structures as of 7 December 2021. However, only 2666 of these released structures are from *Homo sapiens*, where at present, there are about 79,000 known proteins in the proteome as described in the online database UniProt [95]. Furthermore, only 71,489 of these released structures are within a 2 Å resolution. With the currently minimal knowledge of protein structures in the human proteome, in silico molecular docking and screening cannot yet be performed on the entire human proteome. As discussed previously, current gold-standard technologies for solving protein structure are laborious and time consuming, and it can be particularly difficult to obtain accurate representations for some targets [96]. Therefore, the advent of new technologies which are beginning to accurately predict protein structure from a sequence is a particularly exciting prospect for therapeutics development.
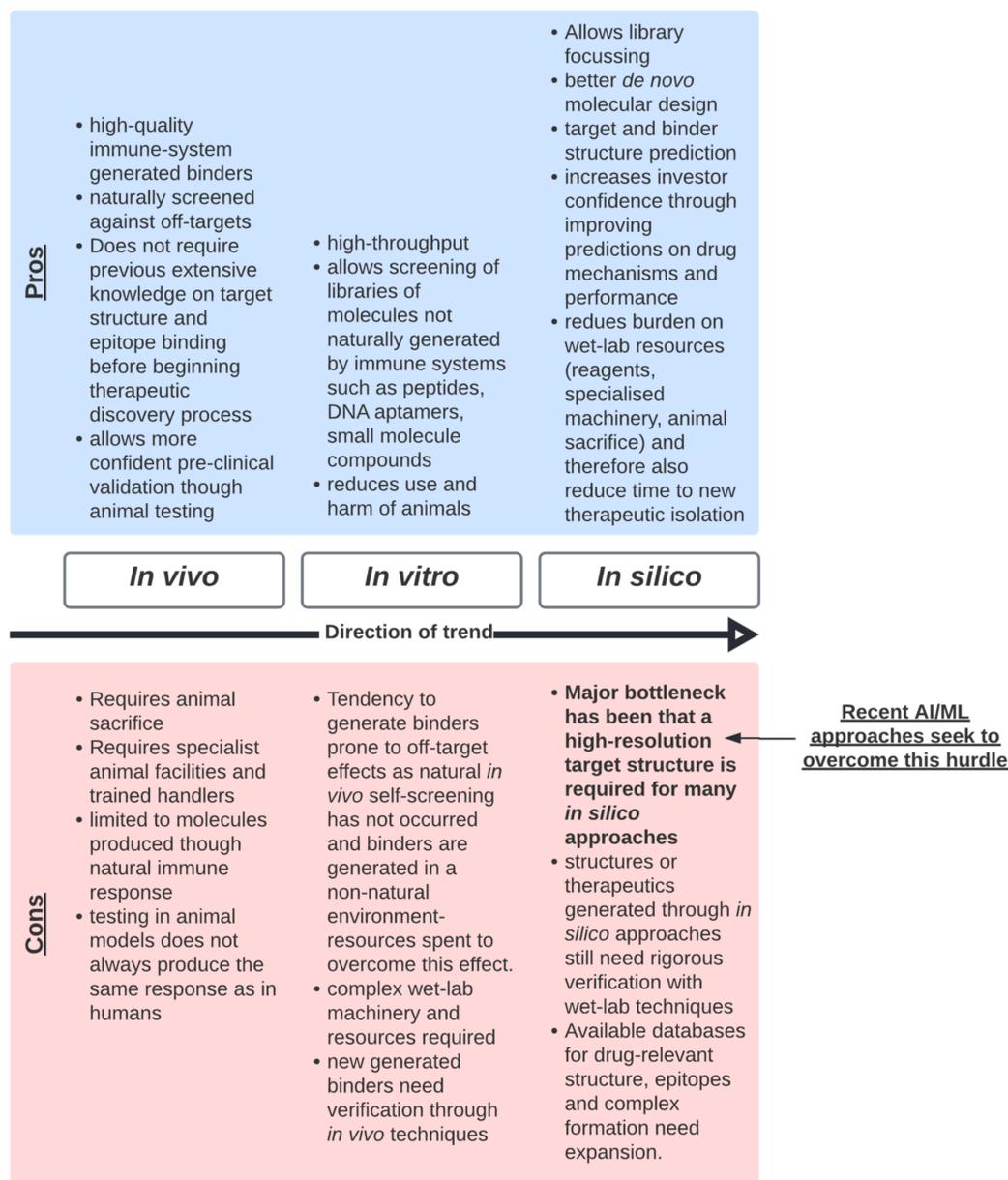
3. The Limitations of Bottom-Up Approaches

The current pipeline for taking new therapeutic molecules from discovery to clinic involves an incredibly lengthy, complex, and expensive process. This is in part because drug development necessitates collaborations between many fields of expertise, from chemists

and biologists to process scientists, clinicians, intellectual property and industry specialists and partners. It has been estimated that the median capitalized development and research cost to bring a new medicine to FDA approval in the United States between 2009 and 2018 was about USD 985 million, with projected costs expected to be in the billions [97]. The median clinical development time for FDA-approved drugs has remained stable at approximately 8.3 years for the past decade [98]. The significant time, expertise and cost associated with drug development usually translates into expensive out-of-pocket costs for patients [99,100]. As such, the process is not as simple as performing a molecular library screen and finding a new molecule that binds to a target.

There is an extraordinary amount of resources used in the current drug development pipeline. It takes years of work in the laboratory to first study and understand the biology of new therapeutic targets. Vast arrays of possible effector molecules are then screened against these new targets and narrowed down to a small cohort of binders, which are then characterised and further tested. Understanding the mechanisms and key targets for certain diseases can take entire careers to research and discover; thus, the next hurdle of developing a new therapeutic and taking it to clinic remains. For example, high-throughput molecular screening approaches for new drug discovery, such as phage-displayed library panning and enrichment [25,26] take molecular libraries, many orders of magnitude in size, and essentially "find what sticks" when searching for a new molecule to bind to or react with a target. In vivo techniques may also be used, where generally a target molecule is inoculated into an animal, and antibodies are produced naturally by the animals' immune system. Advances in this field mostly involve using alternative molecular display vectors, improving the efficiency of molecular panning and enrichment, or genetically engineering animals for better antibody generation. However, these approaches are still generally inefficient. The next step is to characterise these molecules (structure, mechanism of action, solubility, and permeability), optimise the binding properties, and then validate the binder in in vitro and animal models. The gold-standard techniques for solving protein structure—X-ray crystallography [101] and NMR spectroscopy—are capable of solving the crystallised structure of proteins with incredible detail but remain complex and time consuming [96]. In some cases, particularly for membrane-bound proteins, the crystallisation step required to image the protein is not possible. Usually, during the characterisation process, the drug is also tested in an appropriate animal model. In 2015, there were 192.1 million animals estimated to have been used for laboratory purposes globally [102]. There have been questions over the efficacy of the use of animals in drug testing, where occasionally, drugs deemed safe and effective in animals have proven toxic and ineffective in humans [103]. A summary of some of the advantages and disadvantages to in vivo, in vitro and in silico approaches can be found in Figure 5. Thus, current techniques to drug discovery can be costly, and questions over their effectiveness highlight a need to search for alternative approaches.

Once a drug has been isolated, it must then make it to the clinic. If enough funding and support is received, and a candidate molecule has been successfully isolated, trials can then begin in a small group of patients. If the molecule passes safety and efficacy tests, then the trials can expand to larger groups of patients. Because it takes a significant amount of funding to develop new therapeutics, it becomes a substantial gamble and barrier for investors on whether a potential new molecule will have clinical value compared to existing treatments. Unfortunately, this means that many drugs may not make it to the clinic simply because they have not received the required financial support. Once the therapeutic has been shown to have efficacy, it must still be able to be manufactured and distributed in a form that it is conducive to its use (e.g., thermostability or lyophilisation). Therefore, it is of significant interest to develop new techniques for therapeutics development that give investors and prescribers greater confidence in the efficacy of the product. Greater confidence that the molecule is going to behave in a more predictable way through patient trials and in the market will help to streamline the development process, reducing barriers to investment and delivery to patients.

