G4LDB 2.2: a database for discovering and studying G-quadruplex and i-Motif ligands

Yu-Huan Wang[1,†], **Qian-Fan Yang** [1,†], **Xiao Lin**[1], **Die Chen**[2], **Zhi-Yin Wang**[1], **Bin Chen**[1], **Hua-Yi Han**[2], **Hao-Di Chen**[1], **Kai-Cong Cai**[3], **Qian Li**[4], **Shu Yang** [2,*], **Ya-Lin Tang** [4,*] and **Feng Li** [1,*]

[1]Key Laboratory of Green Chemistry and Technology of Ministry of Education, College of Chemistry, Sichuan University, Chengdu610064, China, [2]West China School of Pharmacy, Sichuan University, Chengdu610041, China, [3]College of Chemistry and Materials Science, Fujian Provincial Key Laboratory of Advanced Materials Oriented Chemical Engineering, Fujian Normal University, Fuzhou350007, China and [4]Beijing National Laboratory for Molecular Sciences (BNLMS), Center for Molecular Sciences, State Key Laboratory for Structural Chemistry of Unstable and Stable Species, Institute of Chemistry, Chinese Academy of Sciences, Beijing 100190, China

ABSTRACT

Noncanonical nucleic acid structures, such as G-quadruplex (G4) and i-Motif (iM), have attracted increasing research interests because of their unique structural and binding properties, as well as their important biological activities. To date, thousands of small molecules that bind to varying G4/iM structures have been designed, synthesized and tested for diverse chemical and biological uses. Because of the huge potential and increasing research interests on G4-targeting ligands, we launched the first G4 ligand database G4LDB in 2013. Here, we report a new version, termed G4LDB 2.2 (http://www.g4ldb.com), with upgrades in both content and function. Currently, G4LDB2.2 contains >3200 G4/iM ligands, ~28 500 activity entries and 79 G4–ligand docking models. In addition to G4 ligand library, we have also added a brand new iM ligand library to G4LDB 2.2, providing a comprehensive view of quadruplex nucleic acids. To further enhance user experience, we have also redesigned the user interface and optimized the database structure and retrieval mechanism. With these improvements, we anticipate that G4LDB 2.2 will serve as a comprehensive resource and useful research toolkit for researchers across wide scientific communities and accelerate discovering and validating better binders and drug candidates.

INTRODUCTION

Past decades have witnessed tremendous research interests in biologically important noncanonical (non-B) nucleic acid structures, such as G-quadruplex (G4) and i-Motif (iM), because of their unique structural features and potential biological activities (1–3). G4s are a class of unique nucleic acid structures composed of stacked 'G-quartets' (4–6) that comprise a planar arrangement of four guanines stabilized by Hoogsteen hydrogen bonding. They have received increasing attention since the pioneering studies from the Blackburn, Sundquist and Cech groups in the late 1980s (7–9). It has been discovered that the formation of G4 is prevalent in the human genome and is liable for various biological functions, ranging from transcription (10–13), to translation (14,15) and to cell aging (16–18). Therefore, many G4s are closely related to diseases, such as cancer (19), diabetes, (20,21) and neurodegenerative disease (22,23). First discovered by Gehring *et al.* (24), iMs are another class of cytosine-rich quadruplex nucleic acid structures that have been found to play important physiological functions in multiple key genomic regions, including telomeres, proto-oncogene promoters and HIV proviral genomes (25–29).

Because of their critical physiological functions, G4 and iM are naturally ideal targets for drug development. Numerous small molecules have been designed to recognize, regulate and probe G4/iM in living cells and/or *in vivo*. Remarkably, many G4/iM ligands have shown significant biological activities, including the regulation of oncogene expression (30–32), the inhibition of telomeres extension (33) and the induction of cancer cell senescence and apoptosis (16,18). For example, BRACO-19 is

---

a well-characterized potent and selective ligand targeting telomeric G4, demonstrating *in vivo* anti-tumor effect on mice bearing vulval carcinoma (34). NSC309874 (a benzothiophene-2-carboxamide derivative) is a PDGFR-β i-Motif-interactive compound that was found to downregulate PDGFR-β promoter activity in the neuroblastoma cell line SK-N-SH (35). Some ligands have already entered the stage of clinical trials. For example, a G4 stabilizer CX-5461 is currently in the Phase I/II clinical trials for treating advanced cancers bearing BRCA1/2 deficiencies (36). To date, thousands of G4/iM ligands have been developed and this pool of ligands is expected to keep increasing given the current trend of research.

Given the structural and biological importance of non-B nucleic acid structures, several powerful databases have been launched, including QuadBase2 (37), NALDB (38) and G4IPDB (39), with focuses on nucleic acid topological structures, ligands and binding proteins, respectively. Given the huge potential and increasing research interests in G4-targeting ligands, we have previously built the first exclusive G4 ligand database G4LDB in 2013 (40). G4LDB collected 1105 G4 ligands with 4751 activity records. It also contained an online tool for ligand design and the real-time prediction of binding affinity, offering a convenient way for accelerating ligand discovery. However, since the initial publication of G4LDB, tremendous research efforts have been made for discovering new G4 ligands, for improved the understanding of G4–ligand interactions and subsequent biological impact, as well as for exploring their therapeutic applications (41). Therefore, the first edition of G4LDB no longer meets the current need in G4 research. Beyond G4 ligand database, the discovery of iM *in vivo* (42), as well as the growing research interests in finding iM ligands for therapeutic and regulatory uses (43) also demands a ligand database for iMs, which is currently unavailable.

Herein, we release a new version of G4LDB, termed G4LDB 2.2 (http://www.g4ldb.com). G4LDB 2.2 includes over 3000 G4 and iM ligands and 28 500 activity entries. It also features redesigned database structure, upgraded retrieval mechanism, as well as enhanced web architecture and user-interface. With these advancements, we believe G4LDB 2.2 will not only offer comprehensive information and analyses on existing quadruplex nucleic acid ligands but also facilitate the research and development of new binders and drug candidates.

