



Cite this: *Chem. Sci.*, 2023, **14**, 1443

All publication charges for this article have been paid for by the Royal Society of Chemistry

## AlphaFold accelerates artificial intelligence powered drug discovery: efficient discovery of a novel CDK20 small molecule inhibitor†

Feng Ren,<sup>a</sup> Xiao Ding,<sup>a</sup> Min Zheng,<sup>a</sup> Mikhail Korzinkin,<sup>b</sup> Xin Cai,<sup>a</sup> Wei Zhu,<sup>a</sup> Alexey Mantsyzov,<sup>b</sup> Alex Aliper,<sup>b</sup> Vladimir Aladinskiy,<sup>b</sup> Zhongying Cao,<sup>a</sup> Shanshan Kong,<sup>a</sup> Xi Long,<sup>b</sup> Bonnie Hei Man Liu,<sup>b</sup> Yingtao Liu,<sup>a</sup> Vladimir Naumov,<sup>b</sup> Anastasia Shneyderman,<sup>b</sup> Ivan V. Ozerov,<sup>b</sup> Ju Wang,<sup>a</sup> Frank W. Pun,<sup>b</sup> Daniil A. Polykovskiy,<sup>b</sup> Chong Sun,<sup>c</sup> Michael Levitt,<sup>d</sup> Alán Aspuru-Guzik<sup>\*c</sup> and Alex Zhavoronkov<sup>ID \*ab</sup>

The application of artificial intelligence (AI) has been considered a revolutionary change in drug discovery and development. In 2020, the AlphaFold computer program predicted protein structures for the whole human genome, which has been considered a remarkable breakthrough in both AI applications and structural biology. Despite the varying confidence levels, these predicted structures could still significantly contribute to structure-based drug design of novel targets, especially the ones with no or limited structural information. In this work, we successfully applied AlphaFold to our end-to-end AI-powered drug discovery engines, including a biocomputational platform PandaOmics and a generative chemistry platform Chemistry42. A novel hit molecule against a novel target without an experimental structure was identified, starting from target selection towards hit identification, in a cost- and time-efficient manner. PandaOmics provided the protein of interest for the treatment of hepatocellular carcinoma (HCC) and Chemistry42 generated the molecules based on the structure predicted by AlphaFold, and the selected molecules were synthesized and tested in biological assays. Through this approach, we identified a small molecule hit compound for cyclin-dependent kinase 20 (CDK20) with a binding constant  $K_d$  value of  $9.2 \pm 0.5 \mu\text{M}$  ( $n = 3$ ) within 30 days from target selection and after only synthesizing 7 compounds. Based on the available data, a second round of AI-powered compound generation was conducted and through this, a more potent hit molecule, ISM042-2-048, was discovered with an average  $K_d$  value of  $566.7 \pm 256.2 \text{ nM}$  ( $n = 3$ ). Compound ISM042-2-048 also showed good CDK20 inhibitory activity with an  $IC_{50}$  value of  $33.4 \pm 22.6 \text{ nM}$  ( $n = 3$ ). In addition, ISM042-2-048 demonstrated selective anti-proliferation activity in an HCC cell line with CDK20 overexpression, Huh7, with an  $IC_{50}$  of  $208.7 \pm 3.3 \text{ nM}$ , compared to a counter screen cell line HEK293 ( $IC_{50} = 1706.7 \pm 670.0 \text{ nM}$ ). This work is the first demonstration of applying AlphaFold to the hit identification process in drug discovery.

Received 14th October 2022  
Accepted 5th January 2023

DOI: 10.1039/d2sc05709c  
[rsc.li/chemical-science](http://rsc.li/chemical-science)

## Introduction

The 3D structures of proteins are highly correlated with their functions in cells and the biological impacts caused by amino

<sup>a</sup>Insilico Medicine Shanghai Ltd, Suite 901, Tower C, Changtai Plaza, 2889 Jinke Road, Pudong New District, Shanghai 201203, China. E-mail: alex@insilico.com

<sup>b</sup>Insilico Medicine Hong Kong Ltd, Unit 310, 3/F, Building 8W, Phase 2, Hong Kong Science Park, Pak Shek Kok, Hong Kong, China

<sup>c</sup>Department of Chemistry, Department of Computer Science, University of Toronto, Vector Institute for Artificial Intelligence, Canadian Institute for Advanced Research, Toronto, Ontario, Canada. E-mail: alan@aspuru.com

<sup>d</sup>Department of Structural Biology, Stanford University, Palo Alto, CA, USA

† Electronic supplementary information (ESI) available: Synthesis of compounds. See DOI: <https://doi.org/10.1039/d2sc05709c>

acid mutations. A protein structure is a versatile tool to study the gene-disease association and mode of action (MoA), to evaluate the druggability of a therapeutic target. Structure-based drug discovery (SBDD) has been a mainstay method to identify hit molecules and perform lead optimization, which requires the 3D structure of a target.<sup>2-4</sup> After an endeavor spanning decades, only a small fraction of known proteins have experimentally determined structures. Accurate protein structure prediction has been a longstanding challenge until the appearance of AlphaFold at CASP14.<sup>5</sup> The structures predicted by AlphaFold can reach an accuracy level comparable to those of experimental methods.<sup>6,7</sup> The scientific community celebrated DeepMind's accomplishment<sup>8-12</sup> and the release of proteome-wide AlphaFold DB,<sup>11</sup> which has



now expanded to contain over 804 000 protein structures covering 21 species.<sup>13</sup> Although the protein models predicted by AlphaFold have variable qualities from good, bad to ugly,<sup>10</sup> the predicted local distance difference test score is provided as a confidence metric to guide the usage of 3D structures produced by AlphaFold. AlphaFold models have been used to aid the determination of experimental structures by crystallography<sup>14</sup> and cryo-EM,<sup>15</sup> to guide the functional study of PINK1,<sup>16</sup> to help identify pathogenic mutations,<sup>17,18</sup> and to explore the protein–protein interaction.<sup>19</sup> Public databases include AlphaFold models as references, *e.g.*, UniProt,<sup>20</sup> the therapeutic target database,<sup>21</sup> and APPRIS.<sup>22</sup> The methodology of AlphaFold has inspired RoseTTAFold,<sup>23</sup> a potentially faster and cheaper protein prediction tool with adequate accuracy, and AlphaDesign,<sup>24</sup> a protein design framework. AI-powered protein prediction has been selected as one of the 2021 breakthroughs by both *Science*<sup>25</sup> and *Nature*<sup>26</sup> journals.

In this work, we rapidly identify *de novo* molecules for a novel target by combining the protein structure predicted by AlphaFold with the end-to-end AI-powered drug discovery platforms PandaOmics and Chemistry42. The process embarked on indication, target selection, hit generation and hit identification.<sup>27</sup> While we were aware of the capabilities of AlphaFold2 applied to the scientific community, the application and modification of the algorithm for commercial purposes are still poorly understood. Here, we used the freely-available predicted structures from the AlphaFold DB repository as a starting point.

The general workflow is described in Fig. 1 where hepatocellular carcinoma (HCC) was used as the indication of interest due to its high prevalence in liver cancers and lack of effective treatments. In general, by analysis of text and OMICS data from 10 datasets for HCC, PandaOmics provided a list of the top 20 targets. Afterwards multidimensional filtration was applied including novelty, accessibility by biologics, safety, small molecule accessibility, and tissue specificity. Cyclin-dependent kinase 20 (CDK20) was finally selected as our initial target to work on due to its strong disease association, limited experimental structure information, and shortage of approved drugs or clinical compounds in the context of any disease during the last 3 years. Through Chemistry42 structure-based compound generation using the AlphaFold predicted CDK20 structure, 8918 molecules were generated and, after molecular docking and clustering, 7 were selected for synthesis and biological testing. Among them, compound ISM042-2-001 demonstrated a  $K_d$  value of  $9.2 \pm 0.5 \mu\text{M}$  ( $n = 3$ ) in CDK20 kinase binding assay. Empowered by Chemistry42 and AlphaFold predicted protein structures, it took us only 30 days to discover our first hit. The predicted binding mode was then used as guidance for the second-round compound generation, synthesis, and testing, which resulted in a more potent hit molecule ISM042-2-048 with nanomolar potency. To the best of our knowledge, this work is the first reported example that successfully utilized AlphaFold-predicted protein structures to identify a confirmed hit for a novel target in early drug discovery.

