

Article

OncoSolidDB: An Oncology-Focused Curated Database of Ligand–Target Interactions for Precision Medicine Across Major Solid Cancers

Oussema Khamessi ^{1,2,*} , Rihab Mahjoub ^{1,2} , Ghada Mahjoub ^{1,2}  and Kais Ghedira ^{1,*} 

¹ Laboratory of Bioinformatics, Biomathematics and Biostatistics (BIMS-LAB) (LR16IPT09), Institut Pasteur de Tunis, University of Tunis El Manar, 13 Place Pasteur, BP74, Tunis 1002, Tunisia; rihab.mahjoub@insat.ucar.tn (R.M.); ghada.mahjoub@insat.ucar.tn (G.M.)

² Higher Institute of Biotechnology Sidi Thabet ISBST, BiotechPôle, BP-66, University of Manouba, Ariana 2020, Tunisia

* Correspondence: oussama.khamassi@pasteur.tn (O.K.); kais.ghedira@pasteur.tn (K.G.)

Simple Summary

Cancer treatments increasingly rely on drugs designed to bind specific molecular targets in tumors. However, information about these therapeutic ligands is scattered across multiple databases, making it difficult for researchers to access structured and oncology-focused data in a single place. To address this challenge, we developed OncoSolidDB, a curated and openly accessible database dedicated to ligands targeting solid tumors. The database integrates chemical structures, standardized identifiers, pharmacological annotations, approval history and downloadable protein structural files. Covering 243 ligands across 15 major solid cancer types, OncoSolidDB enables researchers to explore molecular properties, perform computational modeling and support drug repurposing studies. By organizing oncology-specific ligand data into a harmonized platform, this resource aims to facilitate rational drug design, accelerate translational research and improve data accessibility for the scientific community.

Abstract

Background/Objectives: The rapid expansion of targeted therapies has reshaped oncology by exploiting ligand–receptor interactions (LRI) to improve treatment specificity and patient outcomes. However, the data describing these ligands remain fragmented across multiple sources, limiting accessibility for researchers and clinicians. To address this gap, we developed the OncoSolidDB, the first curated and oncology-focused bioinformatics database dedicated to ligands associated with solid malignancies. **Methods:** OncoSolidDB integrates and harmonizes data from reliable repositories, including ChEMBL, DrugBank and the Anti-Cancer Fund, consolidating curated structural, chemical, pharmacological, and clinical annotations along with standardized identifiers. **Results:** The database currently encompasses 243 ligands across 15 major solid tumor types including breast, lung, colorectal, melanoma, prostate, gastric, ovarian, cervical, bladder, esophageal, head and neck, thyroid, pancreatic, renal and liver cancer (Hepatocellular Carcinoma, HCC). Each entry is annotated by standardized identifiers (DrugBank, ChEMBL), approval year, chemical structures (SMILES strings, 2D images), and downloadable protein structure files (PDB format). Temporal coverage spans 1953–2025, enabling exploration of historical trends in oncology drug approvals. The database content is suitable for bioinformatics analysis, molecular docking, virtual screening, ligand-based modeling, and drug repurposing studies. Outputs are available through a freely accessible web interface that supports search browsing by cancer type. **Conclusions:** By consolidating oncology-specific ligand data



Academic Editors: Axel H. Schönthal and Gary Piazza

Received: 3 March 2026

Revised: 12 April 2026

Accepted: 13 April 2026

Published: 12 May 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

into a single, structured platform, OncoSolidDB offers a valuable resource for advancing drug discovery, repurposing strategies, and the rational design of next-generation targeted therapies for solid tumors. OncoSolidDB is accessible via our Bioinformatics Research PortalEinstein.

Keywords: solid tumor ligands; oncology drug database; ligand–receptor interactions; SMILES; bioinformatics; cancer database

1. Introduction

Cancer remains one of the leading causes of morbidity and mortality worldwide, with the global burden continuing to rise at an alarming rate. According to the latest estimates from the World Health Organization and the Global Cancer Observatory (GLOBOCAN), approximately 20 million new cancer cases and 9.7 million cancer-related deaths were recorded in 2022, with projections indicating that these numbers will reach nearly 35 million new cases annually by 2050 [1,2]. Among all cancer types, solid tumors, including breast, lung, colorectal, prostate, and melanoma, account for the majority of cases, representing over 90% of all malignancies and contributing substantially to global clinical and socioeconomic burdens. Regionally, the burden is particularly pronounced in North Africa and the Middle East, where cancer incidence is expected to double by 2040, driven by population growth, aging and lifestyle transitions [3,4]. Despite significant advances in oncology, including the development of chemotherapies, targeted therapies, and immunotherapies, many solid tumors continue to exhibit intrinsic or acquired drug resistance, profound tumor heterogeneity and off-target toxicities that limit long-term efficacy and patient outcomes [5,6].

Bioinformatics has emerged as a key discipline for addressing these challenges by enabling the study of ligand–receptor interactions at the molecular level. Small-molecule ligands, peptides, or biologics that bind to cellular receptors play a central role in modulating oncogenic pathways and form the backbone of precision oncology. Understanding their chemical diversity, structural features, and interaction profiles is critical for rational drug discovery, drug repurposing, and the design of next-generation therapies. However, such efforts require reliable, comprehensive, and oncology-specific datasets that integrate chemical, pharmacological, and structural information in a user-friendly and accessible manner. Several public chemical and pharmacological databases provide valuable information for cancer drug research, including DrugBank [7], ChEMBL [8–10], and the Therapeutic Target Database (TTD) [11].