MATERIALS AND METHODS

Data acquisition

Structures of ligands. Structures of ligands were created by MolView (www.molview.org) (44) based on the descriptions in the literature. The structure descriptors of ligands (isomeric or canonical SMILES strings) were gathered from the NCBI PubChem database (45) or generated by MolView. Molecular identities and descriptors, including PubChem compound ID (CID), IUPAC name and synonyms were gathered from PubChem. Other molecular descriptors, including formula, molecular weight, number of H-bond donors (HBD count), number of H-bond acceptors (HBA count), Alogp (Ghose-Crippen-Viswanadhan

octanol-water partition coefficient) and molecular solubility, were gathered from PubChem or computed by Discovery Studio 2.5 (Accelerys, San Diego, CA, USA). The G4–ligand complex information was obtained from the RCSB Protein Data Bank (PDB) (46).

Structure searching tool. Molecular structures can be drawn directly using the graphical user interface of Ketcher 2.0 (47) (EPAM, Newtown, PA, USA) embedded in web pages. Based on the Indigo toolkit (EPAM, Newtown, PA, USA), three searching modes of structures were provided, including exact, substructure and similarity match. The similarity between the user-designed structure and ligands in the database could be quantified by three algorithms, including Tanimoto, Tversky and Euclid-sub. All thresholds for the similarity retrieval were set as 0.8.

Data of activities. The activity data of G4/iM ligands were collected from the literature. A three-level classification for ligand activity information was employed for quick data locating. Activity 1 was designed to the level of ligand function, which is divided into four categories, including intermolecular interaction, biological activity at the molecular level, activity at the cellular level and *in vivo* activity. Activity 2 was designed to be divided based on the specific structural (e.g. stabilization and structural regulation) or biological (e.g. gene expression and cytotoxicity) function of ligands. This level of activity categories was further divided based on detailed experimental Materials and Methods (e.g. UV-Vis, NMR, TRAP assay and MTT assay). Brief descriptions of experimental conditions for activity data were also recorded in the Comments field to offer methodology information.

Docking models information. The G4–ligand complex structures and binding characteristics were obtained from PDB (46). All of the structures were processed and prepared to construct docking models by AutoDockTools 1.5.6 (48).

System data updates. New data related to G4/iM ligands will be updated semi-annually (usually in January and August).

Molecular visualization

G4LDB 2.2 can display a series of molecular visualization effects. The 2D structure for each ligand is generated by Ketcher 2.0, and the 3D structure is presented by the web-embedded JSmol applet (49). The docking models and the predicted results are also displayed by the JSmol applet. A display control panel is provided for choosing calculated modes and adjusting display styles.

Online docking

To prepare docking models of G-quadruplex, 3D structures of 79 ligand/G-quadruplex complexes recorded in the PDB (from 1994 to 2021) were retrieved. G4LDB 2.2 online docking proceeded in three steps. The first step was to prepare the receptor. Ligands were first removed from the docking models by PyMOL 2.4.0 (Schrodinger) to leave

receptors with empty binding pockets defined as the grid boxes. Gasteiger partial charges and hydrogens were then assigned to the atoms in each receptor by AutoDockTools 1.5.6. The binding site coordinate and the gird box size were set according to the original data in the PDB. Detailed information of the binding site for each model was listed in Supplementary Table S1. The second step requires a user to design and input a customized ligand. The 2D structure of a user-designed ligand will be input online by the Ketcher applet. It will then be converted into a 3D structure and undergone geometry optimization to produce the initial ligand structure. The 2D to 3D transformation, geometry optimization and molecular format transformation will be performed using OpenBabel 3.1.1 (50) and Raccoon 1.0b (51) at the server end. In the last step, the chosen receptor and the processed ligand will be docked online by Autodock Vina 1.1.2 (52). All parameters have been set as default. When searching the conformational and orientational spaces of a structurally flexible ligand with fully rotatable bonds, the structure of the G-quadruplex was kept rigid. For each docking evaluation, at least 20 independent runs were performed to evaluate different ligand poses. Several most favorable poses (default value is 1) were dumped into the result file.

Database server implementation

G4LDB 2.2 has been installed on Spring Cloud server workstations. A Nginx 1.17.10 server has been used as the webserver platform. The website has been built with Vue Js 2. The PostgreSQL 9.6 relational database management system was employed to organize, manage and store data. The graphic chemical editor Ketcher 2.0 was used to build interactive web interfaces. The structure matching and molecular similarity prediction was accomplished by the Indigo toolkit. JSmol applets was embedded in the interface to render 3D structures of ligands and complexes. G4LDB 2.2 site is best viewed by Google Chrome.

RESULTS

Overview of G4LDB 2.2

G4LDB 2.2 is the latest version of the quadruplex nucleic acid ligand database. It aims to provide a comprehensive collection of small molecular ligands for G4 and iM with detailed physical/chemical information and biological activities. It also provides an online ligand design module allowing the prediction of ligand binding affinity, as well as ligand–receptor docking in real-time.

Comparing to the previous version of G4LDB, a series of updates and improvements have been made in both content and function. The latest version of the database includes 3099 G4 ligands and corresponding complex and activity records. A brand new iM ligand library was also built for the first time, including 110 ligands and corresponding activity records. Particularly, the number of cellular activity entries increased from 1691 to 8096, and 351 *in vivo* activities were included for the first time. A detailed comparison between G4LDB and G4LDB 2.2 was shown in Table 1. Because of the drastic increase in the amount of both ligands

Table 1. The comparison between G4LDB and G4LDB 2.2.

		G4LDB	G4LDB 2.2
Data capacity	G4 ligands	1105	3099
	G4–ligand complexes	0	44
	G4 ligand activities	4751	27807
	iM ligands	0	110
	iM ligand activities	0	883
Activity statistics	Molecular interaction	1955	18595
	Biological activity in molecular level	1105	1642
	Biological activity in cellular level	1691	8096
	Biological activity *in vivo* level	0	357
Docking	G4 models	28	79

and activities, we have also introduced a new retrieval mechanism and advanced searching engineer to improve the data accessibility. To also improve compatibility, JavaScript plugins were used to replace the previous JAVA ones. A brand-new user interface was also introduced.

Updated database content

Data in G4LDB 2.2 are continuously updated. The updated information includes new ligands/complexes, actives and docking models. Comparing to the previous version released in 2013, G4LDB 2.2 has included >1900 new G4 ligands and a new iM ligand library. It has also collected 44 reported G4–ligand complexes based on the data in the PDB. Detailed information on the complexes was listed in Supplementary Table S2. Docking models have also been enlarged from 28 to 79.