	In vivo	**In vitro**	**In silico**
Pros	• high-quality immune-system generated binders • naturally screened against off-targets • Does not require previous extensive knowledge on target structure and epitope binding before beginning therapeutic discovery process • allows more confident pre-clinical validation though animal testing	• high-throughput • allows screening of libraries of molecules not naturally generated by immune systems such as peptides, DNA aptamers, small molecule compounds • reduces use and harm of animals	• Allows library focussing • better *de novo* molecular design • target and binder structure prediction • increases investor confidence through improving predictions on drug mechanisms and performance • redues burden on wet-lab resources (reagents, specialised machinery, animal sacrifice) and therefore also reduce time to new therapeutic isolation
Cons	• Requires animal sacrifice • Requires specialist animal facilities and trained handlers • limited to molecules produced though natural immune response • testing in animal models does not always produce the same response as in humans	• Tendency to generate binders prone to off-target effects as natural *in vivo* self-screening has not occurred and binders are generated in a non-natural environment-resources spent to overcome this effect. • complex wet-lab machinery and resources required • new generated binders need verification through *in vivo* techniques	• **Major bottleneck has been that a high-resolution target structure is required for many *in silico* approaches** • structures or therapeutics generated through *in silico* approaches still need rigorous verification with wet-lab techniques • Available databases for drug-relevant structure, epitopes and complex formation need expansion.

Direction of trend →

Recent AI/ML approaches seek to overcome this hurdle

Figure 5. Some of the major advantages and disadvantages to each broad approach to therapeutics discovery, with the trend moving towards increased use of in silico technologies.

The time and costs are not the only negatives associated with the slow and expensive drug development pipeline. Current limitations in the discovery pipeline means that current disease treatment tends towards more generic, rather than personalised approaches, which can affect patient experience and outcomes. With existing technologies, it is not yet feasible to develop molecular therapeutics on a patient-to-patient basis, although individual patient screening and characterisation and the use of global therapeutics that are tailored to a patients' clinical presentation is generally regarded by clinicians as crucial [104–106]. This has implications for patient outcomes, where significant inter- and intra-patient variability [107] means that available therapeutics are not only expensive [100] but can be "unreliable" in that they can cause unforeseen adverse effects and varied or limited impacts on a disease [108–112]. With the example of cancer, global treatments include classic chemotherapy and targeted treatments such as monoclonal antibodies (mAbs) or peptides. Classic chemotherapy involves the use of cytotoxic chemicals, including folate and purine analogues, inhibitors and promoters of microtubule polymerisation, and DNA damaging

agents [113]. These cytotoxic agents act by interfering with cellular division, which affects not only rapidly dividing malignant cells but also any other cells in the body that are undergoing division, such as haematopoietic and immune cells [113–115]. Freely circulating chemotherapy drugs often require administration in high doses, as they accumulate in high concentrations in the kidney and have short circulation times [116]. This results in a need for higher drug doses, which is problematic because classic chemotherapy has limited specificity and even low doses can cause systemic toxicity [117,118], whilst often resulting in only minor tumour response and sometimes tumour drug resistance [111,119,120].

Targeted therapeutics such as mAbs have high specificity for target antigens and have been employed as therapeutic agents for the past two decades [113]. Antibodies can impart therapeutic benefit by more specifically binding to targeted cells presenting specific antigens and marking them for antibody-dependent cell-mediated cytotoxicity (ADCC), complement-dependent cytotoxicity (CDC), or antibody-dependent cellular phagocytosis (ADCP) through opsonisation of target antigens [121]. Antibodies can also bind to targets and inhibit downstream functions through neutralization [113], and they have reportedly been found to induce cell differentiation and induce immune responses to target cells [122]. More recently, conjugation of antibodies and targeting peptides via linker chemicals to cytotoxic or radioactive payloads has allowed for precise delivery of therapeutic cargo to specific target cells [123,124]. Radiolabelling allows for either imaging or treatment (dependent on dosage), where uptake can be visualised through positron emission tomography (PET) scans. This mode of treatment has been coined as "theranostics" (therapeutic diagnostics). Whilst these targeted therapies are more precise and are less likely to impact off-target cells, significant inter- and intra-patient biological variability [107] means that treatments can cause unforeseen adverse effects and varied or limited impacts on a disease, significantly affecting patient outcomes [108–112]. Creating a pipeline where the off-target effects of therapeutics could be more accurately predicted and where therapeutics are more precisely designed for each targeted effect would be invaluable to patient outcomes.

Top-down approaches, where we can design a therapeutic from scratch and be able to predict molecular interactions, would have a massive benefit to streamlining the development pipeline and ultimately to patients. If we could design a therapeutic from scratch to specifically interact with a desired target, then it is possible that we could design novel therapeutics or modify existing ones with their specific interactome in mind. This would enable better prediction, and possibly elimination, of off-target toxicities in patients and better tools for clinicians to predict which patients to prescribe and which patients not to prescribe. For the same reasons, a top-down approach may help to streamline many areas of the therapeutics development pipeline, not only by improving molecular design but by making products more likely to be competitive in the market. Increased confidence in and better prediction of drug performance shifts the risk of drug development to earlier in the pipeline, where there is minimal collateral damage incurred upon the failure of a new product. This could help reduce barriers to financial investment and regulatory approval at later stages of the pipeline by increasing the likelihood and predictability of drug success.

4. Towards a Top-Down Approach

Machine Learning for Predicting Molecular Structure and Complex Formation

Recent advances in computational science are beginning to make the paradigm shift to a top-down approach for molecular discovery possible. Major developments in high-throughput technologies, including next generation sequencing, protein expression and protein characterisation, protein structure deconvolution and computational advances, are the foundation for shifting to top-down technologies. The wealth of information available pertaining to molecular sequence and structural information in online databases has accelerated computational technologies to the forefront of top-down molecular discovery approaches. Whilst many computational approaches have largely focussed on improving known target-ligand interactions or are designed to complement wet-lab techniques, recently, there has been a shift towards developing entirely de novo molecular designs.

A range of de novo approaches to molecular design has been developed to date. Deep and reinforcement learning has been used to generate novel chemical libraries that have desired properties. A novel computational strategy call ReLeaSE (Reinforcement Learning for Structural Evolution) integrates a generative and a predictive deep neural network to generate libraries of de novo structures. These libraries can be geared towards having certain levels of structural complexity, or a specific range of physical properties (melting point, hydrophobicity), or to have inhibitory activity against certain targets. Specifically designing libraries in this manner allows molecular libraries to be targeted towards certain functions and increases the likelihood of finding a molecule with desired properties in wet-lab experiments. The process of obtaining a new binder using this method is not strictly top-down, as the new libraries still need to be specifically screened against the target of interest. However, they still use de novo molecular generation techniques to create desired properties; thus, this could be viewed as a top-down approach in this aspect.

ML algorithms trained using sets of known structures of proteins available online on PDB and other databases have recently made significant advances in predicting protein folding [125–127]. DeepMind's AlphaFold 2 has received a considerable amount of media attention for having "solved" the 50-year-old grand challenge of using the nucleic acid sequence of a protein to predict protein folding. In the 14th Critical Assessment of Protein Structure Prediction (CASP14) in 2020 [128], AlphaFold 2 achieved computational prediction of protein folding to near-experimental accuracy (all-atom protein structure prediction from within 1.2 to 1.6 Å at a 95% confidence interval, where the width of a carbon atom is approximately 1.4 Å) from the analysed sequences [126]. The model's training uses a set of protein structural and sequence data, and multiple sequence alignments (MSAs) of the original sequence against proteins from many species, giving evolutionary information on sequence conservation. These data were derived from the Protein Data Bank [94] and Uniclust30 [129], and MSAs were generated using publicly available software. The building block of the novel neural network—"Evoformer"—produces processed representations of the MSA of the presentation of residues across different species and a processed pairwise representation of the structural relation between residues ($N_{res} \times N_{res}$ array) that are refined as they progress through the model. The final paired and MSA representations are then fed into a model that gives 3D structure in terms of rotation and translation for each residue of the protein, and further iterative refinement processes in the model produces an output of predicted protein structure.

There is a range of groups that have produced variations or alternatives to the technology, particularly since the second iteration of AlphaFold from the CASP14 competition. The team behind the second-best protein structure prediction method after AlphaFold, RoseTTAFold [130], which can operate within a fraction of the time taken by AlphaFold (~10 min on a single GPU as opposed to days with several GPUs), has since produced a combination of AlphaFold and RoseTTAFold to predict protein complexes [131]. Using a paired MSA, they identified 1505 likely-to-interact proteins to form complexes in yeast *Saccharomyces cerevisiae*, 699 of which had been structurally characterised and 700 of which were supported in the literature. However, these are predictions of stable protein complex formations, which would likely have highly conserved and stable pMSAs through different species, due to the nature of the interaction being under significant selective pressure. More transient, weaker interactions that are under lower selective pressure were noted by the authors as likely to be difficult to predict, as well as binary protein complexes from high-order obligate protein associations. It was also noted that predictions would be inaccurate for species-unique protein complexes and that there would be an overrepresentation of hydrophobic or amphipathic domains. As well as this, recent approaches involving deep network hallucination have investigated the generation of new proteins with entirely novel backbone structures and amino acid sequences [132].