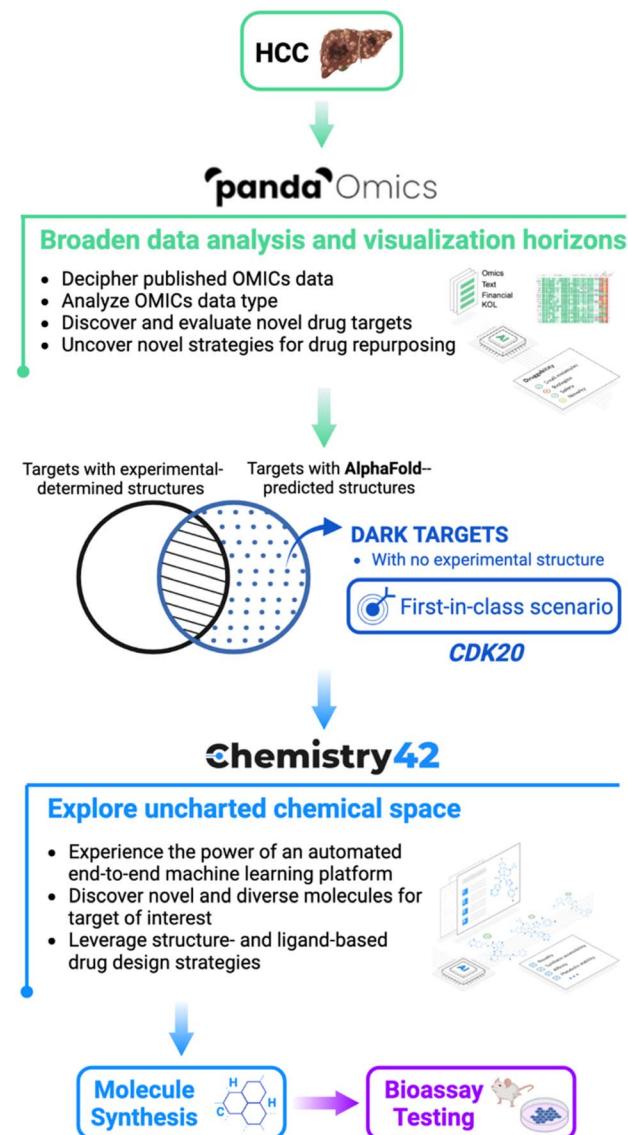


Fig. 1 The pipeline to combine AlphaFold with Insilico Medicine end-to-end, and AI-powered drug discovery platforms PandaOmics and Chemistry42 in the drug discovery for hepatocellular carcinoma from target selection and hit generation to hit identification. A novel therapeutic target was identified from a pool of dark targets that have AlphaFold-predicted but lack experimentally determined structures. Such a target represents a first-in-class novel target which is revealed for the first time to treat HCC.

## Target selection and identification

Primary liver cancer is the sixth most frequently occurring cancer and the third most common cause of worldwide cancer mortality according to the GLOBOCAN 2020 update released by the International Agency for Research on Cancer (IARC). Hepatocellular carcinoma (HCC) is the dominant type of liver cancer, accounting for approximately 75% of the total patient population. The incidence rate of liver cancer is very close to its mortality rate due to very poor prognosis in all regions around the world.<sup>28</sup> PD-L1 inhibitor atezolizumab in combo with bevacizumab has become



the new standard-of-care (SoC) first-line treatment for advanced HCC after demonstrating improvement in both 12 months overall survival (OS) and progression-free survival (PFS) compared to the previous SoC Nexavar,<sup>29</sup> but there's still a huge unmet medical need for HCC patients.

PandaOmics is an automated drug discovery AI engine to accelerate and optimize key steps in the early stages of drug discovery.<sup>30</sup> This biocomputational platform combines bioinformatics methods for data analysis, visualization and interpretation with advanced multimodal deep learning approaches for target identification.<sup>31–35</sup> The PandaOmics therapeutic target and biomarker identification system is based on the combination of multiple scores derived from text and OMICs data associating genes with a disease of interest. Text evidence prioritization (text, financial and key opinion leader (KOL) score families) singles out the genes, which are extensively mentioned across the scientific literature and grant description. OMICs-based scores, in contrast, explore the molecular connection of genes with diseases based on differential expression, gene variants, interactome topology, signaling pathway perturbation analysis algorithms,<sup>36</sup> knockout/overexpression experiments and more. This approach allows users to unveil hidden hypotheses that might not be obvious from common general knowledge or simple bioinformatics analysis. AI tools are extremely helpful for efficient target hypothesis generation. The overall scoring approach results in a ranked list of target hypotheses for a given disease which can be subsequently filtered according to their novelty, accessibility by small molecules and antibodies, safety, tissue specificity, crystal structure availability and major biological structures.<sup>33,34</sup>

Another unique feature of the PandaOmics platform is its ability to combine the data from different experiments into a single meta-analysis and leverage the insights from all the datasets together for precise target prioritization. We created a meta-analysis for each of the diseases of interest composed of 10 datasets for HCC (1133 disease samples and 674 healthy controls). After obtaining the ranked list of target hypotheses we applied PandaOmics filters in order to get a list of the most promising targets that satisfy the first-in-class scenario (see the Methods section) characteristics and share the current unavailability of crystal structures but have structure folds predicted by AlphaFold. The final list of top 20 targets was then manually curated as the most promising candidates. For the HCC case, CDK20 was chosen due to its highest scores aligned with the first-in-class scenario. The proposed therapeutic target CDK20 was passed to the Chemistry42 platform for the automated generation of small molecule inhibitors.

## CDK20 as a promising target for cancer treatment

CDK20, also known as cell cycle-related kinase (CCRK), is the latest identified member of the cyclin-dependent kinase family, which has attracted great attention in recent years due to its functions (both cell cycle-dependent and -independent) in a variety of human tissues.<sup>37</sup> CDK20 is widely expressed at a comparable translational level in many human tissues

including the brain, lung, liver, pancreas, and gastrointestinal tract.<sup>38</sup> More importantly, increasing preclinical evidence suggested that CDK20 is overexpressed in many tumor cell lines including tumor samples from patients with different types of cancer, such as colorectal cancer, hepatocellular carcinoma (HCC), lung cancer, and ovarian carcinoma.<sup>39–42</sup> *In vitro* studies showed that the androgen receptor (AR), CDK20, and  $\beta$ -catenin constitute a positive feedback circuit to promote cell cycle progression in HCC cells, and CDK20 overexpression frequently correlates with ectopic expression of the AR and  $\beta$ -catenin in primary HCC tissue samples and with tumor staging and poor overall survival of patients.<sup>40</sup> In lung cancer cells, CDK20 competes with nuclear factor erythroid 2-related factor 2 (NRF2) for kelch-like ECH associated protein 1 (KEAP1) binding, which prevents degradation of NRF2 and enhances its transcriptional activity, therefore lowering the cellular reactive oxygen species (ROS) level. Moreover, CDK20 depletion in lung cancer cells demonstrates impaired cell proliferation, defective G2/M arrest, and increased radiochemosensitivity.<sup>41</sup> In addition to its protumorigenic role through modulation of the cell cycle and oncogenic signaling, CDK20 is also involved in immunosuppression in certain types of tumors. Zhou *et al.* reported that by activating the EZH2-NF- $\kappa$ B pathway, CDK20 expressed in HCC cells increased IL-6 production and induced immunosuppressive MDSC expansion from human peripheral blood mononuclear cells; inhibition of tumorous CDK20 increased IFN- $\gamma$  + TNF- $\alpha$  + CD8 $^+$  T cell infiltration and upregulated PD-L1 expression level in tumors, providing a greater chance of combination therapy with PD-L1 blockade to eradicate HCC tumors.<sup>43</sup> Hence emerging scientific evidence suggested that CDK20 inhibition could be considered a promising therapeutic approach for cancer treatment, especially for HCC.