To accommodate the massive growth in data and to satisfy diversified needs of users, we have also improved the database structure. Molecular descriptors from PubChem, including PubChemCID, IUPAC name, SMILES string, formula, molecular weight (Mw) and synonyms, have been included for available ligands. An external hyperlink for each ligand to PubChem has also been added. Typically, the formula and Mw of ligands do not include coordinated ions or salts in G4LDB 2.2. However, in the case of metal complexes, the coordinated ions have been included in the formula and Mw.

We have also redesigned the ligand activity table to facilitate better search and browse of the critical information. A three-level-classification of activity was applied for quick data locating. Sequence information was also enriched by adding Sequence name found in the literature and Nucleic acid type (DNA/RNA). Given that ligands may occasionally demonstrate varied activities in the presence of different counter ions, we have also combined counter ion information in the ligand activity table. Detailed experimental conditions have also been collected and presented in the comments field.

Updated database function

Comparing to the previous version, G4LDB 2.2 provides more utility functions to better facilitate users for searching, browsing, retrieving, analyzing and exporting interested information.

Figure 1. Two new modules on the main page. (**A**) Latest entries module demonstrates 6 star G4/iM ligands periodically. (**B**) News module displays latest significant progresses in the G4/iM ligand research field monthly.

The main page. Two new modules, Latestentries and News, have been included for the first time (Figure 1). Six selected newly published ligands are displayed in the Latestentries column periodically with information including the ligand structures and/or binding complexes, brief introductions of activity and the external hyperlinks to references. By clicking the ligand ID, users can access to the detailed information of the ligand. News column displays latest progresses in the G4/iM ligand research field and important updates about the database at a frequency of each month.

The search module. In G4LDB 2.2, fuzzy search and advanced search can both be applied according to the specific needs of users (Figure 2). The retrieval operation can be performed by simply inputting any key words in the search box on the main page (Figure 2A), or using the advanced search module that is further divided into Ligand Search and Structure Search. There are three ways to access to the advanced search module, including using the hyperlink under the fuzzy search box, clicking the Advanced Search icon on the main page or using the Search menu on the navigation bar at the top of the pages (Figure 2A).

On the Ligand Search page (Figure 2B), multiple search conditions, including quadruplex type (G4 and/or iM), ligand descriptors, nucleic acid sequence, reference information, ligand property and ligand activity, can be combined with AND/OR logical relationship to retrieve the interested entries. Particularly, we have enhanced the sequence search function (Figure 2C) so that users are able to search nucleic acid targets with exact or upstream/downstream-extended sequence information. It also allows the search for sequences with a few (no >10) mutant bases. To allow more flexible search, symbols for uncertain nucleic acid bases are also supported, such as 'N' presents A, T, C or G while 'R' presents only A or G.

The StructureSearch module provides a convenient approach to retrieve ligand information according to 2D molecular or fragment structures (Figure 2D). Several acquisition methods and search patterns have been provided for the structure search. Users can build a retrieval molecule in three ways, including drawing and editing a structure directly in the Drawing window, inputting a SMILES string or uploading a molecular file. The supported file formats include mol, rxn, smi, smiles, cxsmi, cxsmiles, smarts, inchi and mrv. The detailed instruction can be found by clicking the question mark. There are three modes for structure-based querying, including Exact, Sub Structures and Similarity match. For searching similar structures, three algorithms, including Tanimoto, Tversky and Euclid-sub, can be chosen to achieve better matches.

The browse module. In G4LDB 2.2, the browse function is integrated in the SearchResult page (Figure 3A). Several new functions have been introduced. A sidebar was designed to display the number of retrieved entries and the search conditions. Eligible entries are listed in the right column with three modes, including Ligand, Activity and Sequence. Users can easily switch the mode of listing to browse specific summary of information on entries (Figure 3B). To facilitate the quick location of desired information, entries can be ranked in several different ways in different listing modes (Figure 3B). For example, entries can be sorted by Ligand ID, PubChem CID, Mw or First reported year in the ligand listing mode. They could also be ranked by Activity, Value, Ligand ID or Publication year in the activity mode. In the ligand listing modes, up to three entries (if available) matching the name and synonyms fields via fuzzy search would be recommended on the top of the page (Figure 3C). This function may better facilitate users to find desired results. Users can also use the SimilarStructuresSearch button under each ligand to expand their browse (Figure 3D). The refine function has also been integrated into the sidebar to help narrow down the search results. The refine conditions vary with different listing modes automatically. G4 ligands with or without target–ligand complex information can be screened in the ligand listing mode, while entries with DNA or RNA sequences can be filtered in the sequence and activity listing modes. Furthermore, the Export and Analysis
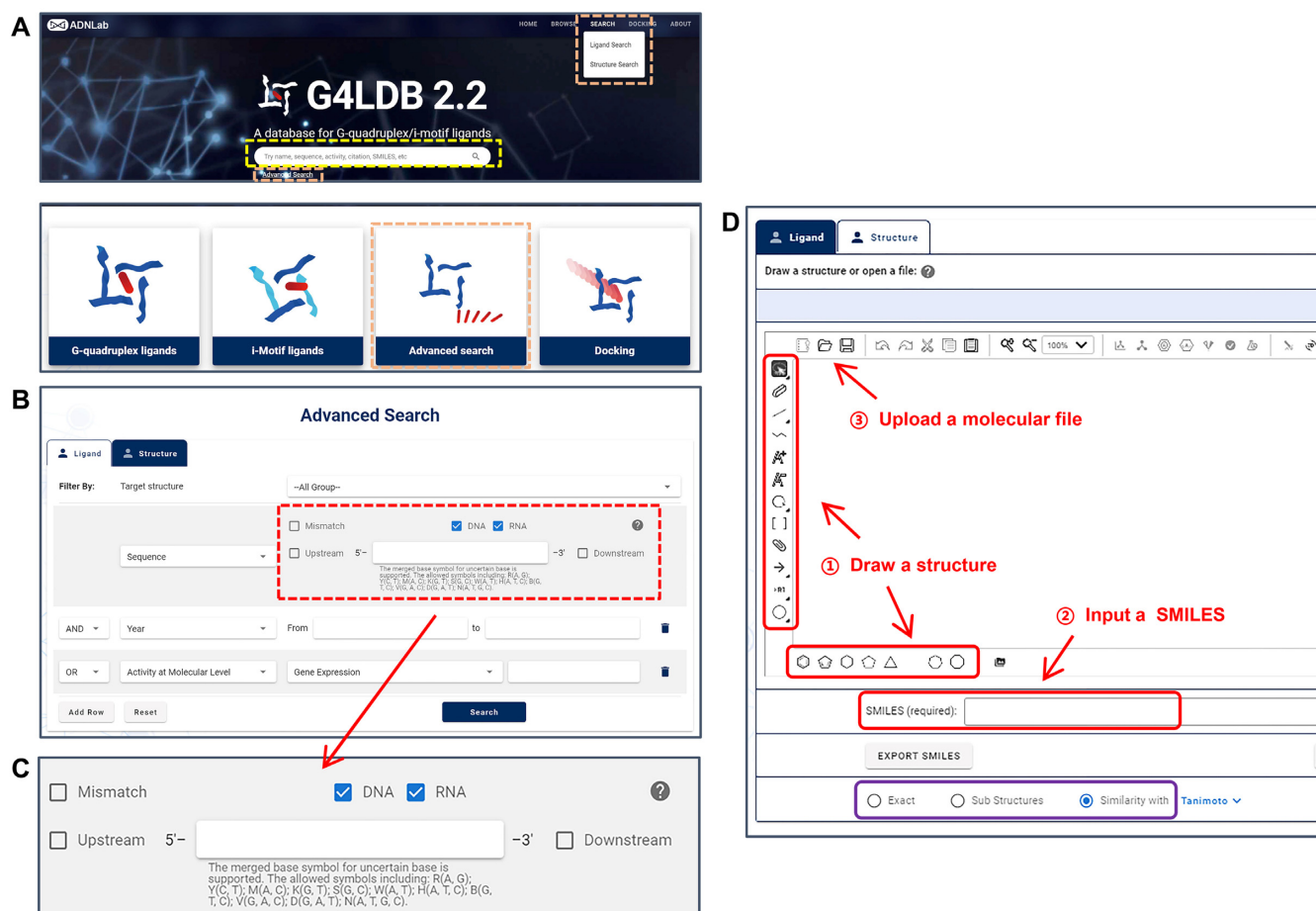
Figure 2. Two search ways in G4LDB 2.2. (**A**) Entries of the fuzzy search (yellow dashed box) and advanced search (orange dashed boxes) on the main page. (**B**) The ligand search module allows combining multiple types of conditions with AND/OR logical relationship. (**C**) Enhanced sequence search condition. (**D**) The structure search module provides three ways to build a retrieval molecule (red arrows) and three matching modes (purple box).