Other groups have made significant steps towards de novo design of target-binding proteins given nothing but the target structure. The design of proteins given no other information (such as homologous target-binding proteins) other than the target structure has

been a major challenge for in silico therapeutics development. As previously described, it is challenging to predict interactions that are more transient, non-covalent and under lower selective pressure. The David Baker lab at EMBL-EBI is one of the first groups claiming to be able to perform de novo design for more transient interactions of protein binders and targets. In this work, the group has moved beyond using a native or homologous protein scaffold and previously identified hot-spot residues to identify and predict proteins with favourable and more transient interactions [133]. This approach begins by performing target docking of a library of "disembodied" amino acids and creating what is termed a "rotamer interaction field" (RIF). This RIF is the stored backbone coordinates and binding energies of billions of amino acids that are screened in this docking approach that have favourable more transient (hydrogen bonding and non-polar) interactions. This RIF is used as a field which targets docking of large protein scaffold libraries that can be more rapidly scored (due to reduced algorithmic complexity created through the RIF) based on only backbone amino acid coordinates. Focussed screening and motif clustering and selection allows the deconvolution of motifs that are predicted to bind to the target through hydrogen bonding or non-covalent interactions. Whilst the RIF is designed to attempt to reduce the required computational power, the process still requires many computationally demanding steps, and there is a low success rate for producing designs that bind the target successfully. Even among molecules that were found to bind in some capacity to targets, it was still required to make amino acid substitutions to create a strong binder. Therefore, whilst this approach is one of the first promising solutions to de novo design of therapeutics, more work is required to make this method deliverable.

Spirited claims have been made that the recent advances in machine learning models are one of the greatest developments in structural biology in the past 20 years, which is also one of the biggest achievements in artificial intelligence in recent history [134]. Some have even asserted that that AlphaFold's predictions have virtually solved the structure of the entire human proteome [135]. Since CASP14, the AlphaFold team in conjunction with EMBL-EBI has recently released their predictions for 98.5% of the human proteome through building on the AlphaFold model, where they state to have predicted 58% of proteins with confidence and 38% with high confidence [136]. As previously described, ligand docking is already used to computationally model how chemical therapeutics interact with structurally solved targets and to screen libraries of potential chemical therapeutics against these targets. Therefore, the implications for being able to predict protein folding without the need for experimental validation of protein structures—which is one of the biggest current bottlenecks to in silico approaches to therapeutics discovery—would quickly and significantly extend the possibilities in the field of therapeutics discovery. It is possible that eventually all library screening could be performed in silico without the need for experimental determination of target—or protein and antibody library—structures. In addition, instead of a bottom-up screening approach where an investigator is trying to "find the needle in the haystack", you might be able to flip this to a top-down approach, where one could build a molecule with the exact characteristics you desire entirely de novo using computational modelling.

Whilst the prospects of this computational achievement are exciting, it remains a question of how realistically useful and reliable these models are in their current state, particularly for molecular discovery applications. Foremost, this technology is not yet user friendly or easy to access by anyone who is not an expert in the field. For example, whilst the open-source code for AlphaFold has been released and is publicly available (https://github.com/deepmind/alphafold, accessed on 28 February 2022), the value in ML lies not in the code but in the training dataset. ML algorithms will have markedly different outputs depending on the data they are trained on. Whilst it is possible to recreate the AlphaFold training dataset, it remains highly computationally challenging and time consuming to build and mimic. This means that whilst the code may have been released for these ML algorithms, it is not in a form that allows this technology to be readily accessed by researchers to predict new structures or interactions. As well as

this, whilst data for protein structure are readily available, the complexity of molecular interactions and a limit to online datasets specific to molecular interactions and therapeutics make AI for molecular discovery and interaction predictions challenging. Regardless of the challenge of compiling a robust and accessible drug-related database, proteins are constantly in motion in cells [137]. Not only are they being altered by post-translational modifications, but they are changing conformations depending on the function being performed and ligand-bound states. Generally, to model ligand docking, the ligand-bound (holo) protein structure is preferred to the unbound (apo) structure, and there is some question as to whether current AI protein structure predictions can differentiate between holo and apo states. Some studies have shown that peptide-bound protein complexes could be predicted accurately by concatenating the sequences of receptors and peptides using a poly-glycine linker [138]. Google's Deepmind has also recently released the source code for their "AlphaFold-Multimer", which has been reported to be able to accurately predict heteromeric protein docking interfaces in 67% of cases, and homomeric interfaces in 69% of cases, but were not able to predict antibody interactions [139]. However, other researchers have remained uncertain about the accuracy of predictions in protein active sites, as when AlphaFold and RoseTTAFold predictions were overlayed onto the experimentally solved structures of 20 GPCRs, only about 50% of the predicted structures were useful [140]. AlphaFold is expected to have poor predictive capability for the intrinsically disordered active sites or those that are unstructured in isolation [140]. It is also not expected to be able to capture the protein destabilising effects of point mutations, which would be important for being able to accurately model disease states.

Although there are many caveats to in silico predictive technologies that need development, there is still much promise in the field. Similar issues to accurate structure and protein interaction prediction exist with any protein structure imaging technique. Whether an image is obtained in the holo or apo state, it is still only a snapshot of the possible conformations that this protein could undertake. As previously mentioned, there are also a significant number of protein structures in PDB that are relatively low resolution due to limitations to wet-lab experimental approaches; thus, ML techniques such as AlphaFold could help to improve the ability to perform molecular docking with these structures. Post-translational modifications, such as glycosylation, and flexible, less constrained domains are also still difficult to structurally determine using X-ray crystallography [141]. It may also be possible that ML models could be trained without the need to obtain protein structures. Early studies have suggested that it is possible to train deep generative models exclusively on antibody sequence data to design new epitope-specific antibodies [142]. As previously discussed, sequence data can be used to make statistical inference or train ML algorithms, enabling the improvement of enzymatic activity or antibody affinity [91–93]. If sequence data could be used in lieu of structural data to train ML algorithms to predict new therapeutics, they could significantly expand this field of therapeutics discovery and off-target prediction. Sequence-based data are significantly more accessible than structural data because experimental validation of the interaction can be derived from sequencing and relatively simple laboratory tests such as ELISA rather than X-ray crystallography or cryo-EM.

It is not yet clear how this new AI technology can be adopted by drug developers, and in which role it will play in therapeutics development. It can be hypothesised that the technology will streamline and accelerate the journey to progress a therapeutic from discovery to clinic. If successful, the technology could reduce the time taken to obtain new therapeutics, by reducing or perhaps eliminating the need for bottom-up or library screening approaches. Blue-sky thinking would encourage the idea that we may begin to create drugs on an individual patient basis. Potentially, the technology could allow for identification and screening of off-target binders that produce toxicity for patients by allowing whole proteome screening. Being able to reduce the risk of patient toxicity would certainly improve investor confidence that the drug candidate will be successful in clinic, as well as decreasing the likelihood of failure through clinical trials. These things would

allow for the acceleration of therapeutics through the development journey. Whilst the recent predictive technologies are a shift in the top-down approach to molecular discovery, in the current state, they are not yet positioned to replace the existing gold standard in in vivo and in vitro techniques. As well as this, biological diversity is thus large that the amount of information that would need to be processed for AI to predict off-targets for a single existing drug for each patient with high accuracy would be vast, let alone creating new drugs on a patient-by-patient basis. Predictive models such as AlphaFold still require significant development and further prospective validation experiments, but existing data are promising.

Before computational approaches can enable a top-down approach to drug discovery, a united effort between all areas of bottom-up therapeutics discovery will likely be required. Expansive, high-quality wet-lab data are generally considered as critical for the development of successful ML algorithms for molecular discovery [143]. Large volumes of binding motif data are highly valuable to predict paratope–epitope features [144,145]. Therefore, the use and further development of bottom-up wet lab techniques designed to feed into more extensive online databases and computational analysis would be likely to help drive predictive outputs for top-down approaches. If it is possible to predict paratope–epitope interactions from sequence only, then it may be suggested that high-throughput wet-lab experimental techniques that can quickly and consistently identify binder-target sequences—not only holo structures—of molecular interactions in a high-throughput manner would be especially useful. This means that high-quality experimental design and consistent and high-throughput assay technology that can be used to create data from a broad range of samples and conditions will be a driving factor behind enabling top-down therapeutics design. Selected techniques will need to minimise noise and batch effects (e.g., variation in data from performing the same experiment in two different laboratories), as this would likely have a significant effect on ML-based predictions.