## Generation of novel hits targeting CDK20 by using AlphaFold predicted structures

As of today, there are limited CDK20 inhibitors reported (displayed in Fig. 2) despite great success being achieved with inhibitors against other members of the CDK family. One possible reason is that there is no available 3D-structure information for this target. In 2005, Nikolai and co-workers described a potential CDK20 inhibitor named RGB-286147 without reporting any binding affinity data.<sup>44</sup> Eurofins disclosed BMS-357075 as a CDK20 binder with a Kd value of 56 nM.<sup>45</sup> AAPK-25 was also reported as a CDK20 inhibitor with a Kd value of 8020 nM.<sup>46</sup> Moreover, Mueller *et al.* described several potential CDK20 inhibitors in their recent poster, including Palbociclib, Flavopiridol, Dinaciclib, and Roscovitine with IC<sub>50</sub> values ranging from 1260 nM to 8680 nM.<sup>47</sup> Besides, they also identified another potent CDK20 inhibitor MER-128 with an IC<sub>50</sub> value of 2 nM; however, the structure of this molecule was not disclosed.<sup>47</sup> Fig. 3 describes our generative procedures for the identification of CDK20 inhibitors starting from structure extraction to hit generation through the SBDD approach by utilizing Chemistry42.<sup>27,48–55</sup>



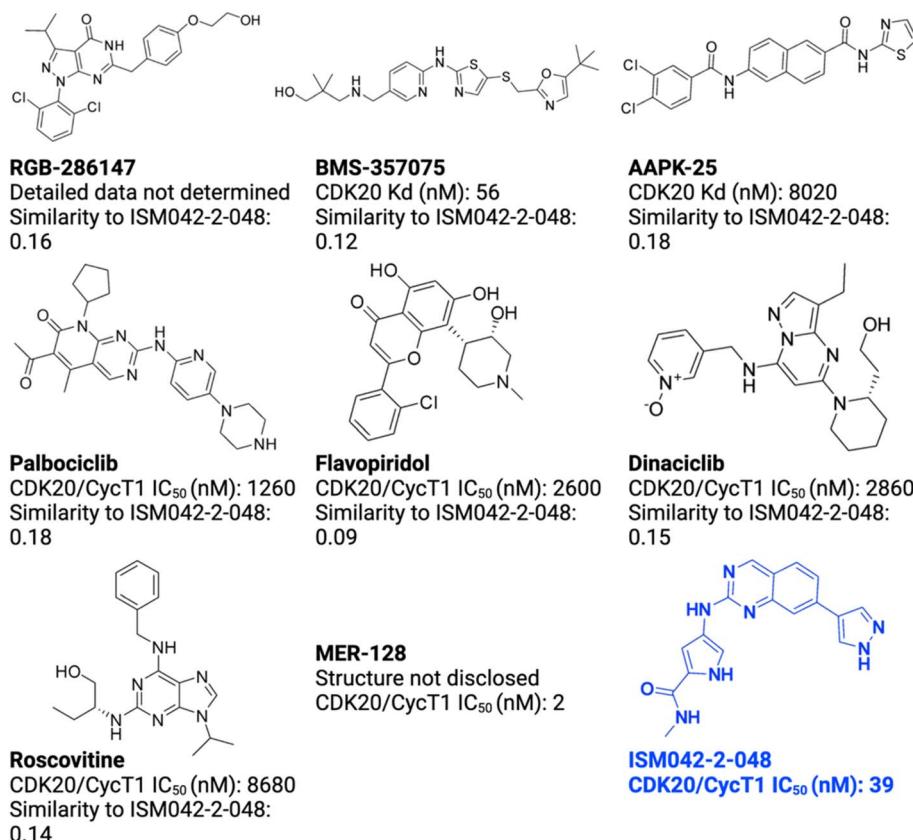


Fig. 2 Reported CDK20 inhibitors from the literature and the novel inhibitor ISM042-2-048 discovered in this paper. The Tanimoto similarities of reported molecules to ISM042-2-048 are calculated from Morgan fingerprints using RDKit.<sup>1</sup>



Fig. 3 In silico Medicine generative procedures for CDK20 hits.

After uploading a protein structure to the Chemistry42 platform, the built-in energy-based approach is automatically used in order to determine putative binding sites. The surface of the protein is evenly covered with probes (methyl), and the energy of non-covalent interactions with the receptor atoms is calculated for each probe. Probes having energy lower than a user-defined threshold are clustered into separate pockets. Each identified cavity is scored using pocket volume, surface, and depth descriptors. Based on these descriptors, Chemistry42 provides a list of identified binding sites.

The CDK20 structure predicted by AlphaFold (AF-Q8IZL9-F1-model\_v1) has a high confidence level overall except for the C-terminal as displayed in Fig. 4A. The C-terminal conformation in the AlphaFold predicted structure with very low confidence blocks the solvent-exposed region of the protein and the residue

Arg305 on the C-terminal occupies the ATP pocket as shown in Fig. 4A. The C-terminal has a flexible loop that has various conformations. The C-terminal in the AlphaFold predicted structure is not in a favorable conformation for the design of an inhibitor by occupying the ATP pocket. Hence the C-terminal (Pro303-Gly346) is removed and only the structure from residue Met1 to residue Ile302 is used as an input for molecule generation in Chemistry42. For this structural model, Chemistry42 identified a shallow ATP binding pocket with an estimated volume of around 150 Å<sup>3</sup> as shown in Fig. 4B. Near the hinge residue Met84, residue Phe81 occupies the gatekeeper and stops a ligand from reaching the back pocket. The predicted binding pocket has a DFG-in conformation and two acidic centers (Asp87 and Glu90) in the solvation region. A pocket-based generation approach was then utilized to generate



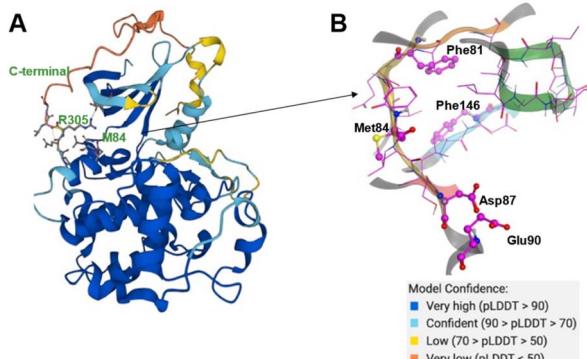


Fig. 4 (A) The AlphaFold predicted structure of CDK20 (AF-Q8IZL9-F1-model\_v1); (B) ATP pocket of CDK20 with a DFG-in (residue Phe146) conformation. Met84 is the hinge residue. P-loop is colored in green. Two acid centers Asp87 and Glu90 are located in the solvent-exposed region of the protein.

novel molecule structures. The hinge residue Met84 is defined as the required binding point. Other 3D structural information from the ATP pocket has been used to guide the generation of molecules towards better fitting the targeted pocket, such as the 3D shape of the pocket, the pocket volume, and the spatial arrangements of atoms in the pocket. In total 8918 molecules were designed by Chemistry42. After molecular docking and clustering, 54 molecules with diverse hinge core structures were prioritized and 7 compounds were selected for synthesis.