functions have been added. Entries can be selected individually or in bulk for subsequent exporting or statistical analysis. The export options have also been customized with the listing mode.

The LigandDetail page has also been reorganized in the latest version of the database (Figure 4). Detailed information on ligand structure, property and activity have been separated by tabs. 2D and simulated 3D structures (if available) of ligands are displayed in the Structure tab (Figure 4A), accompanied by ligand identity information, including the Name in the reference, IUPAC Name (if available), SMILES, PubChem CID (if available) and synonyms (if available). External hyperlinks to the PubChem database have also been added to provide source information about ligands. The physicochemical properties of ligands including formula, Mw, HBD and HBA count, AlogP and molecular solubility are provided in the Properties tab line by line (Figure 4B). The activity table has been reorganized (Figure 4C), in which target sequences (with name and type if available) and counter ions (if available) have been added. Reference information has also been supplemented by adding author names, journal of publication, year of publication and DOI. Hyperlinks have been added to Sequence, Activity 1, Activity 2, Methods and DOI for quick extended

search. Moreover, complex information available for some ligands has been added as an independent tab (Figure 4D). Detailed information on receptor–ligand complex, including reference, receptor sequence, released date and a hyperlink to the PBD have been provided. The complexes can also be used directly as docking models.

The docking module. Comparing to the previous version, the docking module of G4LDB 2.2 has been enhanced by greater number of available models, more flexible ligand design and faster calculation. Molecular dynamics docking is highly empirical. As a result, automatic docking tools may lead to unreliable prediction results. To achieve more reliable prediction, we prepared 79 models from the PDB database, each of which contains a high-resolution 3D structure and experimentally confirmed binding sites. Instructions and warnings have also been added for guiding the better selection of docking models/sites and for evaluating the docking results. In the current version of the database, the docking process has been optimized as a 4-step process.

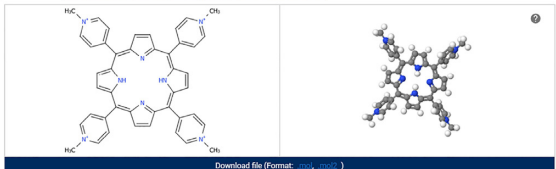- Step 1 is to choose an appropriate docking model. Similar to the search result page, the docking model browse

Figure 3. (**A**) Redesigned search result page. The sidebar displays the number of retrieved entries and the search conditions (red arrows), and integrates refine function (purple box). (**B**) Eligible entries are listed with three modes, including Ligand, Activity and Sequence, and each mode provides unique summary information and rank options. (**C**) Recommend entries matching the name and synonyms fields (if available) would be presented at the top of the page. (**D**) The similar structure search function.

Figure 4. The ligand detail page. Detailed information on a ligand is separated by tabs, including structure (**A**), properties (**B**), activities and receptor–ligand complex. (**C**) The reorganized activity information table. (**D**) The complex information table.

page has been integrated with the refine and rank functions. Detailed information on models has been provided, including complex description, receptor/ligand properties and binding site/gridbox. The 3D structures of models have been embedded in the page by the JS-mol applet 3D visualization tools, allowing users to rotate and zoom in/out the complexes. Hyperlinks to PDB ID, ligand ID and PubChem CID have also been pro-