5. Conclusions

The recent advances in therapeutics discovery have the potential to be highly beneficial to all areas of the drug discovery journey. In summary, the critical points from this review article are:

- There is a clear need for a more efficient pipeline for the journey of therapeutics, from discovery to the clinic, as highlighted by the recent novel coronavirus pandemic. Emerging AI technologies may enable a smoother transition through the various stages of the pipeline, including reducing barriers such as regulatory hurdles and market performance.
- Bottom-up techniques are slower and more resource exhaustive than top-down techniques and, due to the nature of the approach, requires much downstream characterisation and validation. It is recommended that we should continue to improve the quality of data generated from bottom-up technologies. This will be a critical step to move to top-down technologies, as these data will feed these new approaches.
- The earlier that the risk of failure of the development of a new drug is addressed and predicted, the more likely it is that the drug will successfully and quickly reach the clinic with minimal losses. It is recommended that we move towards top-down approaches to drug discovery that enable stronger understanding of molecular behaviour, as investors are more likely to support a drug that they are confident will be successful and competitive in the market.
- It is hypothesised that access to more personalised medicine will enable clinicians to effectively compete with the biological intricacies of complex disease.
- Emerging top-down AI technologies are improving the prediction of molecular structures, molecular behaviour and optimal de novo drug design. It is recommended that we continue to incorporate and improve these techniques into therapeutics discovery, as they will help to streamline the development pipeline, although further work is needed to contend with the massive complexity of biological systems.

- Encouraging the sharing and improved accessibility to ML approaches for people who are not specialists in the field may help to drive innovation and discovery with new ideas and perspectives.

Author Contributions: S.E.K., D.K. and C.B.H. conceptualised and outlined the format of the manuscript; S.E.K. performed most of the writing of the manuscript; D.K., F.A. and W.J.L. assisted in revising and editing the manuscript; M.T. and C.B.H. reviewed and provided valuable feedback and comments; M.T. approved the manuscript for submission. All authors have read and agreed to the published version of the manuscript.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Mullard, A. 2020 FDA drug approvals. *Nat. Rev. Drug Discov.* **2021**, *20*, 85–90. [CrossRef]
2. Muttenthaler, M.; King, G.F.; Adams, D.J.; Alewood, P.F. Trends in peptide drug discovery. *Nat. Rev. Drug Discov.* **2021**, *20*, 309–325. [CrossRef]
3. Grilo, A.L.; Mantalaris, A. The increasingly human and profitable monoclonal antibody market. *Trends Biotechnol.* **2019**, *37*, 9–16. [CrossRef]
4. U.S. Food and Drug Administration FDA COVID-19 Response. Available online: https://www.fda.gov/emergency-preparedness-and-response/counterterrorism-and-emerging-threats/coronavirus-disease-2019-covid-19 (accessed on 18 November 2021).
5. Agostini, M.L.; Andres, E.L.; Sims, A.C.; Graham, R.L.; Sheahan, T.P.; Lu, X.; Smith, E.C.; Case, J.B.; Feng, J.Y.; Jordan, R.; et al. Coronavirus susceptibility to the antiviral remdesivir (GS-5734) is mediated by the viral polymerase and the proofreading exoribonuclease. *mBio* **2018**, *9*, e00221-18. [CrossRef]
6. Madelain, V.; Baize, S.; Jacquot, F.; Reynard, S.; Fizet, A.; Barron, S.; Solas, C.; Lacarelle, B.; Carbonnelle, C.; Mentré, F.; et al. Ebola viral dynamics in nonhuman primates provides insights into virus immuno-pathogenesis and antiviral strategies. *Nat. Commun.* **2018**, *9*, 4013. [CrossRef]
7. Beigel, J.H.; Tomashek, K.M.; Dodd, L.E.; Mehta, A.K.; Zingman, B.S.; Kalil, A.C.; Hohmann, E.; Chu, H.Y.; Luetkemeyer, A.; Kline, S.; et al. Remdesivir for the Treatment of COVID-19—Final Report. *N. Engl. J. Med.* **2020**, *383*, 1813–1826. [CrossRef]
8. Cohen, J. The 'very, very bad look' of remdesivir, the first FDA-approved COVID-19 drug. *Science* **2020**, *370*, 642–643. [CrossRef]
9. Szemiel, A.M.; Merits, A.; Orton, R.J.; MacLean, O.A.; Pinto, R.M.; Wickenhagen, A.; Lieber, G.; Turnbull, M.L.; Wang, S.; Furnon, W.; et al. In vitro selection of Remdesivir resistance suggests evolutionary predictability of SARS-CoV-2. *PLOS Pathog.* **2021**, *17*, e1009929. [CrossRef]
10. Lee, E.C.; Liang, Q.; Ali, H.; Bayliss, L.; Beasley, A.; Bloomfield-Gerdes, T.; Bonoli, L.; Brown, R.; Campbell, J.; Carpenter, A.; et al. Complete humanization of the mouse immunoglobulin loci enables efficient therapeutic antibody discovery. *Nat. Biotechnol.* **2014**, *32*, 356–363. [CrossRef]
11. Traggiai, E.; Becker, S.; Subbarao, K.; Kolesnikova, L.; Uematsu, Y.; Gismondo, M.R.; Murphy, B.R.; Rappuoli, R.; Lanzavecchia, A. An efficient method to make human monoclonal antibodies from memory B cells: Potent neutralization of SARS coronavirus. *Nat. Med.* **2004**, *10*, 871–875. [CrossRef]
12. Wrammert, J.; Smith, K.; Miller, J.; Langley, W.A.; Kokko, K.; Larsen, C.; Zheng, N.Y.; Mays, I.; Garman, L.; Helms, C.; et al. Rapid cloning of high-affinity human monoclonal antibodies against influenza virus. *Nature* **2008**, *453*, 667–671. [CrossRef]
13. Tonegawa, S. Somatic generation of antibody diversity. *Nature* **1983**, *302*, 575–581. [CrossRef]
14. Chi, X.; Li, Y.; Qiu, X. V(D)J recombination, somatic hypermutation and class switch recombination of immunoglobulins: Mechanism and regulation. *Immunology* **2020**, *160*, 233–247. [CrossRef]
15. Glanville, J.; Zhai, W.; Berka, J.; Telman, D.; Huerta, G.; Mehta, G.R.; Ni, I.; Mei, L.; Sundar, P.D.; Day, G.M.R.; et al. Precise determination of the diversity of a combinatorial antibody library gives insight into the human immunoglobulin repertoire. *Proc. Natl. Acad. Sci. USA* **2009**, *106*, 20216–20221. [CrossRef]

16. Köhler, G.; Milstein, C. Continuous cultures of fused cells secreting antibody of predefined specificity. *Nature* **1975**, *256*, 495–497. [CrossRef]

17. Kashmiri, S.V.S.; De Pascalis, R.; Gonzales, N.R.; Schlom, J. SDR grafting—A new approach to antibody humanization. *Methods* **2005**, *36*, 25–34. [CrossRef]

18. Zhu, F.; Nair, R.R.; Fisher, E.M.C.; Cunningham, T.J. Humanising the mouse genome piece by piece. *Nat. Commun.* **2019**, *10*, 1845. [CrossRef]

19. Agarwal, Y.; Beatty, C.; Ho, S.; Thurlow, L.; Das, A.; Kelly, S.; Castronova, I.; Salunke, R.; Biradar, S.; Yeshi, T.; et al. Development of humanized mouse and rat models with full-thickness human skin and autologous immune cells. *Sci. Rep.* **2020**, *10*, 14598. [CrossRef]

20. Xu, J.; Xu, K.; Jung, S.; Conte, A.; Lieberman, J.; Muecksch, F.; Lorenzi, J.C.C.; Park, S.; Schmidt, F.; Wang, Z.; et al. Nanobodies from camelid mice and llamas neutralize SARS-CoV-2 variants. *Nature* **2021**, *595*, 278–282. [CrossRef]

21. Rogers, T.F.; Zhao, F.; Huang, D.; Beutler, N.; Burns, A.; He, W.T.; Limbo, O.; Smith, C.; Song, G.; Woehl, J.; et al. Isolation of potent SARS-CoV-2 neutralizing antibodies and protection from disease in a small animal model. *Science* **2020**, *369*, 956–963. [CrossRef]

22. Schardt, J.S.; Pornnoppadol, G.; Desai, A.A.; Park, K.S.; Zupancic, J.M.; Makowski, E.K.; Smith, M.D.; Chen, H.; Garcia de Mattos Barbosa, M.; Cascalho, M.; et al. Discovery and characterization of high-affinity, potent SARS-CoV-2 neutralizing antibodies via single B cell screening. *Sci. Rep.* **2021**, *11*, 20738. [CrossRef] [PubMed]