## Results and discussion

Fig. 5 shows the chemical structures for the 7 compounds selected to synthesize and to assess the binding abilities towards CDK20. Among the selected compounds, one

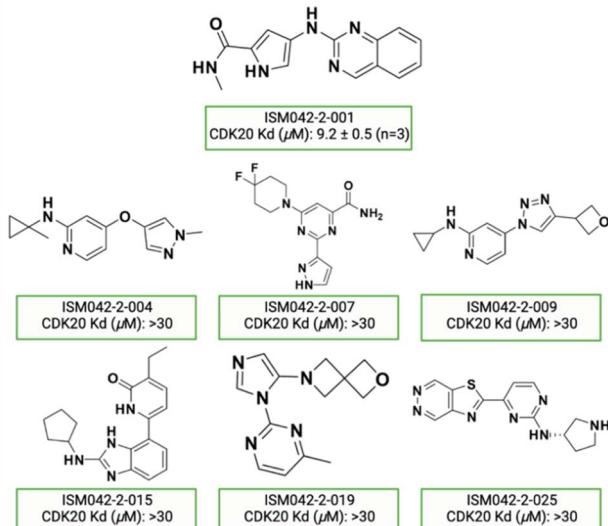


Fig. 5 Chemical structures for the selected 7 molecules from the first-round Chemistry42 generation for synthesis and testing in CDK20 binding assay.

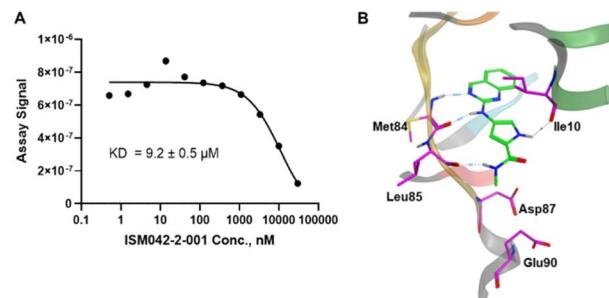


Fig. 6 (A) Representative binding affinity curve of ISM042-2-001 in CDK20 kinase binding assay. Data points are presented as the mean of duplicate wells in one experiment. Similar results were obtained in three independent experiments and the  $K_D$  is the mean  $\pm$  SD of three independent experiments. (B) Predicted binding pose for ISM042-2-001 with CDK20.

compound ISM042-2-001 demonstrated a  $K_D$  value of  $9.2 \pm 0.5 \mu\text{M}$  ( $n = 3$ , one representative binding curve is shown in Fig. 6A) in CDK20 kinase binding assay and an  $IC_{50}$  value of  $>6000 \text{ nM}$  in CDK20 kinase activity assay. It took us only 30 days to discover the hit molecule. We also proposed the binding mode for ISM042-2-001 *via* molecular docking as depicted in Fig. 6B: the four hydrogen bond interactions are represented as dashed lines. Besides the two hydrogen bonds formed with the hinge residue Met84, ISM042-2-001 also interacts with residue Leu85 *via* the amide-NH group and with residue Ile10 in the P-loop *via* the pyrrole-NH group. Alternatively, the amide-NH group or the pyrrole-NH group may form hydrogen bonds with the two acid centers Asp87 and Glu90 in the solvation region.

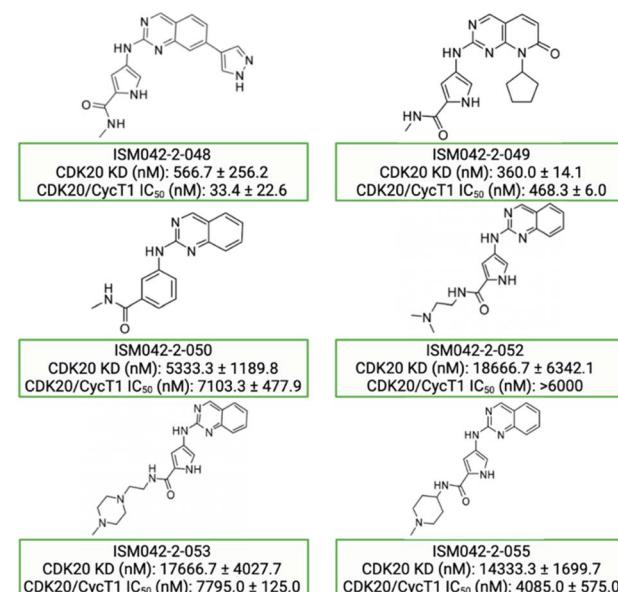


Fig. 7 Chemical structures for the second-round Chemistry42 generation for synthesis and testing in CDK20 binding and kinase activity assays. Biological data with a standard deviation are presented from three independent experiments.



Based on the predicted binding pose and potency data, we conducted a second round of compound generation utilizing our generative AI tool Chemistry42. 16 novel molecules were generated aiming to improve the binding affinity based on two approaches: (1) occupation of the hydrophobic pocket near the gatekeeper region with functional groups on the quinazoline ring; (2) modifications of the pyrrole-2-carboxamide group to access the solvation region and interact with acidic residues Asp87 or Glu90. With the above strategies, 6 out of 16 generated molecules were synthesized and tested as shown in Fig. 7, of which ISM042-2-048 and ISM042-2-049 displayed 15 and 24 fold improvements in binding affinity compared to ISM042-2-001, with measured  $K_d$  values of  $566.7 \pm 256.2$  nM and  $360.0 \pm 14.1$  nM, respectively. A predicted binding mode of ISM042-2-048 with CDK20 was shown in Fig. 8B. Based on the proposed binding mode, in addition to the interactions in the hinge and solvent areas, the pyrazole group of ISM042-2-048 forms a hydrogen bond with residue Lys33, which explains the significant improvement of its binding affinity. ISM042-2-048 is different from the reported CDK20 inhibitors with a novel scaffold and low molecular similarity as shown in Fig. 2. Furthermore, the inhibition of CDK20 kinase activity by ISM042-2-048 was confirmed with an average  $IC_{50}$  of  $33.4 \pm 22.6$  nM ( $n = 3$ ) and showed selective anti-proliferation activity in Huh7 ( $IC_{50} = 208.7$  nM) and

$\pm 3.3$  nM), an HCC cell line with CDK20 overexpression, compared to a counter screen cell line HEK293 ( $IC_{50} = 1706.7 \pm 670.0$  nM), as displayed in Fig. 9. Next-round of optimization will be initiated soon to further improve potency, and the ADME properties and kinase selectivity will also be evaluated.

## Conclusions

Structure-based drug discovery (SBDD) has been a mainstay method to identify hit molecules and perform lead optimization. And the predicted protein structure by AlphaFold has been considered a powerful tool to identify hits for novel targets with no or limited structure information. Herein, we present an example of rapid identification of a CDK20 hit molecule by using AlphaFold predictions as inputs to our automated drug discovery AI engines PandaOmics and Chemistry42 within 30 days covering target selection, molecule generation, compound synthesis and biological testing. Among the 7 compounds synthesized, ISM042-2-001 demonstrated a  $K_d$  value of  $9.2 \pm 0.5$   $\mu$ M ( $n = 3$ ) in CDK20 kinase binding assay. Based on the preliminary SAR, a second round of AI-powered compound generation was conducted and 6 more compounds were synthesized and tested within 30 days from the discovery of the first hit ISM042-2-001. To our delight, a more potent hit molecule, ISM042-2-048, was discovered with an average  $K_d$  value of  $566.7 \pm 256.2$  nM ( $n = 3$ ) and an average  $IC_{50}$  value of  $33.4 \pm 22.6$  nM ( $n = 3$ ). Furthermore, ISM042-2-048 also demonstrated good anti-proliferation activity in an HCC cell line Huh7 with high expression levels of CDK20 ( $IC_{50} = 208.7 \pm 3.3$  nM) whereas less effect was seen in the counter screen cell line HEK293 ( $IC_{50} = 1706.7 \pm 670.0$  nM). This preliminary result indicated that our CDK20 inhibitor didn't induce indiscriminate cyto-toxicity but rather showed a stronger effect on CDK20-overexpressing HCC cells and therefore could serve as a tool molecule to evaluate biological functions for this target. Further optimization of this molecule as well as the evaluation of ADME properties are in progress. Moreover, this work represents the first example of successfully utilizing AlphaFold predicted protein structures for hit identification for a novel target. Further applications of this approach to other target classes such as GPCR and E3 ligase are ongoing.