vided for extra information about the complexes and the ligands (Figure 5A). Seventy-nine models are included in the current version and will be updated frequently. For models having multiple binding sites, users have the option to choose different sites as needed. The corresponding binding modes and gridboxes would be presented automatically associated with the chosen site. In the current version, we have only provided fixed
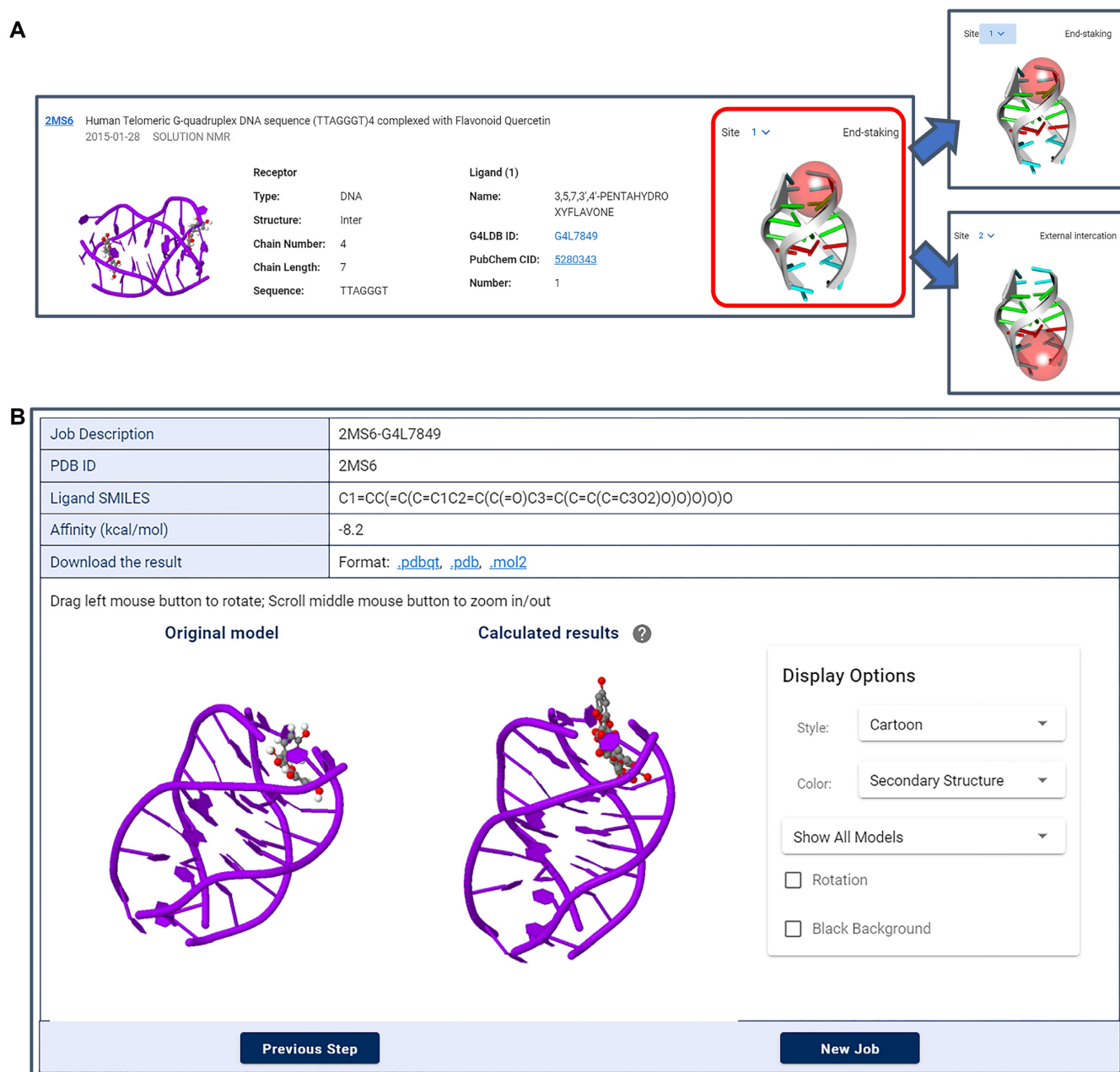
Figure 5. (**A**) The detailed information for docking models. For models with multiple binding sites, different binding gridboxes and binding modes would be presented associated with the chosen site selected in the drop-down menu. (**B**) The prediction result page.

gridbox coordinates and rigid docking algorithm. Customized gridbox setting and more complex algorithms (such as semi-flexible docking) would be supplied in the future.

- Step 2 is to design a ligand. Users have the option to build a ligand molecule online, or directly load an existing ligand in the database. The original ligand for a chosen docking model is also presented for help evaluate the reliability of the chosen model. In order to improve the accuracy of the prediction, multiple preferred conformations can be calculated in one docking job in the latest version.

- Step 3 is to confirm and submit a job. By doing so, the process of molecular docking, binding evaluation and results extraction will be automatically carried out at the server end. The docking job typically takes 1–2 min. Given that some docking jobs would take a longer time, users have the option to leave email addresses and a notification message will be sent when the job is done. Both the original model and the predicted binding modes, as well as molecular interactions are rendered directly by the JSmol applet 3D visualization tools in step 4 for better observation and comparation. The predicted affinity of G4–ligand complex is presented in the resulting table

and the structure is downloadable in the format of pdbqt, pdb and mol2 (Figure 5B). To confirm the docking result is reliable, the value of the calculated docking free energy would be checked automatically. There would be a warning to inform users if the calculated free energy is positive.

The feedback module. G4LDB 2.2 provides a feedback module for users to submit any suggestion and correction about the database. Users can contact us by email or via the newly designed feedback page. A feedback hyperlink has been added at the bottom of the main page and a Feedback button has been added on the detail page for any ligand. The ligand ID would be recorded automatically and be displayed on the title of the feedback page when users submit suggestions or corrections for a specific ligand.

Database and web server architecture

The website was completely rebuilt with Vue.js for better stability. The PostgreSQL 9.6 relational database management system is employed to organize, manage and store data. To improve the security and compatibility, the JavaScript is used instead of the previously used Java plugins. Moreover, the user interface was also re-constructed to improve user experience.

DISCUSSION

With a series of much expanded resources and newly added functions, G4LDB 2.2 offers a highly convenient platform for the design and discovery of G4/iM-targeting ligands/drugs. Using the examples below, we illustrate some of the key applications of G4LDB 2.2.

Browsing G4/iM ligands and checking information for a certain ligand

Currently, G4LDB 2.2 collects >3000 G4 ligands and 100 iM ligands. By clicking the Browse menu of the navigation bar, users can browse all G4/iM ligands on the search result page. By clicking the ligand ID or View detail button, the page with detailed information would pop up. Users have the option to visit the structural, property, activity and G4–ligand complex (if available) information for the selected ligand by clicking the corresponding tabs. Taking a well-established G4 ligand, TMPyP4, as an example, users can access its 2D/3D structure, the physicochemical properties, as well as >550 records of activities summarized from >60 literature. The comprehensive information offered by G4LDB 2.2 reflects the discovery and development track of TMPyP4 in the past 20 years. Users can also retrieve four G4–TMPyP4 complexes in the Receptor–Ligand Complex tab page. The 3D structure and related literatures of the complexes are provided, as well as hyperlinks to PDB. Users can choose one complex as the model to perform docking calculations via the DOCKING button.