23. Prado, N.D.R.; Pereira, S.S.; Da Silva, M.P.; Morais, M.S.S.; Kayano, A.M.; Moreira-Dill, L.S.; Luiz, M.B.; Zanchi, F.B.; Fuly, A.L.; Huacca, M.E.F.; et al. Inhibition of the myotoxicity induced by Bothrops jararacussu venom and isolated phospholipases A2 by specific camelid single-domain antibody fragments. *PLoS ONE* **2016**, *11*, e0151363. [CrossRef]

24. Ramos, H.R.; Junqueira-de-Azevedo, I.d.L.M.; Novo, J.B.; Castro, K.; Duarte, C.G.; Machado-de-Ávila, R.A.; Chavez-Olortegui, C.; Ho, P.L. A heterologous multiepitope DNA prime/recombinant protein boost immunisation strategy for the development of an antiserum against micrurus corallinus (*Coral Snake*) venom. *PLoS Negl. Trop. Dis.* **2016**, *10*, e0004484. [CrossRef]

25. Smith, G.P. Filamentous fusion phage: Novel expression vectors that display cloned antigens on the virion surface. *Science* **1985**, *228*, 1315–1317. [CrossRef]

26. Clackson, T.; Hoogenboom, H.R.; Griffiths, A.D.; Winter, G. Making antibody fragments using phage display libraries. *Nature* **1991**, *352*, 624–628. [CrossRef]

27. Gunneriusson, E.; Samuelson, P.; Uhlén, M.; Nygren, P.Å.; Ståhl, S. Surface display of a functional single-chain Fv antibody on staphylococci. *J. Bacteriol.* **1996**, *178*, 1341–1346. [CrossRef]

28. Schneewind, O.; Fowler, A.; Faull, K.F. Structure of the cell wall anchor of surface proteins in Staphylococcus aureus. *Science* **1995**, *268*, 103–106. [CrossRef]

29. Schreuder, M.P.; Brekelmans, S.; Van Den Ende, H.; Klis, F.M. Targeting of a heterologous protein to the cell wall of *Saccharomyces cerevisiae*. *Yeast* **1993**, *9*, 399–409. [CrossRef]

30. Ueda, M.; Tanaka, A. Genetic immobilization of proteins on the yeast cell surface. *Biotechnol. Adv.* **2000**, *18*, 121–140. [CrossRef]

31. Murai, T.; Ueda, M.; Atomi, H.; Shibasaki, Y.; Kamasawa, N.; Osumi, M.; Kawaguchi, T.; Arai, M.; Tanaka, A. Genetic immobilization of cellulase on the cell surface of Saccharomyces cerevisiae. *Appl. Microbiol. Biotechnol.* **1997**, *48*, 499–503. [CrossRef]

32. Parthiban, K.; Perera, R.L.; Sattar, M.; Huang, Y.; Mayle, S.; Masters, E.; Griffiths, D.; Surade, S.; Leah, R.; Dyson, M.R.; et al. A comprehensive search of functional sequence space using large mammalian display libraries created by gene editing. *mAbs* **2019**, *11*, 884–898. [CrossRef] [PubMed]

33. Hanes, J.; Plückthun, A. In vitro selection and evolution of functional proteins by using ribosome display. *Proc. Natl. Acad. Sci. USA* **1997**, *94*, 4937–4942. [CrossRef] [PubMed]

34. Roberts, R.W.; Szostak, J.W. RNA-peptide fusions for the in vitro selection of peptides and proteins. *Proc. Natl. Acad. Sci. USA* **1997**, *94*, 12297–12302. [CrossRef] [PubMed]

35. Nemoto, N.; Miyamoto-Sato, E.; Husimi, Y.; Yanagawa, H. In vitro virus: Bonding of mRNA bearing puromycin at the 3′-terminal end to the C-terminal end of its encoded protein on the ribosome in vitro. *FEBS Lett.* **1997**, *414*, 405–408. [CrossRef]

36. Brenner, S.; Lerner, R.A. Encoded combinatorial chemistry. *Proc. Natl. Acad. Sci. USA* **1992**, *89*, 5381–5383. [CrossRef]

37. Bunin, B.A.; Ellman, J.A. A general and expedient method for the solid-phase synthesis of 1,4-benzodiazepine derivatives. *J. Am. Chem. Soc.* **1992**, *114*, 10997–10998. [CrossRef]

38. Dewitt, S.H.; Kiely, J.S.; Stankovic, C.J.; Schroeder, M.C.; Cody, D.M.R.; Pavia, M.R. "Diversomers": An approach to nonpeptide, nonoligomeric chemical diversity. *Proc. Natl. Acad. Sci. USA* **1993**, *90*, 6909–6913. [CrossRef]

39. Dolle, R.E. Historical overview of chemical library design. *Methods Mol. Biol.* **2011**, *685*, 3–25. [PubMed]

40. Stylli, C.; Beckey, S.S.; Shumate, C.B.; Coassin, P.J. Systems and Methods for Rapidly Identifying Useful Chemicals in Liquid Samples. U.S. Patent US6472218B1, 31 August 2004.

41. Michael, S.; Auld, D.; Klumpp, C.; Jadhav, A.; Zheng, W.; Thorne, N.; Austin, C.P.; Inglese, J.; Simeonov, A. A robotic platform for quantitative high-throughput screening. *Assay Drug Dev. Technol.* **2008**, *6*, 637–657. [CrossRef]

42. Kornienko, O.; Lacson, R.; Kunapuli, P.; Schneeweis, J.; Hoffman, I.; Smith, T.; Alberts, M.; Inglese, J.; Strulovici, B. Miniaturization of whole live cell-based GPCR assays using microdispensing and detection systems. *J. Biomol. Screen.* **2004**, *9*, 186–195. [CrossRef] [PubMed]

43. Marks, J.D.; Hoogenboom, H.R.; Bonnert, T.P.; McCafferty, J.; Griffiths, A.D.; Winter, G. By-passing immunization. Human antibodies from V-gene libraries displayed on phage. *J. Mol. Biol.* **1991**, *222*, 581–597. [CrossRef]

44. Winter, G.; Griffiths, A.D.; Hawkins, R.E.; Hoogenboom, H.R. Making antibodies by phage display technology. *Annu. Rev. Immunol.* **1994**, *12*, 433–455. [CrossRef] [PubMed]

45. Knappik, A.; Ge, L.; Honegger, A.; Pack, P.; Fischer, M.; Wellnhofer, G.; Hoess, A.; Wölle, J.; Plückthun, A.; Virnekäs, B. Fully synthetic human combinatorial antibody libraries (HuCAL) based on modular consensus frameworks and CDRs randomized with trinucleotides. *J. Mol. Biol.* **2000**, *296*, 57–86. [CrossRef] [PubMed]

46. Cesaro-Tadic, S.; Lagos, D.; Honegger, A.; Rickard, J.H.; Partridge, L.J.; Blackburn, G.M.; Plückthun, A. Turnover-based in vitro selection and evolution of biocatalysts from a fully synthetic antibody library. *Nat. Biotechnol.* **2003**, *21*, 679–685. [CrossRef] [PubMed]

47. Seelig, B. MRNA display for the selection and evolution of enzymes from in vitro-translated protein libraries. *Nat. Protoc.* **2011**, *6*, 540–552. [CrossRef]

48. Jones, M.L.; Alfaleh, M.A.; Kumble, S.; Zhang, S.; Osborne, G.W.; Yeh, M.; Arora, N.; Hou, J.J.C.; Howard, C.B.; Chin, D.Y.; et al. Targeting membrane proteins for antibody discovery using phage display. *Sci. Rep.* **2016**, *6*, 26240. [CrossRef]

49. Franzini, R.M.; Randolph, C. Chemical space of DNA-encoded libraries: Miniperspective. *J. Med. Chem.* **2016**, *59*, 6629–6644. [CrossRef]

50. Almagro, J.C.; Pedraza-Escalona, M.; Arrieta, H.I.; Pérez-Tapia, S.M. Phage display libraries for antibody therapeutic discovery and development. *Antibodies* **2019**, *8*, 44. [CrossRef] [PubMed]

51. Boder, E.T.; Wittrup, K.D. Yeast surface display for screening combinatorial polypeptide libraries. *Nat. Biotechnol.* **1997**, *15*, 553–557. [CrossRef]

52. Cherf, G.M.; Cochran, J.R. Applications of yeast surface display for protein engineering. *Methods Mol. Biol.* **2015**, *1319*, 155–175. [CrossRef] [PubMed]

53. Francisco, J.A.; Campbell, R.; Iverson, B.L.; Georgiou, G. Production and fluorescence-activated cell sorting of Escherichia coli expressing a functional antibody fragment on the external surface. *Proc. Natl. Acad. Sci. USA* **1993**, *90*, 10444–10448. [CrossRef]

54. Bessette, P.H.; Rice, J.J.; Daugherty, P.S. Rapid isolation of high-affinity protein binding peptides using bacterial display. *Protein Eng. Des. Sel.* **2004**, *17*, 731–739. [CrossRef] [PubMed]