## Materials and methods

### Target ID and target proposal

The PandaOmics platform was used to conduct hypothesis generation for hepatocellular carcinoma, limiting the target list to the proteins whose structures were predicted by AlphaFold2. HCC analysis combined the data from ten experiments: GSE36376,<sup>56</sup> GSE107170,<sup>56,57</sup> GSE102079,<sup>58</sup> GSE45267,<sup>58,59</sup> GSE133039,<sup>60</sup> GSE104766,<sup>61</sup> GSE77314,<sup>61,62</sup> GSE60502,<sup>63</sup> E-MTAB-5905<sup>64</sup> and TCGA-LIHC,<sup>65</sup> resulting in 1133 disease samples and 674 healthy controls. Targets are proposed by following our first-in-class scenario. The first-in-class scenario is defined as follows: the protein is druggable by small molecules, the target is considered novel by the PandaOmics system, and the target was not in phase 1 clinical trials in the context of any disease

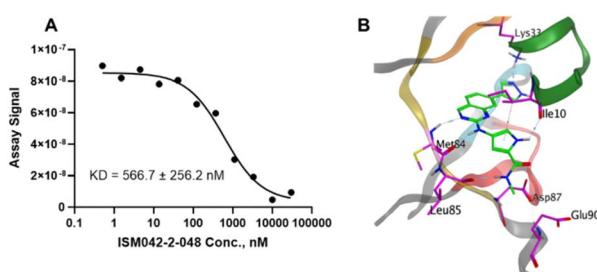


Fig. 8 (A) Representative binding affinity curve for ISM042-2-048 in CDK20 kinase binding assay. Data points are presented as the mean of duplicate wells in one experiment. Similar results were obtained in three independent experiments and the  $K_D$  is the mean  $\pm$  SD of the three independent experiments. (B) Predicted binding pose for ISM042-2-048 in CDK20.

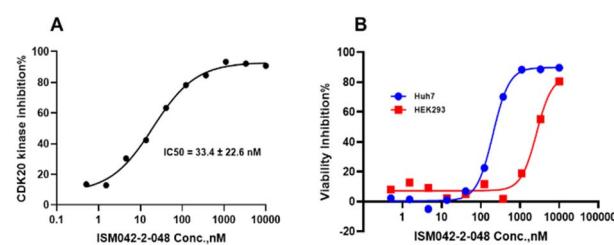


Fig. 9 (A) Representative dose-response curve for ISM042-2-048 in CDK20 kinase activity assay. Similar results were obtained in three independent experiments and the  $IC_{50}$  is the mean  $\pm$  SD of the three independent experiments. (B) Cell viability curves for ISM042-2-048 in cell line Huh7 and counter screen cell line HEK293. Data points are presented as the mean of duplicate wells in one experiment. Similar results were obtained in three independent experiments.



during the last 3 years and is not a target of previously approved drugs.

### CDK20 human CMGC kinase binding assay

The assay was available as a product of the KINOMEscan service from DiscoverX/Eurofins. In brief, CDK20 proteins were produced in HEK-293 cells and subsequently tagged with DNA for qPCR detection. Streptavidin-coated magnetic beads were treated with biotinylated small molecule ligands for 30 minutes at room temperature to generate affinity resins. The liganded beads were blocked with excess biotin and washed with blocking buffer (SeaBlock (Pierce), 1% BSA, 0.05% Tween 20, and 1 mM DTT) to remove the unbound ligand and to reduce non-specific binding. Binding reactions were assembled by combining kinases, liganded affinity beads, and test compounds in 1× binding buffer (20% SeaBlock, 0.17× PBS, 0.05% Tween 20, and 6 mM DTT). Test compounds were prepared as 111X stocks in 100% DMSO. Binding constants (Kds) were determined using an 11-point 3-fold compound dilution series with three DMSO control points. All compounds for Kd measurements are distributed by acoustic transfer (non-contact dispensing) in 100% DMSO. The compounds were then diluted directly into the assays such that the final concentration of DMSO was 0.9%. All reactions were performed in polypropylene 384-well plates. Each has a final volume of 0.02 ml. The assay plates were incubated at room temperature with shaking for 1 hour and the affinity beads were washed with washing buffer (1× PBS and 0.05% Tween 20). The beads were then re-suspended in elution buffer (1× PBS, 0.05% Tween 20, and a 0.5  $\mu$ M non-biotinylated affinity ligand) and incubated at room temperature under shaking for 30 minutes. The kinase concentration in the eluates was measured by qPCR. Kds were calculated with a standard dose-response curve. The curves were fitted using a non-linear least squares fit with the Levenberg–Marquardt algorithm.

### CDK20 kinase activity assay

A radiometric protein kinase assay (<sup>33</sup>PanQinase<sup>®</sup> ActivityAssay, available as a service product from Reaction Biology Corp.) was used for measuring the kinase activity of the CDK20 kinases. This assay was performed in 96-well FlashPlates<sup>TM</sup> from PerkinElmer (Boston, MA, USA) in a 50  $\mu$ l reaction volume. The reaction cocktail was pipetted in four steps in the following order: (a) 25  $\mu$ l of assay buffer (standard buffer/[ $\gamma$ -<sup>33</sup>P]-ATP); (b) 10  $\mu$ l of ATP solution (in H<sub>2</sub>O); (c) 5  $\mu$ l of test compound (in 10% DMSO); (d) 10  $\mu$ l of enzyme/substrate mixture. The assay for CDK20 kinase contained 70 mM HEPES-NaOH pH 7.5, 3 mM MgCl<sub>2</sub>, 3 mM MnCl<sub>2</sub>, 3  $\mu$ M Na-orthovanadate, 1.2 mM DTT, 50  $\mu$ g ml<sup>-1</sup> PEG<sub>20000</sub>, 1.0  $\mu$ M ATP [ $\gamma$ -<sup>33</sup>P]-ATP (approx. 7  $\times$  10<sup>5</sup> cpm per well), and 200 ng/50  $\mu$ l kinase protein, and the substrate was 4.0  $\mu$ g/50  $\mu$ l. The compounds were dissolved to 1  $\times$  10<sup>-3</sup> M in volumes of 100% DMSO. The 1  $\times$  10<sup>-3</sup> M stock solutions were subjected to a serial, semi-logarithmic dilution using 100% DMSO as a solvent. The final volume of the assay was 50  $\mu$ L All compounds were tested at 10 final assay concentrations in the range from 1  $\times$  10<sup>-5</sup> M to 3  $\times$  10<sup>-10</sup> M. The final DMSO concentration in the reaction cocktails was 1% in all cases. The

“low control” was defined as the value that reflects unspecific binding of radioactivity to the plate in the absence of a protein kinase but in the presence of the substrate. The “high control” was defined as the full activity in the absence of any inhibitor. The difference between high and low controls was taken as 100% activity. As part of the data evaluation the low control value from a particular plate was subtracted from the high control value as well as from all 80 “compound values” of the corresponding plate.

The residual activities for each concentration and the compound’s IC<sub>50</sub> values were calculated using Quattro Workflow V3.1.1 (Quattro Research GmbH, Munich, Germany; <http://www.quattro-research.com/>). The fitting model for the IC<sub>50</sub> determinations was “sigmoidal response (variable slope)” with parameters “top” fixed at 100% and “bottom” at 0%. The fitting method used was a least-squares fit.