Retrieving specific information and establishing statistical relationships between search results

G4LDB 2.2 provides two types of search modules: fuzzy search and advanced search. Taking the human telomere 22 nt sequence AGGGTTAGGGTTAGGGTTAGGG (hTel22) as an example, all entries related to the sequence would be returned upon inputting the sequence in the search box of the main page. The result includes entries related to 63 sequences. More accurate/specific retrieval can be achieved using the advanced search function. By clicking the hyperlink under the search box, users will reach the page of ligand search. By further clicking the Add Row button → choosing the Sequence condition → typing the sequence into the search box → clicking the Search button, users will find the specific information related only to the exact sequence. Users may also expand their exploration by using upstream/downstream-extended or mismatch options. For example, by checking the Upstream box in the last step, entries related to 32 sequences would be returned. The result can be further refined by checking the 24 nt sequence TTAGGGTTAGGGTTAGGGTTAGGG (hTel24) on the sidebar and clicking Refine button. In this way, users can easily change the retrieve object from hTel22 to hTel24.

Having performed the search function, users can further analyze the search results based on the year of publication, Mw, HBD/HBA count, AlogP or molecular solubility using the analysis function. For example, once the search results for hTel24 related ligands, users can find the distribution of publications over the years for the 152 retrieved ligands by clicking the Analysis button → All records → ANALYSIS. Similar distributions of these 152 ligands as a function of Mw, HBD/HBA count, AlogP and molecular solubility can also be achieved by selecting the corresponding function in the menu. All analysis results can be downloaded as images or CSV files.

Designing small molecular ligands and predicting molecular interactions between small molecules and G-quadruplexes

Docking module is an important part of G4LDB, which can predict binding strength of newly designed ligands online. This module incorporated 79 docking models, covering nearly all known binding modes reported in the PDB database.

Here we take a docking job based on the model 2MS6 as an example to illustrate the docking function. On the page of docking, detailed information of all 79 models is listed. 2MS6 with its original ligand Quercetin (G4L7849) can be found by browsing the listed models or search its name at the sidebar. As 2MS6 contains two different binding sites, the users have the option to choose the specific binding site by checking the dropdown menu of the Site function (Figure 5A). The site 1 corresponds to the ligand end-stacks on one end of the G4 receptor, whereas site 2 corresponds to external interaction. Once choosing the model and site, users can process to the next step by clicking Next step button. In this step, users can either draw a ligand from scratch or import an existing ligand from the database. Here we demonstrate the design of a ligand based on the original ligand Quercetin in the model. This can be achieved by following G4LDB Ligand → inputting ligand ID 7849 → DRAW button. Both the structure and SMILES string of the ligand G4L7849 would be loaded. Upon following Draw Ligand → DRAW SMILES button, users can further customize the ligand through free modification. By following EXPORT SMILES → inputting Job Description → inputting

Number Modes → Next Step button, users can preview the calculation job, check the docking model and the binding site, and compare the structural difference between the original ligand and designed ones. The docking job usually takes 1–2 min, which is much faster than that in the previous version of the database. Both the original model and the predicted binding modes are presented and a display control panel is provided for choosing the calculated modes and adjusting display styles (Figure 5B).

Building quantitative structure–activity relationship models based on selected activity

The quantitative structure–activity relationship (QSAR) approach is an essential part of drug development. Our G4LDB 2.2 serves as an ideal source for building predictive QSAR models. In the latest version, the updated refine and export function makes it much easier to acquire relationships between ligands and selected activities.

Here we use berberine derivates and their abilities to stabilize G4 measured by Circular Dichroism (CD) as an example to illustrate the QSAR application. First, on the Ligand Search page, users can find berberine (G4L1014) by using combined search condition, Target structure as G-quadruplex and Name as Berberine. By following the Similar Structures Search button → Turn to list button, one can retrieve all berberine derivates in the database. By further refining the Ligand type as G-quadruplex, Ligand activity 2 as Stabilization, and Method as Circular Dichroism (CD), 7 interested G4 ligands would be returned. By following Export button → Records from 1 to 7 → Ligand ID and SMILES → EXPORT, users can export a CSV file containing all the ligands with structure information. Another file containing the activity information can also be obtained by following the Activity listing mode → Export button → Records from 1 to 16 → Ligand ID, Sequence, Activity 1, Activity 2, Method and Value → EXPORT. The QSAR studies can then be performed on the basis of the two sets of data.

CONCLUSION

Recent advances in discovering and studying G4-targeting ligands have yielded huge number of novel ligands, research methods and activity information. iMs have also attracted significant research interests in recent years. Motivated by the current needs for indexing and sorting these important resources, we have released the updated G4 and iM ligand database, G4LDB 2.2. Comparing to the previous version, the size of G4-ligand library has increased over 6-fold, and an iM ligands library has also been included for the first time. The data organization has been optimized for better usability and the search module has also been enhanced with much improved accessibility. Moreover, the web server architecture and user interface have also been redesigned with much enhanced user experience. By continuously offering up-to-date and comprehensive information for G4 and iM ligands, we anticipate that G4LDB 2.2 will serve as a unique and widely used research tool to accelerate quadruplex nucleic acid research.