55. Ho, M.; Nagata, S.; Pastan, I. Isolation of anti-CD22 Fv with high affinity by Fv display on human cells. *Proc. Natl. Acad. Sci. USA* **2006**, *103*, 9637–9642. [CrossRef]

56. Beerli, R.R.; Bauer, M.; Buser, R.B.; Gwerder, M.; Muntwiler, S.; Maurer, P.; Saudan, P.; Bachmann, M.F. Isolation of human monoclonal antibodies by mammalian cell display. *Proc. Natl. Acad. Sci. USA* **2008**, *105*, 14336–14341. [CrossRef] [PubMed]

57. Bowers, P.M.; Horlick, R.A.; Kehry, M.R.; Neben, T.Y.; Tomlinson, G.L.; Altobell, L.; Zhang, X.; Macomber, J.L.; Krapf, I.P.; Wu, B.F.; et al. Mammalian cell display for the discovery and optimization of antibody therapeutics. *Methods* **2014**, *65*, 44–56. [CrossRef] [PubMed]

58. Kamalinia, G.; Grindel, B.J.; Takahashi, T.T.; Millward, S.W.; Roberts, R.W. Directing evolution of novel ligands by mRNA display. *Chem. Soc. Rev.* **2021**, *50*, 9055–9103. [CrossRef] [PubMed]

59. Omidfar, K.; Daneshpour, M. Advances in phage display technology for drug discovery. *Expert Opin. Drug Discov.* **2015**, *10*, 651–669. [CrossRef] [PubMed]

60. Bábíčková, J.; Tóthová, Ľ.; Boor, P.; Celec, P. In vivo phage display—A discovery tool in molecular biomedicine. *Biotechnol. Adv.* **2013**, *31*, 1247–1259. [CrossRef] [PubMed]

61. Takahashi, T.T.; Austin, R.J.; Roberts, R.W. mRNA display: Ligand discovery, interaction analysis and beyond. *Trends Biochem. Sci.* **2003**, *28*, 159–165. [CrossRef]

62. Breous-Nystrom, E.; Schultze, K.; Meier, M.; Flueck, L.; Holzer, C.; Boll, M.; Seibert, V.; Schuster, A.; Blanusa, M.; Schaefer, V.; et al. Retrocyte Display®technology: Generation and screening of a high diversity cellular antibody library. *Methods* **2014**, *65*, 57–67. [CrossRef] [PubMed]

63. Brinkerhoff, H.; Kang, A.S.W.; Liu, J.; Aksimentiev, A.; Dekker, C. Multiple rereads of single proteins at single—Amino acid resolution using nanopores. *Science* **2021**, *374*, 1509–1513. [CrossRef] [PubMed]

64. Rosen, C.B.; Rodriguez-Larrea, D.; Bayley, H. Single-molecule site-specific detection of protein phosphorylation with a nanopore. *Nat. Biotechnol.* **2014**, *32*, 179–181. [CrossRef] [PubMed]

65. Kuo, S.C.; Lauffenburger, D.A. Relationship between receptor/ligand binding affinity and adhesion strength. *Biophys. J.* **1993**, *65*, 2191–2200. [CrossRef]

66. Kou, S.C.; Hammer, D.A.; Lauffenburger, D.A. Simulation of detachment of specifically bound particles from surfaces by shear flow. *Biophys. J.* **1997**, *73*, 517–531. [CrossRef]

67. Pierres, A.; Benoliel, A.M.; Bongrand, P. Use of a laminar flow chamber to study the rate of bond formation and dissociation between surface-bound adhesion molecules: Effect of applied force and distance between surfaces. *Faraday Discuss.* **1999**, *111*, 321–330. [CrossRef] [PubMed]

68. Park, H.S.; Lee, W.; Nam, Y.S. Elution dynamics of M13 bacteriophage bound to streptavidin immobilized in a microfluidic channel. *BioChip J.* **2016**, *10*, 48–55. [CrossRef]

69. Wang, J.; Liu, Y.; Teesalu, T.; Sugahara, K.N.; Kotamrajua, V.R.; Adams, J.D.; Ferguson, B.S.; Gong, Q.; Oh, S.S.; Csordas, A.T.; et al. Selection of phage-displayed peptides on live adherent cells in microfluidic channels. *Proc. Natl. Acad. Sci. USA* **2011**, *108*, 6909–6914. [CrossRef] [PubMed]

70. Gérard, A.; Woolfe, A.; Mottet, G.; Reichen, M.; Castrillon, C.; Menrath, V.; Ellouze, S.; Poitou, A.; Doineau, R.; Briseno-Roa, L.; et al. High-throughput single-cell activity-based screening and sequencing of antibodies using droplet microfluidics. *Nat. Biotechnol.* **2020**, *38*, 715–721. [CrossRef] [PubMed]

71. Raftery, L.J.; Howard, C.B.; Grewal, Y.S.; Vaidyanathan, R.; Jones, M.L.; Anderson, W.; Korbie, D.; Duarte, T.; Cao, M.D.; Nguyen, S.H.; et al. Retooling phage display with electrohydrodynamic nanomixing and nanopore sequencing. *Lab. Chip* **2019**, *19*, 4083–4092. [CrossRef]

72. Wuethrich, A.; Howard, C.B.; Trau, M. Geometric optimisation of electrohydrodynamic fluid flows for enhanced biosensing. *Microchem. J.* **2018**, *137*, 231–237. [CrossRef]

73. Kamil Reza, K.; Wang, J.; Vaidyanathan, R.; Dey, S.; Wang, Y.; Trau, M. Electrohydrodynamic-Induced SERS Immunoassay for Extensive Multiplexed Biomarker Sensing. *Small* **2017**, *13*, 1602902. [CrossRef]

74. Vaidyanathan, R.; Naghibosadat, M.; Rauf, S.; Korbie, D.; Carrascosa, L.G.; Shiddiky, M.J.A.; Trau, M. Detecting exosomes specifically: A multiplexed device based on alternating current electrohydrodynamic induced nanoshearing. *Anal. Chem.* **2014**, *86*, 11125–11132. [CrossRef] [PubMed]

75. Shiddiky, M.J.A.; Vaidyanathan, R.; Rauf, S.; Tay, Z.; Trau, M. Molecular nanoshearing: An innovative approach to shear off molecules with AC-induced nanoscopic fluid flow. *Sci. Rep.* **2014**, *4*, 3716. [CrossRef] [PubMed]

76. Wang, J.; Kao, Y.C.; Zhou, Q.; Wuethrich, A.; Stark, M.S.; Schaider, H.; Soyer, H.P.; Lin, L.L.; Trau, M. An Integrated Microfluidic-SERS Platform Enables Sensitive Phenotyping of Serum Extracellular Vesicles in Early Stage Melanomas. *Adv. Funct. Mater.* **2022**, *32*, 2010296. [CrossRef]

77. Li, J.; Wuethrich, A.; Edwardraja, S.; Lobb, R.J.; Puttick, S.; Rose, S.; Howard, C.B.; Trau, M. Amplification-Free SARS-CoV-2 detection using nanoyeast-scFv and ultrasensitive plasmonic nanobox-integrated nanomixing microassay. *Anal. Chem.* **2021**, *93*, 10251–10260. [CrossRef]

78. Wang, J.; Wuethrich, A.; Lobb, R.J.; Antaw, F.; Sina, A.A.I.; Lane, R.E.; Zhou, Q.; Zieschank, C.; Bell, C.; Bonazzi, V.F.; et al. Characterizing the heterogeneity of small extracellular vesicle populations in multiple cancer typesviaan ultrasensitive chip. *ACS Sens.* **2021**, *6*, 3182–3194. [CrossRef] [PubMed]

79. Sutera, S.P.; Skalak, R. The history of poiseuille's law. *Annu. Rev. Fluid Mech.* **2003**, *25*, 1–20. [CrossRef]

80. Wang, J.; Guo, J.; Zhao, K.; Ruan, W.; Li, L.; Ling, J.; Peng, R.; Zhang, H.; Yang, C.; Zhu, Z. Auto-panning: A highly integrated and automated biopanning platform for peptide screening. *Lab. Chip* **2021**, *21*, 2702–2710. [CrossRef]

81. Bender, B.J.; Gahbauer, S.; Luttens, A.; Lyu, J.; Webb, C.M.; Stein, R.M.; Fink, E.A.; Balius, T.E.; Carlsson, J.; Irwin, J.J.; et al. A practical guide to large-scale docking. *Nat. Protoc.* **2021**, *16*, 4799–4832. [CrossRef]