### Cell viability assay

The Huh7 and HEK293 cells were maintained in DMEM with 10% FBS and 1% streptomycin and penicillin in 5% CO<sub>2</sub>. When the confluence of these two cell lines reached 80–90%, cells were harvested and resuspended, and the proper cells were added into a 384-well plate as below: Huh-7: 650 cells per well; HEK-293 : 300 cells per well. The final testing concentrations of the compounds were: 10 000, 3333, 1111, 370, 123, 41.2, 13.7, 4.57, 1.52, and 0.51 nM. The cells were incubated in a 37 °C, 5% CO<sub>2</sub> incubator for 3 days, and then a reagent was added (Cell-titer Glo assay kit) to the cells for testing and a Multiplate reader was used to record the chemiluminescence value. And GraphPad Prism 9 software was used to calculate IC<sub>50</sub> and plot the effect-dose curve of compounds. The assays were conducted at Pharmaron Inc. in a fee-for-service mode.

### Code availability

Two products were used in this study: PandaOmics and Chemistry42. These platforms are commercially available online at <https://www.pandaomics.com/> and <https://www.chemistry42.com/> correspondingly. Several algorithms used on the platform are publicly available, including the IPANDA algorithm used in PandaOmics (<https://github.com/insilicomedicine/ipanda>), GENTRL model used as part of Chemistry42 (<https://github.com/insilicomedicine/GENTRL>), and VAE-TRIP used in Chemistry42 (<https://github.com/insilicomedicine/TRIP>). Demo of both platforms can be requested on the platforms under the “request demo” link.

### Data availability

AlphaFold structures are available online at <https://alphafold.ebi.ac.uk/>. Gene series data are publicly available at the GEO database (<https://www.ncbi.nlm.nih.gov/geo/>), ArrayExpress (<https://www.ebi.ac.uk/arrayexpress/>), and the Cancer Genome Atlas (TCGA, <https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga>).



## Author contributions

F. R. led the project. X. D. conducted compound selection and synthesis and drafted the initial version of the manuscript. M. Z. performed compound generation and docking. M. K. performed target identification. X. C. conducted biology and assay development and execution. W. Z. performed compound selection and synthesis. A. M. provided support to Chemistry42 for compound generation. A. A. developed Chemistry42 for compound generation. V. A. generated compounds from Chemistry42. Z. C. developed and executed assays. S. K. performed target selection and identification. X. L. and B. H. M. L. drew figures. Y. L. conducted compound docking. V. N., A. S. and I. V. O. performed target selection and identification. J. W. performed target identification and developed the assays. F. W. P. drew figures. D. A. P. provided support to Chemistry42 for compound generation. C. S. wrote, reviewed and edited the manuscript. M. L. and A. A.-G. conceived the project. A. Z. conceived and drove the project.

## Conflicts of interest

Insilico Medicine is a company developing an AI-based end-to-end integrated pipeline for drug discovery and development and engaged in aging and cancer research. Alán Aspuru-Guzik is co-founder and Chief Vision officer of Kebotix, an AI-powered materials and molecular discovery company and co-founder and Chief Scientific Officer of Zapata Computing, a quantum software computing company. Alán Aspuru-Guzik is a scientific advisor to Insilico Medicine. Michael Levitt is an advisor to Insilico Medicine.

## Acknowledgements

Alán Aspuru-Guzik would like to thank the Canada 150 Research Chairs Program for their generous support, as well as Anders G. Frøseth.

## References

- 1 RDKit, *RDKit: Open-Source Cheminformatics Software*, 2022, <https://rdkit.org/>.
- 2 M. Batoor, B. Ahmad and S. Choi, *Int. J. Mol. Sci.*, 2019, **20**(11), 2783.
- 3 K. Nyiri, G. Koppány and B. G. Vertessy, *Cancer Metastasis Rev.*, 2020, **39**, 1091–1105.
- 4 J. J. Marineau, K. B. Hamman, S. Hu, S. Alnemy, J. Mihalich, A. Kabro, K. M. Whitmore, D. K. Winter, S. Roy, S. Ciblat, N. Ke, A. Savinainen, A. Wilsily, G. Malojcic, R. Zahler, D. Schmidt, M. J. Bradley, N. J. Waters and C. Chuaqui, *J. Med. Chem.*, 2022, **65**, 1458–1480.
- 5 J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Zidek, A. Potapenko, A. Bridgland, C. Meyer, S. A. A. Kohl, A. J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A. W. Senior, K. Kavukcuoglu, P. Kohli and D. Hassabis, *Proteins*, 2021, **89**, 1711–1721.
- 6 J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Zidek, A. Potapenko, A. Bridgland, C. Meyer, S. A. A. Kohl, A. J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A. W. Senior, K. Kavukcuoglu, P. Kohli and D. Hassabis, *Nature*, 2021, **596**, 583–589.
- 7 R. Evans, M. O'Neill, A. Pritzel, N. Antropova, A. Senior, T. Green, A. Žídek, R. Bates, S. Blackwell, J. Yim, O. Ronneberger, S. Bodenstein, M. Zielinski, A. Bridgland, A. Potapenko, A. Cowie, K. Tunyasuvunakool, R. Jain, E. Clancy, P. Kohli, J. Jumper and D. Hassabis, *bioRxiv* 2022, preprint, DOI: [10.1101/2021.10.04.463034](https://doi.org/10.1101/2021.10.04.463034).
- 8 M. Akdel, D. E. V. Pires, E. P. Pardo, J. Jänes, A. O. Zalevsky, B. Mészáros, P. Bryant, L. L. Good, R. A. Laskowski, G. Pozzati, A. Shenoy, W. Zhu, P. Kundrotas, V. R. Serra, C. H. M. Rodrigues, A. S. Dunham, D. Burke, N. Borkakoti, S. Velankar, A. Frost, K. Lindorff-Larsen, A. Valencia, S. Ovchinnikov, J. Durairaj, D. B. Ascher, J. M. Thornton, N. E. Davey, A. Stein, A. Elofsson, T. I. Croll and P. Beltrao, *bioRxiv*, 2021, preprint, DOI: [10.1101/2021.09.26.461876](https://doi.org/10.1101/2021.09.26.461876).
- 9 A. Perrakis and T. K. Sixma, *EMBO Rep.*, 2021, **22**, e54046.
- 10 J. M. Thornton, R. A. Laskowski and N. Borkakoti, *Nat. Med.*, 2021, **27**, 1666–1669.
- 11 M. Varadi, S. Anyango, M. Deshpande, S. Nair, C. Natassia, G. Yordanova, D. Yuan, O. Stroe, G. Wood, A. Laydon, A. Zidek, T. Green, K. Tunyasuvunakool, S. Petersen, J. Jumper, E. Clancy, R. Green, A. Vora, M. Lutfi, M. Figurnov, A. Cowie, N. Hobbs, P. Kohli, G. Kleywegt, E. Birney, D. Hassabis and S. Velankar, *Nucleic Acids Res.*, 2022, **50**, D439–D444.
- 12 Y. Zhang, P. Li, F. Pan, H. Liu, P. Hong, X. Liu and J. Zhang, *bioRxiv*, 2021, preprint, DOI: [10.1101/2021.11.03.467194](https://doi.org/10.1101/2021.11.03.467194).
- 13 EMBI-EBI, *AlphaFold Protein Structure Database*, 2022, <https://wwwalphafold.ebi.ac.uk/>.
- 14 T. G. Flower and J. H. Hurley, *Protein Sci.*, 2021, **30**, 728–734.
- 15 M. F. Peter, P. Depping, N. Schneberger, E. Severi, K. Gatterdam, S. Tindall, A. Durand, V. Heinz, P.-A. Koenig, M. Geyer, C. Ziegler, G. H. Thomas and G. Hagelueken, *bioRxiv*, 2021, preprint, DOI: [10.1101/2021.12.03.471092](https://doi.org/10.1101/2021.12.03.471092).
- 16 P. Kakade, H. Ojha, O. G. Raimi, A. Shaw, A. D. Waddell, J. R. Ault, S. Burel, K. Brockmann, A. Kumar, M. S. Ahangar, E. M. Krysztofinska, T. Macartney, R. Bayliss, J. C. Fitzgerald and M. M. K. Muqit, *Open Biol.*, 2022, **12**, 210264.
- 17 C. M. Lin, J. H. Yang, H. J. Lee, Y. P. Lin, L. P. Tsai, C. S. Hsu, G. W. G. Luxton and C. F. Hu, *Life*, 2021, **11**.
- 18 N. Sen, I. Anishchenko, N. Bordin, I. Sillitoe, S. Velankar, D. Baker and C. Orengo, *Briefings Bioinf.*, 2022, **23**(4), bbac187.
- 19 I. R. Humphreys, J. Pei, M. Baek, A. Krishnakumar, I. Anishchenko, S. Ovchinnikov, J. Zhang, T. J. Ness,



S. Banjade, S. R. Bagde, V. G. Stancheva, X. H. Li, K. Liu, Z. Zheng, D. J. Barrero, U. Roy, J. Kuper, I. S. Fernandez, B. Szakal, D. Branzei, J. Rizo, C. Kisker, E. C. Greene, S. Biggins, S. Keeney, E. A. Miller, J. C. Fromme, T. L. Hendrickson, Q. Cong and D. Baker, *Science*, 2021, **374**, eabm4805.