DATA AVAILABILITY

The database is now publicly accessible through the URL http://www.g4ldb.com/.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

REFERENCES

1. Ahmed,S., Kintanar,A. and Henderson,E. (1994) Human telomeric C-Strand tetraplexes. *Nat. Struct. Biol.*, **1**, 83–88.
2. Brazier,J.A., Shah,A. and Brown,G.D. (2012) I-Motif formation in gene promoters: unusually stable formation in sequences complementary to known G-quadruplexes. *Chem. Commun.*, **48**, 10739–10741.
3. Suseela,Y.V., Narayanaswamy,N., Pratihar,S. and Govindaraju,T. (2018) Far-red fluorescent probes for canonical and non-canonical nucleic acid structures: current progress and future implications. *Chem. Soc. Rev.*, **47**, 1098–1131.
4. Gellert,M., Lipsett,M.N. and Davies,D.R. (1962) Helix formation by guanylic acid. *Proc. Natl. Acad. Sci. USA.*, **48**, 2013–2018.
5. Gray,L.T., Puig Lombardi,E., Verga,D., Nicolas,A., Teulade-Fichou,M.P., Londono-Vallejo,A. and Maizels,N. (2019) G-quadruplexes sequester free heme in living cells. *Cell Chem. Biol.*, **26**, 1681–1691.
6. Sen,D. and Gilbert,W. (1988) Formation of parallel 4-Stranded complexes by guanine-rich motifs in DNA and its implications for meiosis. *Nature*, **334**, 364–366.
7. Henderson,E., Hardin,C.C., Walk,S.K., Tinoco,I. and Blackburn,E.H. (1987) Telomeric DNA oligonucleotides form novel intramolecular structures containing guanine guanine base-pairs. *Cell*, **51**, 899–908.
8. Sundquist,W.I. and Klug,A. (1989) Telomeric DNA dimerizes by formation of guanine tetrads between hairpin loops. *Nature*, **342**, 825–829.
9. Williamson,J.R., Raghuraman,M.K. and Cech,T.R. (1989) Mono-Valent cation induced structure of telomeric DNA - the G-Quartet model. *Cell*, **59**, 871–880.
10. Cogoi,S. and Xodo,L.E. (2006) G-quadruplex formation within the promoter of the KRAS proto-oncogene and its effect on transcription. *Nucleic Acids Res.*, **34**, 2536–2549.
11. Ogasawara,S. (2018) Transcription driven by reversible photocontrol of hyperstable G-Quadruplexes. *ACS Synth. Biol.*, **7**, 2507–2513.
12. Kim,N. (2019) The Interplay between G-quadruplex and Transcription. *Curr. Med. Chem.*, **26**, 2898–2917.
13. Cui,X., Chen,H., Zhang,Q., Xu,M., Yuan,G. and Zhou,J. (2019) Exploration of the Structure and Recognition of a G-quadruplex in the her2 Proto-oncogene promoter and its transcriptional regulation. *Sci. Rep.*, **9**, 3966.
14. Kumari,S., Bugaut,A., Huppert,J.L. and Balasubramanian,S. (2007) An RNA G-quadruplex in the 5 ′ UTR of the NRAS proto-oncogene modulates translation. *Nat. Chem. Biol.*, **3**, 218–221.
15. Qin,Y. and Hurley,L.H. (2008) Structures, folding patterns, and functions of intramolecular DNA G-quadruplexes found in eukaryotic promoter regions. *Biochimie*, **90**, 1149–1171.
16. Pennarun,G., Granotier,C., Gauthier,L.R., Gomez,D., Hoffschir,F., Mandine,E., Riou,J.F., Mergny,J.L., Mailliet,P. and Boussin,F.D. (2005) Apoptosis related to telomere instability and cell cycle alterations in human glioma cells treated by new highly selective G-quadruplex ligands. *Oncogene*, **24**, 2917–2928.
17. Biffi,G., Tannahill,D., McCafferty,J. and Balasubramanian,S. (2013) Quantitative visualization of DNA G-quadruplex structures in human cells. *Nat. Chem.*, **5**, 182–186.

18. Sengupta,P., Banerjee,N., Roychowdhury,T., Dutta,A., Chattopadhyay,S. and Chatterjee,S. (2018) Site-specific amino acid substitution in dodecameric peptides determines the stability and unfolding of c-MYC quadruplex promoting apoptosis in cancer cells. *Nucleic Acids Res.*, **46**, 9932–9950.

19. Neidle,S. and Parkinson,G. (2002) Telomere maintenance as a target for anticancer drug discovery. *Nat. Rev. Drug Discov.*, **1**, 383–393.

20. Connor,A.C., Frederick,K.A., Morgan,E.J. and McGown,L.B. (2006) Insulin capture by an insulin-linked polymorphic region G-quadruplex DNA oligonucleotide. *J. Am. Chem. Soc.*, **128**, 4986–4991.

21. Schonhoft,J.D., Bajracharya,R., Dhakal,S., Yu,Z.B., Mao,H.B. and Basu,S. (2009) Direct experimental evidence for quadruplex-quadruplex interaction within the human ILPR. *Nucleic Acids Res.*, **37**, 3310–3320.

22. Haeusler,A.R., Donnelly,C.J., Periz,G., Simko,E.A.J., Shaw,P.G., Kim,M.S., Maragakis,N.J., Troncoso,J.C., Pandey,A., Sattler,R. *et al.* (2014) C9orf72 nucleotide repeat structures initiate molecular cascades of disease. *Nature*, **507**, 195–200.

23. Cammas,A. and Millevoi,S. (2017) RNA G-quadruplexes: emerging mechanisms in disease. *Nucleic Acids Res.*, **45**, 1584–1595.

24. Gehring,K., Leroy,J.L. and Gueron,M. (1993) A tetrameric DNA-structure with protonated cytosine·cytosine base-pairs. *Nature*, **363**, 561–565.

25. Garavis,M., Escaja,N., Gabelica,V., Villasante,A. and Gonzalez,C. (2015) Centromeric alpha-satellite DNA adopts dimeric i-Motif structures capped by at hoogsteen base pairs. *Chem-Eur. J.*, **21**, 9816–9824.

26. Ruggiero,E., Lago,S., Sket,P., Nadai,M., Frasson,I., Plavec,J. and Richter,S.N. (2019) A dynamic i-motif with a duplex stem-loop in the long terminal repeat promoter of the HIV-1 proviral genome modulates viral transcription. *Nucleic. Acids. Res.*, **47**, 11057–11068.

27. Abou Assi,H., Garavis,M., Gonzalez,C. and Damha,M.J. (2018) i-Motif DNA: structural features and significance to cell biology. *Nucleic Acids Res.*, **46**, 8038–8056.

28. Shu,B., Cao,J.J., Kuang,G.T., Qiu,J., Zhang,M.L., Zhang,Y., Wang,M.X., Li,X.Y., Kang,S.S., Ou,T.M. *et al.* (2018) Syntheses and evaluation of new acridone derivatives for selective binding of oncogene c-myc promoter i-motifs in gene transcriptional regulation. *Chem. Commun.*, **54**, 2036–2039.

29. Takahashi,S., Brazier,J.A. and Sugimoto,N. (2017) Topological impact of noncanonical DNA structures on Klenow fragment of DNA polymerase. *Proc. Natl. Acad. Sci. USA.*, **114**, 9605–9610.