82. Bohacek, R.S.; McMartin, C.; Guida, W.C. The art and practice of structure-based drug design: A molecular modeling perspective. *Med. Res. Rev.* **1996**, *16*, 3–50. [CrossRef]

83. Lyu, J.; Wang, S.; Balius, T.E.; Singh, I.; Levit, A.; Moroz, Y.S.; O'Meara, M.J.; Che, T.; Algaa, E.; Tolmachova, K.; et al. Ultra-large library docking for discovering new chemotypes. *Nature* **2019**, *566*, 224–229. [CrossRef]

84. Sadybekov, A.A.; Sadybekov, A.V.; Liu, Y.; Iliopoulos-Tsoutsouvas, C.; Huang, X.-P.; Pickett, J.; Houser, B.; Patel, N.; Tran, N.K.; Tong, F.; et al. Synthon-based ligand discovery in virtual libraries of over 11 billion compounds. *Nature* **2022**, *601*, 452–459. [CrossRef] [PubMed]

85. Sterling, T.; Irwin, J.J. ZINC 15—ligand discovery for everyone. *J. Chem. Inf. Model.* **2015**, *55*, 2324–2337. [CrossRef] [PubMed]

86. Hansch, C.; Fujita, T. p-σ-π Analysis. A method for the correlation of biological activity and chemical structure. *J. Am. Chem. Soc.* **1964**, *86*, 1616–1626. [CrossRef]

87. Cherkasov, A.; Muratov, E.N.; Fourches, D.; Varnek, A.; Baskin, I.I.; Cronin, M.; Dearden, J.; Gramatica, P.; Martin, Y.C.; Todeschini, R.; et al. QSAR modeling: Where have you been? Where are you going to? *J. Med. Chem.* **2014**, *57*, 4977–5010. [CrossRef] [PubMed]

88. Leem, J.; Dunbar, J.; Georges, G.; Shi, J.; Deane, C.M. ABodyBuilder: Automated antibody structure prediction with data–driven accuracy estimation. *mAbs* **2016**, *8*, 1259–1268. [CrossRef] [PubMed]

89. Weitzner, B.D.; Jeliazkov, J.R.; Lyskov, S.; Marze, N.; Kuroda, D.; Frick, R.; Adolf-Bryfogle, J.; Biswas, N.; Dunbrack, R.L.; Gray, J.J. Modeling and docking of antibody structures with Rosetta. *Nat. Protoc.* **2017**, *12*, 401–416. [CrossRef]

90. Swindells, M.B.; Porter, C.T.; Couch, M.; Hurst, J.; Abhinandan, K.R.; Nielsen, J.H.; Macindoe, G.; Hetherington, J.; Martin, A.C.R. abYsis: Integrated antibody sequence and structure—management, analysis, and prediction. *J. Mol. Biol.* **2017**, *429*, 356–364. [CrossRef]

91. Saka, K.; Kakuzaki, T.; Metsugi, S.; Kashiwagi, D.; Yoshida, K.; Wada, M.; Tsunoda, H.; Teramoto, R. Antibody design using LSTM based deep generative model from phage display library for affinity maturation. *Sci. Rep.* **2021**, *11*, 5852. [CrossRef]

92. Fox, R.J.; Davis, S.C.; Mundorff, E.C.; Newman, L.M.; Gavrilovic, V.; Ma, S.K.; Chung, L.M.; Ching, C.; Tam, S.; Muley, S.; et al. Improving catalytic function by ProSAR-driven enzyme evolution. *Nat. Biotechnol.* **2007**, *25*, 338–344. [CrossRef]

93. Liao, J.; Warmuth, M.K.; Govindarajan, S.; Ness, J.E.; Wang, R.P.; Gustafsson, C.; Minshull, J. Engineering proteinase K using machine learning and synthetic genes. *BMC Biotechnol.* **2007**, *7*, 16. [CrossRef] [PubMed]

94. Berman, H.M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T.N.; Weissig, H.; Shindyalov, I.N.; Bourne, P.E. The protein data bank. *Nucleic Acids Res.* **2000**, *28*, 235–242. [CrossRef] [PubMed]

95. Bateman, A.; Martin, M.J.; O'Donovan, C.; Magrane, M.; Alpi, E.; Antunes, R.; Bely, B.; Bingley, M.; Bonilla, C.; Britto, R.; et al. UniProt: The universal protein knowledgebase. *Nucleic Acids Res.* **2004**, *32*, D115–D119. [CrossRef]

96. Kermani, A.A. A guide to membrane protein X-ray crystallography. *FEBS J.* **2021**, *288*, 5788–5804. [CrossRef] [PubMed]

97. Wouters, O.J.; McKee, M.; Luyten, J. Estimated research and development investment needed to bring a new medicine to market, 2009–2018. *JAMA* **2020**, *323*, 844–853. [CrossRef]

98. Brown, D.G.; Wobst, H.J.; Kapoor, A.; Kenna, L.A.; Southall, N. Clinical development times for innovative drugs. *Nat. Rev. Drug Discov.* **2021**. [CrossRef]

99. DeFrancesco, L. COVID-19 antibodies on trial. *Nat. Biotechnol.* **2020**, *38*, 1242–1252. [CrossRef]

100. Hernandez, I.; Bott, S.W.; Patel, A.S.; Wolf, C.G.; Hospodar, A.R.; Sampathkumar, S.; Shrank, W.H. Pricing of monoclonal antibody therapies: Higher if used for cancer? *Am. J. Manag. Care* **2018**, *24*, 109–112.

101. Otwinowski, Z.; Minor, W. Processing of X-ray diffraction data collected in oscillation mode. *Methods Enzymol.* **1997**, *276*, 307–326. [CrossRef]

102. Taylor, K.; Alvarez, L.R. An estimate of the number of animals used for scientific purposes worldwide in 2015. *Altern. Lab. Anim.* **2019**, *47*, 196–213. [CrossRef]

103. Van Norman, G.A. Limitations of animal studies for predicting toxicity in clinical trials: Is it time to rethink our current approach? *JACC Basic Transl. Sci.* **2019**, *4*, 845–854. [CrossRef] [PubMed]

104. Lillie, E.O.; Patay, B.; Diamant, J.; Issell, B.; Topol, E.J.; Schork, N.J. The n-of-1 clinical trial: The ultimate strategy for individualizing medicine? *Pers. Med.* **2011**, *8*, 161–173. [CrossRef] [PubMed]

105. Afghahi, A.; Sledge, G.W. Targeted therapy for cancer in the genomic era. *Cancer J.* **2015**, *21*, 294–298. [CrossRef] [PubMed]

106. Mayer, I.A.; Abramson, V.G.; Lehmann, B.D.; Pietenpol, J.A. New strategies for triple-negative breast cancer-deciphering the heterogeneity. *Clin. Cancer Res.* **2014**, *20*, 782–790. [CrossRef] [PubMed]

107. Füzéry, A.K.; Levin, J.; Chan, M.M.; Chan, D.W. Translation of proteomic biomarkers into FDA approved cancer diagnostics: Issues and challenges. *Clin. Proteomics* **2013**, *10*, 13. [CrossRef]

108. Campbell, P.J.; Getz, G.; Korbel, J.O.; Stuart, J.M.; Jennings, J.L.; Stein, L.D.; Perry, M.D.; Nahal-Bose, H.K.; Ouellette, B.F.F.; Li, C.H.; et al. Pan-cancer analysis of whole genomes. *Nature* **2020**, *578*, 82–93. [CrossRef]

109. Melo, F.D.S.E.; Vermeulen, L.; Fessler, E.; Medema, J.P. Cancer heterogeneity—A multifaceted view. *EMBO Rep.* **2013**, *14*, 686–695. [CrossRef]

110. Wang, X.; Markowetz, F.; De Sousa, E.M.F.; Medema, J.P.; Vermeulen, L. Dissecting cancer heterogeneity—An unsupervised classification approach. *Int. J. Biochem. Cell Biol.* **2013**, *45*, 2574–2579. [CrossRef]

111. Hong, D.; Fritz, A.J.; Zaidi, S.K.; van Wijnen, A.J.; Nickerson, J.A.; Imbalzano, A.N.; Lian, J.B.; Stein, J.L.; Stein, G.S. Epithelial-to-mesenchymal transition and cancer stem cells contribute to breast cancer heterogeneity. *J. Cell. Physiol.* **2018**, *233*, 9136–9144. [CrossRef]

112. Hanash, S.; Taguchi, A. The grand challenge to decipher the cancer proteome. *Nat. Rev. Cancer* **2010**, *10*, 652–660. [CrossRef]

113. Panowski, S.; Bhakta, S.; Raab, H.; Polakis, P.; Junutula, J.R. Site-specific antibody drug conjugates for cancer therapy. *mAbs* **2014**, *6*, 34–45. [CrossRef] [PubMed]