20 UniProt, *UniProt Database*, 2022, <https://www.uniprot.org/>.

21 Y. Zhou, Y. Zhang, X. Lian, F. Li, C. Wang, F. Zhu, Y. Qiu and Y. Chen, *Nucleic Acids Res.*, 2022, **50**, D1398–D1407.

22 J. M. Rodriguez, F. Pozo, D. Cerdan-Velez, T. Di Domenico, J. Vazquez and M. L. Tress, *Nucleic Acids Res.*, 2022, **50**, D54–D59.

23 M. Baek, F. DiMaio, I. Anishchenko, J. Dauparas, S. Ovchinnikov, G. R. Lee, J. Wang, Q. Cong, L. N. Kinch, R. D. Schaeffer, C. Millan, H. Park, C. Adams, C. R. Glassman, A. DeGiovanni, J. H. Pereira, A. V. Rodrigues, A. A. van Dijk, A. C. Ebrecht, D. J. Opperman, T. Sagmeister, C. Buhlmann, T. Pavkov-Keller, M. K. Rathinaswamy, U. Dalwadi, C. K. Yip, J. E. Burke, K. C. Garcia, N. V. Grishin, P. D. Adams, R. J. Read and D. Baker, *Science*, 2021, **373**, 871–876.

24 M. Jendrusch, J. O. Korbel and S. K. Sadiq, bioRxiv, 2021, DOI: [10.1101/2021.10.11.463937](https://doi.org/10.1101/2021.10.11.463937).

25 Science's 2021 Breakthrough of the Year: AI brings protein structures to all, <https://www.science.org/content/article/breakthrough-2021>, 2022.

26 *Nature*, 2021, DOI: [10.1038/d41586-021-03734-6](https://doi.org/10.1038/d41586-021-03734-6).

27 Y. A. Ivanenkov, A. Zhebrak, D. Bezrukov, B. Zagribelnyy, V. Aladinskiy, D. Polykovskiy, E. Putin, P. Kamya, A. Aliper and A. Zhavoronkov, *arXiv*, 2021, preprint, arXiv:2101.09050, DOI: [10.48550/arXiv.2101.09050](https://doi.org/10.48550/arXiv.2101.09050).

28 H. Sung, J. Ferlay, R. L. Siegel, M. Laversanne, I. Soerjomataram, A. Jemal and F. Bray, *Ca-Cancer J. Clin.*, 2021, **71**, 209–249.

29 A. Jain, S. Chitturi, G. Peters and D. Yip, *World J. Hepatol.*, 2021, **13**, 1132–1142.

30 A. Zhavoronkov, F. Pun and B. Belli, *Bioengineering*, 2022, <https://go.nature.com/3QuoY8F>.

31 R. Ravi, K. A. Noonan, V. Pham, R. Bedi, A. Zhavoronkov, I. V. Ozerov, E. Makarev, A. V. Artemov, P. T. Wysocki, R. Mehra, S. Nimmagadda, L. Marchionni, D. Sidransky, I. M. Borrello, E. Izumchenko and A. Bedi, *Nat. Commun.*, 2018, **9**(1), 741.

32 P. Mamoshina, M. Volosnikova, I. V. Ozerov, E. Putin, E. Skibina, F. Cortese and A. Zhavoronkov, *Front. Genet.*, 2018, **9**, 242.

33 F. W. Pun, B. H. M. Liu, X. Long, H. W. Leung, G. H. D. Leung, Q. T. Mewborne, J. Gao, A. Shneyderman, I. V. Ozerov, J. Wang, F. Ren, A. Aliper, E. Bischof, E. Izumchenko, X. Guan, K. Zhang, B. Lu, J. D. Rothstein, M. E. Cudkowicz and A. Zhavoronkov, *Front. Aging Neurosci.*, 2022, **14**, 914017.

34 F. W. Pun, G. H. D. Leung, H. W. Leung, B. H. M. Liu, X. Long, I. V. Ozerov, J. Wang, F. Ren, A. Aliper, E. Izumchenko, A. Moskalev, J. P. de Magalhaes and A. Zhavoronkov, *Aging*, 2022, **14**, 2475–2506.

35 J. L. Chao, M. Korzinkin, A. Zhavoronkov, I. V. Ozerov, M. T. Walker, K. Higgins, M. W. Lingen, E. Izumchenko and P. A. Savage, *Cell Rep. Med.*, 2021, **2**, 100399.

36 I. V. Ozerov, K. V. Lezhnina, E. Izumchenko, A. V. Artemov, S. Medintsev, Q. Vanhaelen, A. Aliper, J. Vijg, A. N. Osipov, I. Labat, M. D. West, A. Buzdin, C. R. Cantor, Y. Nikolsky, N. Borisov, I. Irincheeva, E. Khokhlovich, D. Sidransky, M. L. Camargo and A. Zhavoronkov, *Nat. Commun.*, 2016, **7**, 13427.

37 M. T. Mok, J. Zhou, W. Tang, X. Zeng, A. W. Oliver, S. E. Ward and A. S. Cheng, *Pharmacol. Ther.*, 2018, **186**, 138–151.

38 M. Uhlen, L. Fagerberg, B. M. Hallstrom, C. Lindskog, P. Oksvold, A. Mardinoglu, A. Sivertsson, C. Kampf, E. Sjostedt, A. Asplund, I. Olsson, K. Edlund, E. Lundberg, S. Navani, C. A. Szigyarto, J. Odeberg, D. Djureinovic, J. O. Takanen, S. Hober, T. Alm, P. H. Edqvist, H. Berling, H. Tegel, J. Mulder, J. Rockberg, P. Nilsson, J. M. Schwenk, M. Hamsten, K. von Feilitzen, M. Forsberg, L. Persson, F. Johansson, M. Zwahlen, G. von Heijne, J. Nielsen and F. Ponten, *Science*, 2015, **347**, 1260419.

39 X. An, S. S. Ng, D. Xie, Y. X. Zeng, J. Sze, J. Wang, Y. C. Chen, B. K. Chow, G. Lu, W. S. Poon, H. F. Kung, B. C. Wong and M. C. Lin, *Eur. J. Cancer*, 2010, **46**, 1752–1761.

40 H. Feng, A. S. Cheng, D. P. Tsang, M. S. Li, M. Y. Go, Y. S. Cheung, G. J. Zhao, S. S. Ng, M. C. Lin, J. Yu, P. B. Lai, K. F. To and J. J. Sung, *J. Clin. Invest.*, 2011, **121**, 3159–3175.

41 Q. Wang, J. Ma, Y. Lu, S. Zhang, J. Huang, J. Chen, J. X. Bei, K. Yang, G. Wu, K. Huang, J. Chen and S. Xu, *Oncogene*, 2017, **36**, 5321–5330.

42 G. Q. Wu, D. Xie, G. F. Yang, Y. J. Liao, S. J. Mai, H. X. Deng, J. Sze, X. Y. Guan, Y. X. Zeng, M. C. Lin and H. F. Kung, *Int. J. Cancer*, 2009, **125**, 2631–2642.