30. Liu,W., Lin,C., Wu,G., Dai,J., Chang,T.C. and Yang,D. (2019) Structures of 1:1 and 2:1 complexes of BMVC and MYC promoter G-quadruplex reveal a mechanism of ligand conformation adjustment for G4-recognition. *Nucleic Acids Res.*, **47**, 11931–11942.

31. Ciszewski,L., Ngoc,L.-N., Slater,A., Brennan,A., Williams,H.E.L., Dickson,G., Searle,M.S. and Popplewell,L. (2020) G-quadruplex ligands mediate downregulation of DUX4 expression. *Nucleic Acids Res.*, **48**, 4179–4194.

32. Yang,F., Sun,X., Wang,L., Li,Q., Guan,A., Shen,G. and Tang,Y. (2017) Selective recognition of c-myc promoter G-quadruplex and down-regulation of oncogene c-myc transcription in human cancer cells by 3,8a-disubstituted indolizinone. *RSC Adv.*, **7**, 51965–51969.

33. Zhang,L., Zhang,K.X., Rauf,S., Dong,D., Liu,Y. and Li,J.H. (2016) Single-Molecule analysis of human telomere sequence interactions with G-quadruplex ligand. *Anal. Chem.*, **88**, 4533–4540.

34. Di Somma,S., Amato,J., Iaccarino,N., Pagano,B., Randazzo,A., Portella,G. and Malfitano,A.M. (2019) G-Quadruplex binders induce immunogenic cell death markers in aggressive breast cancer cells. *Cancers*, **11**, 1797.

35. Brown,R.V., Wang,T., Chappeta,V.R., Wu,G.H., Onel,B., Chawla,R., Quijada,H., Camp,S.M., Chiang,E.T., Lassiter,Q.R. *et al.* (2017) The consequences of overlapping G-Quadruplexes and i-Motifs in the platelet-derived growth factor receptor beta core promoter nuclease hypersensitive element can explain the unexpected effects of mutations and provide opportunities for selective targeting of both structures by small molecules to downregulate gene expression. *J. Am. Chem. Soc.*, **139**, 7456–7475.

36. Xu,H., Di Antonio,M., McKinney,S., Mathew,V., Ho,B., O'Neil,N.J., Dos Santos,N., Silvester,J., Wei,V., Garcia,J. *et al.* (2017) CX-5461 is a DNA G-quadruplex stabilizer with selective lethality in BRCA1/2 deficient tumours. *Nat. Commun.*, **8**, 14432.

37. Dhapola,P. and Chowdhury,S. (2016) QuadBase2: web server for multiplexed guanine quadruplex mining and visualization. *Nucleic Acids Res.*, **44**, W277–W283.

38. Mishra,S.K. and Kumar,A. (2016) NALDB: nucleic acid ligand database for small molecules targeting nucleic acid. *Database*, **2016**, baw002.

39. Mishra,S.K., Tawani,A., Mishra,A. and Kumar,A. (2016) G4IPDB: A database for G-quadruplex structure forming nucleic acid interacting proteins. *Sci. Rep.*, **6**, 38144.

40. Li,Q., Xiang,J.F., Yang,Q.F., Sun,H.X., Guan,A.J. and Tang,Y.L. (2013) G4LDB: a database for discovering and studying G-quadruplex ligands. *Nucleic Acids Res.*, **41**, D1115–D1123.

41. Porru,M., Zizza,P., Franceschin,M., Leonetti,C. and Biroccio,A. (2017) EMICORON: A multi-targeting G4 ligand with a promising preclinical profile. *Biochim. Biophys. Acta-Gen. Subj.*, **1861**, 1362–1370.

42. Zeraati,M., Langley,D.B., Schofield,P., Moye,A.L., Rouet,R., Hughes,W.E., Bryan,T.M., Dinger,M.E. and Christ,D. (2018) I-motif DNA structures are formed in the nuclei of human cells. *Nat. Chem.*, **10**, 631–637.

43. Kaiser,C.E., Van Ert,N.A., Agrawal,P., Chawla,R., Yang,D.Z. and Hurley,L.H. (2017) Insight into the complexity of the i-Motif and G-Quadruplex DNA structures formed in the KRAS promoter and subsequent drug induced gene repression. *J. Am. Chem. Soc.*, **139**, 8522–8536.

44. Smith,T.J. (1995) Molview - a program for analyzing and displaying atomic structures on the macintosh personal-computer. *J. Mol. Graph. Model.*, **13**, 122–125.

45. Kim,S., Chen,J., Cheng,T.J., Gindulyte,A., He,J., He,S.Q., Li,Q.L., Shoemaker,B.A., Thiessen,P.A., Yu,B. *et al.* (2021) PubChem in 2021: new data content and improved web interfaces. *Nucleic Acids Res.*, **49**, D1388–D1395.

46. Burley,S.K., Berman,H.M., Christie,C., Duarte,J.M., Feng,Z.K., Westbrook,J., Young,J. and Zardecki,C. (2018) RCSB Protein Data Bank: Sustaining a living digital data resource that enables breakthroughs in scientific research and biomedical education. *Protein Sci.*, **27**, 316–330.

47. Kotov,S., Tremouilhac,P., Jung,N. and Brase,S. (2018) Chemotion-ELN part 2: adaption of an embedded Ketcher editor to advanced research applications. *J. Cheminform.*, **10**, 38.

48. Sanner,M.F. (1999) Python: A programming language for software integration and development. *J. Mol. Graph. Model.*, **17**, 57–61.

49. Hanson,R.M. and Lu,X.J. (2017) DSSR-enhanced visualization of nucleic acid structures in Jmol. *Nucleic Acids Res.*, **45**, W528–W533.

50. O'Boyle,N.M., Banck,M., James,C.A., Morley,C., Vandermeersch,T. and Hutchison,G.R. (2011) Open Babel: An open chemical toolbox. *J. Cheminformatics*, **3**, 33.

51. Forli,S., Huey,R., Pique,M.E., Sanner,M.F., Goodsell,D.S. and Olson,A.J. (2016) Computational protein-ligand docking and virtual drug screening with the AutoDock suite. *Nat. Protoc.*, **11**, 905–919.

52. Trott,O. and Olson,A.J. (2010) Software news and update autodock vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J. Comput. Chem.*, **31**, 455–461.