114. Koh, A.J.; Sinder, B.P.; Entezami, P.; Nilsson, L.; McCauley, L.K. The skeletal impact of the chemotherapeutic agent etoposide. *Osteoporos. Int.* **2017**, *28*, 2321–2333. [CrossRef]

115. Woods, D.; Turchi, J.J. Chemotherapy induced DNA damage response convergence of drugs and pathways. *Cancer Biol. Ther.* **2013**, *14*, 379–389. [CrossRef] [PubMed]

116. Nguyen, T.; Tekrony, A.; Yaehne, K.; Cramb, D.T. Designing a better theranostic nanocarrier for cancer applications. *Nanomedicine* **2014**, *9*, 2371–2386. [CrossRef]

117. Diamantis, N.; Banerji, U. Antibody-drug conjugates—An emerging class of cancer treatment. *Br. J. Cancer* **2016**, *114*, 362–367. [CrossRef] [PubMed]

118. Sau, S.; Alsaab, H.O.; Kashaw, S.K.; Tatiparti, K.; Iyer, A.K. Advances in antibody–drug conjugates: A new era of targeted cancer therapy. *Drug Discov. Today* **2017**, *22*, 1547–1556. [CrossRef] [PubMed]

119. Yu, H.; Braun, P.; Yildirim, M.A.; Lemmens, I.; Venkatesan, K.; Sahalie, J.; Hirozane-Kishikawa, T.; Gebreab, F.; Li, N.; Simonis, N.; et al. High-quality binary protein interaction map of the yeast interactome network. *Science* **2008**, *322*, 104–110. [CrossRef]

120. Yao, V.J.; D'Angelo, S.; Butler, K.S.; Theron, C.; Smith, T.L.; Marchiò, S.; Gelovani, J.G.; Sidman, R.L.; Dobroff, A.S.; Brinker, C.J.; et al. Ligand-targeted theranostic nanomedicines against cancer. *J. Control. Release* **2016**, *240*, 267–286. [CrossRef]

121. Suzuki, M.; Kato, C.; Kato, A. Therapeutic antibodies: Their mechanisms of action and the pathological findings they induce in toxicity studies. *J. Toxicol. Pathol.* **2015**, *28*, 133–139. [CrossRef]

122. Yea, K.; Zhang, H.; Xie, J.; Jones, T.M.; Lin, C.W.; Francesconi, W.; Berton, F.; Fallahi, M.; Sauer, K.; Lerner, R.A. Agonist antibody that induces human malignant cells to kill one another. *Proc. Natl. Acad. Sci. USA* **2015**, *112*, E6158–E6165. [CrossRef]

123. Beck, A.; Goetsch, L.; Dumontet, C.; Corvaïa, N. Strategies and challenges for the next generation of antibody-drug conjugates. *Nat. Rev. Drug Discov.* **2017**, *16*, 315–337. [CrossRef]

124. Larson, S.M.; Carrasquillo, J.A.; Cheung, N.K.V.; Press, O.W. Radioimmunotherapy of human tumours. *Nat. Rev. Cancer* **2015**, *15*, 347–360. [CrossRef] [PubMed]

125. Senior, A.W.; Evans, R.; Jumper, J.; Kirkpatrick, J.; Sifre, L.; Green, T.; Qin, C.; Žídek, A.; Nelson, A.W.R.; Bridgland, A.; et al. Improved protein structure prediction using potentials from deep learning. *Nature* **2020**, *577*, 706–710. [CrossRef] [PubMed]

126. Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, A.; Potapenko, A.; et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **2021**, *596*, 583–589. [CrossRef]

127. Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, A.; Potapenko, A.; et al. Applying and improving AlphaFold at CASP14. *Proteins Struct. Funct. Bioinform.* **2021**, *89*, 1711–1721. [CrossRef]

128. Kryshtafovych, A.; Schwede, T.; Topf, M.; Fidelis, K.; Moult, J. Critical assessment of methods of protein structure prediction (CASP)—Round XIV. *Proteins Struct. Funct. Bioinform.* **2021**, *89*, 1607–1617. [CrossRef] [PubMed]

129. Mirdita, M.; Von Den Driesch, L.; Galiez, C.; Martin, M.J.; Soding, J.; Steinegger, M. Uniclust databases of clustered and deeply annotated protein sequences and alignments. *Nucleic Acids Res.* **2017**, *45*, D170–D176. [CrossRef] [PubMed]

130. Baek, M.; DiMaio, F.; Anishchenko, I.; Dauparas, J.; Ovchinnikov, S.; Lee, G.R.; Wang, J.; Cong, Q.; Kinch, L.N.; Dustin Schaeffer, R.; et al. Accurate prediction of protein structures and interactions using a three-track neural network. *Science* **2021**, *373*, 871–876. [CrossRef]

131. Humphreys, I.R.; Pei, J.; Baek, M.; Krishnakumar, A.; Anishchenko, I.; Ovchinnikov, S.; Zhang, J.; Ness, T.J.; Banjade, S.; Bagde, S.R.; et al. Computed structures of core eukaryotic protein complexes. *Science* **2021**, *374*, eabm4805. [CrossRef] [PubMed]

132. Anishchenko, I.; Pellock, S.J.; Chidyausiku, T.M.; Ramelot, T.A.; Ovchinnikov, S.; Hao, J.; Bafna, K.; Norn, C.; Kang, A.; Bera, A.K.; et al. De novo protein design by deep network hallucination. *Nature* **2021**, *600*, 547–552. [CrossRef]

133. Cao, L.; Coventry, B.; Goreshnik, I.; Huang, B.; Park, J.S.; Jude, K.M.; Marković, I.; Kadam, R.U.; Verschueren, K.H.G.; Verstraete, K.; et al. Robust de novo design of protein binding proteins from target structural information alone. *bioRxiv* **2021**. [CrossRef]

134. Callaway, E. 'It will change everything': DeepMind's AI makes gigantic leap in solving protein structures. *Nature* **2020**, *588*, 203–204. [CrossRef] [PubMed]

135. Porta-Pardo, E.; Ruiz-Serra, V.; Valencia, A. The structural coverage of the human proteome before and after AlphaFold. *PLoS Comput. Biol.* **2022**, *18*, e1009818. [CrossRef]

136. Tunyasuvunakool, K.; Adler, J.; Wu, Z.; Green, T.; Zielinski, M.; Žídek, A.; Bridgland, A.; Cowie, A.; Meyer, C.; Laydon, A.; et al. Highly accurate protein structure prediction for the human proteome. *Nature* **2021**, *596*, 590–596. [CrossRef] [PubMed]

137. Vinson, V.J. Proteins in motion. *Science* **2009**, *324*, 197. [CrossRef] [PubMed]

138. Junsu, K.; Lee, J. Can AlphaFold2 predict protein-peptide complex structures accurately. *bioRxiv* **2021**. [CrossRef]

139. Evans, R.; O'Neill, M.; Pritzel, A.; Antropova, N.; Senior, A.W.; Green, T.; Žídek, A.; Bates, R.; Blackwell, S.; Yim, J.; et al. Protein complex prediction with AlphaFold-Multimer. *bioRxiv* **2021**. [CrossRef]

140. Mullard, A. What does AlphaFold mean for drug discovery? *Nat. Rev. Drug Discov.* **2021**, *20*, 725–727. [CrossRef] [PubMed]

141. Smyth, M.S.; Martin, J.H.J. x Ray crystallography. *J. Clin. Pathol. Mol. Pathol.* **2000**, *53*, 8–14. [CrossRef] [PubMed]

142. Akbar, R.; Robert, P.A.; Weber, C.R.; Widrich, M.; Frank, R.; Pavlović, M.; Scheffer, L.; Chernigovskaya, M.; Snapkov, I.; Slabodkin, A.; et al. In silico proof of principle of machine learning-based antibody design at unconstrained scale. *bioRxiv* **2021**. [CrossRef]

143. Patel, L.; Shukla, T.; Huang, X.; Ussery, D.W.; Wang, S. Machine learning methods in drug discovery. *Molecules* **2020**, *25*, 5277. [CrossRef] [PubMed]

144. Akbar, R.; Robert, P.A.; Pavlović, M.; Jeliazkov, J.R.; Snapkov, I.; Slabodkin, A.; Weber, C.R.; Scheffer, L.; Miho, E.; Haff, I.H.; et al. A compact vocabulary of paratope-epitope interactions enables predictability of antibody-antigen binding. *Cell Rep.* **2021**, *34*, 108856. [CrossRef] [PubMed]

145. Braun, R. Systems analysis of high-throughput data. *Adv. Exp. Med. Biol.* **2014**, *844*, 153–187. [CrossRef] [PubMed]