43 J. Zhou, M. Liu, H. Sun, Y. Feng, L. Xu, A. W. H. Chan, J. H. Tong, J. Wong, C. C. N. Chong, P. B. S. Lai, H. K. Wang, S. W. Tsang, T. Goodwin, R. Liu, L. Huang, Z. Chen, J. J. Sung, K. L. Chow, K. F. To and A. S. Cheng, *Gut*, 2018, **67**, 931–944.

44 M. Caligiuri, F. Becker, K. Murthi, F. Kaplan, S. Dedier, C. Kaufmann, A. Machl, G. Zybarth, J. Richard, N. Bockovich, A. Kluge and N. Kley, *Chem. Biol.*, 2005, **12**, 1103–1115.

45 Eurofins, CDK20, <https://www.discoverx.com/kinase-data-sheets/cdk20>, 2022.

46 ChEMBL, Assay Report Card, [https://www.ebi.ac.uk/chembl/assay\\_report\\_card/CHEMBL4375310/](https://www.ebi.ac.uk/chembl/assay_report_card/CHEMBL4375310/), 2022.

47 D. Mueller, F. Totzke, T. Weber, C. Beissenherz-Huss, D. Kraemer, C. Heidemann-Dinger, C. Ketterer, C. Eckert and M. H. G. Kubbutat, *Cancer Res.*, 2016, **76**, 2821.

48 A. Zhavoronkov, Y. A. Ivanenkov, A. Aliper, M. S. Veselov, V. A. Aladinskiy, A. V. Aladinskaya, V. A. Terentiev, D. A. Polykovskiy, M. D. Kuznetsov, A. Asadulaev, Y. Volkov, A. Zholus, R. R. Shayakhmetov, A. Zhebrak, L. I. Minaeva, B. A. Zagribelnyy, L. H. Lee, R. Soll, D. Madge, L. Xing, T. Guo and A. Aspuru-Guzik, *Nat. Biotechnol.*, 2019, **37**, 1038–1040.



49 Q. Vanhaelen, Y. C. Lin and A. Zhavoronkov, *ACS Med. Chem. Lett.*, 2020, **11**, 1496–1505.

50 A. Kadurin, S. Nikolenko, K. Khrabrov, A. Aliper and A. Zhavoronkov, *Mol. Pharm.*, 2017, **14**, 3098–3104.

51 D. Polykovskiy, A. Zhebrak, D. Vetrov, Y. Ivanenkov, V. Aladinskiy, P. Mamoshina, M. Bozdaganyan, A. Aliper, A. Zhavoronkov and A. Kadurin, *Mol. Pharm.*, 2018, **15**, 4398–4405.

52 D. Polykovskiy, A. Zhebrak, B. Sanchez-Lengeling, S. Golovanov, O. Tatanov, S. Belyaev, R. Kurbanov, A. Artamonov, V. Aladinskiy, M. Veselov, A. Kadurin, S. Johansson, H. Chen, S. Nikolenko, A. Aspuru-Guzik and A. Zhavoronkov, *Front. Pharmacol.*, 2020, **11**, 565644.

53 E. Putin, A. Asadulaev, Q. Vanhaelen, Y. Ivanenkov, A. V. Aladinskaya, A. Aliper and A. Zhavoronkov, *Mol. Pharm.*, 2018, **15**, 4386–4397.

54 E. Putin, A. Asadulaev, Y. Ivanenkov, V. Aladinskiy, B. Sanchez-Lengeling, A. Aspuru-Guzik and A. Zhavoronkov, *J. Chem. Inf. Model.*, 2018, **58**, 1194–1204.

55 M. Kuznetsov and D. Polykovskiy, *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021, vol. 35, pp. 8226–8234.

56 Y. A. Cho, S. Choi, S. Park, C. K. Park and S. Y. Ha, *Cancer Genomics Proteomics*, 2020, **17**, 747–755.

57 G. Diaz, R. E. Engle, A. Tice, M. Melis, S. Montenegro, J. Rodriguez-Canales, J. Hanson, M. R. Emmert-Buck, K. W. Bock, I. N. Moore, F. Zamboni, S. Govindarajan, D. E. Kleiner and P. Farci, *Mol. Cancer Res.*, 2018, **16**, 1406–1419.

58 N. Chiyonobu, S. Shimada, Y. Akiyama, K. Mogushi, M. Itoh, K. Akahoshi, S. Matsumura, K. Ogawa, H. Ono, Y. Mitsunori, D. Ban, A. Kudo, S. Arii, T. Suganami, S. Yamaoka, Y. Ogawa, M. Tanabe and S. Tanaka, *Am. J. Pathol.*, 2018, **188**, 1213–1224.

59 H. W. Wang, T. H. Hsieh, S. Y. Huang, G. Y. Chau, C. Y. Tung, C. W. Su and J. C. Wu, *BMC Genomics*, 2013, **14**, 736.

60 J. Carrillo-Reixach, L. Torrens, M. Simon-Coma, L. Royo, M. Domingo-Sabat, J. Abril-Fornaguera, N. Akers, M. Sala, S. Ragull, M. Arnal, N. Villalmanzo, S. Cairo, A. Villanueva, R. Kappler, M. Garrido, L. Guerra, C. Sabado, G. Guillen, M. Mallo, D. Pineyro, M. Vazquez-Vitali, O. Kuchuk, M. E. Mateos, G. Ramirez, M. L. Santamaria, Y. Mozo, A. Soriano, M. Grotzer, S. Brachereau, N. G. de Andoin, B. Lopez-Ibor, R. Lopez-Almaraz, J. A. Salinas, B. Torres, F. Hernandez, J. J. Uriz, M. Fabre, J. Blanco, C. Paris, V. Bajcova, G. Laureys, H. Masnou, A. Clos, C. Belendez, C. Guettier, L. Sumoy, R. Planas, M. Jorda, L. Nonell, P. Czauderna, B. Morland, D. Sia, B. Losic, M. A. Buendia, M. R. Sarrias, J. M. Llovet and C. Armengol, *J. Hepatol.*, 2020, **73**, 328–341.

61 K. B. Hooks, J. Audoux, H. Fazli, S. Lesjean, T. Ernault, N. Dugot-Senant, T. Leste-Lasserre, M. Hagedorn, B. Rousseau, C. Danet, S. Brachereau, L. Brugieres, S. Taque, C. Guettier, M. Fabre, A. Rullier, M. A. Buendia, T. Commes, C. F. Grosset and A. A. Raymond, *Hepatology*, 2018, **68**, 89–102.

62 G. Liu, G. Hou, L. Li, Y. Li, W. Zhou and L. Liu, *Oncotarget*, 2016, **7**, 32607–32616.

63 Y. H. Wang, T. Y. Cheng, T. Y. Chen, K. M. Chang, V. P. Chuang and K. J. Kao, *BMC Cancer*, 2014, **14**, 815.

64 B. Losic, A. J. Craig, C. Villacorta-Martin, S. N. Martins-Filho, N. Akers, X. Chen, M. E. Ahsen, J. von Felden, I. Labgaa, D. D'Avola, K. Allette, S. A. Lira, G. C. Furtado, T. Garcia-Lezana, P. Restrepo, A. Stueck, S. C. Ward, M. I. Fiel, S. P. Hiotis, G. Gunasekaran, D. Sia, E. E. Schadt, R. Sebra, M. Schwartz, J. M. Llovet, S. Thung, G. Stolovitzky and A. Villanueva, *Nat. Commun.*, 2020, **11**, 291.

65 B. J. Erickson, S. Kirk, Y. Lee, O. Bathe, M. Kearns, C. Gerdes, K. Rieger-Christ and J. Lemmerman, *The Cancer Imaging Archive*, 2016, DOI: [10.7937/K9/TCIA.2016.IMGQW8UQ](https://doi.org/10.7937/K9/TCIA.2016.IMGQW8UQ).