While these repositories offer rich annotations on drug molecules, protein targets and bioactivity data, they are designed as general-purpose resources rather than oncology-specialized platforms. Consequently, cancer-related ligands are often mixed with non-cancer compounds, making it difficult to extract oncology-specific datasets systematically. Moreover, structural bioinformatics features such as standardized protein ligand complex files (in PDB format), curated SMILES strings, and cancer-type specific drug associations are often incomplete, scattered or missing [12,13].

This lack of oncology-focused integration hinders researchers who need streamlined access to curated ligand datasets for structural modeling, virtual screening, or comparative pharmacology studies. To address these gaps, we developed OncoSolidDB, a dedicated and integrative database focusing exclusively on ligands associated with solid tumors [14–16]. OncoSolidDB consolidates data from reliable repositories, including DrugBank, ChEMBL, and the Anti-Cancer Fund and harmonizes them into a cancer-specific framework. The database currently hosts 243 ligands covering 15 major solid tumor types, including breast,

cervical, bladder, colorectal, esophageal, gastric, head and neck, liver (Hepatocellular Carcinoma), lung, melanoma, ovarian, pancreatic, prostate, renal and thyroid cancers. Each ligand entry is enriched with structural (SMILES, 2D images, and PDB files), chemical, pharmacological, and clinical annotations, as well as cross-references to major identifiers (DrugBank and ChEMBL). In addition, the database captures temporal approval data (1953–2025), enabling exploration of historical trends in oncology drug development. Uniquely, OncoSolidDB provides an interactive web-based interface that supports searching and browsing by cancer type, ligand annotation, and visualization of ligand overlaps across cancers, making the resource practical for both exploratory and hypothesis-driven research [17–19]. The overarching objective of this work is to present the solid tumor section of OncoSolidDB and demonstrate its potential as a specialized, oncology-focused ligand repository for structural bioinformatics and drug discovery applications.

To the best of our knowledge, OncoSolidDB represents the first curated and oncology-focused database that integrates chemical, structural, pharmacological, and clinical data specifically for ligands targeting solid tumors, providing a unified platform not available in general-purpose chemical repositories such as DrugBank or ChEMBL. By offering curated, cancer-specific ligand datasets, the database facilitates downstream computational workflows such as molecular docking, virtual screening, ligand-based modeling, and receptor-ligand network analysis [20–22]. Its oncology-specific orientation provides researchers with a more targeted resource than general-purpose chemical databases, thereby reducing time and effort in data preprocessing. Ultimately, OncoSolidDB, publicly available via https://liganddb.pythonanywhere.com/solid_cancer (accessed on 14 January 2026) and through our Bioinformatic Research Portal (public web server titled “Einstein,” <http://196.203.66.81/>, accessed on 14 January 2026) aims to accelerate the discovery of novel therapeutic strategies, support drug repurposing, and contribute to the rational design of ligands in the fight against solid tumors.

2. Materials and Methods

2.1. Data Collection and Source Integration

OncoSolidDB was developed through a systematic and comprehensive collection of ligands associated with solid tumors. The primary source of ligand information was the Anti-Cancer Fund (<https://www.anticancerfund.org/>) database [14], which provides data on anti-cancer agents and their clinical indications. Additional data were retrieved from DrugBank (<https://go.drugbank.com/> v5.0) and ChEMBL (<https://www.ebi.ac.uk/chembl/>, accessed on 14 January 2026, v33) by querying cancer-specific indications, such as “breast cancer”, “lung cancer”, and “pancreatic cancer”.

Each ligand entry underwent rigorous manual curation, ensuring correct association with the corresponding cancer type, verification of availability in the DrugBank and ChEMBL databases, accurate chemical identifiers and SMILES strings, and the removal of duplicate entries and harmonization of identifiers across sources following established curation practices for chemical databases [23–26].

For each ligand, the following data were collected and standardized: ligand name, approval year, DrugBank and ChEMBL IDs, cancer type, SMILES representation, 2D chemical structure and downloadable 3D PDB files.

2.2. Selection Criteria

A rigorous selection process was implemented to guarantee that only high-confidence, oncology-relevant ligands were included in OncoSolidDB. The following inclusion and exclusion criteria were applied to all candidate compounds retrieved from the source databases.

2.2.1. Inclusion Criteria for Ligand Selection

To ensure the reliability, consistency, and scientific relevance of the data integrated into OncoSolidDB, a set of stringent inclusion criteria was established for ligand selection. Only compounds meeting all predefined conditions were retained for further analysis and database integration. Specifically, ligands were required to be either approved therapeutic agents or investigational molecules associated with solid tumors, with at least one validated cancer indication reported in curated resources such as DrugBank, ChEMBL, and Anticancer Fund. Additionally, each compound had to possess a valid and standardized chemical structure representation, including a SMILES string, to enable structural normalization and computational analyses such as docking and QSAR modeling. Cross-referencing across databases was ensured through the use of unique identifiers (DrugBank ID and ChEMBL ID), guaranteeing data traceability and interoperability. Finally, only ligands linked to at least one of the fifteen predefined solid tumor types included in the scope of OncoSolidDB were considered, thereby maintaining the thematic coherence and applicability of the database.

2.2.2. Exclusion Criteria

To maintain the quality, accuracy, and biological relevance of the dataset, specific exclusion criteria were applied during the curation process. Entries were discarded if they were exclusively associated with hematological malignancies rather than solid tumors, ensuring alignment with the database's scope. Compounds lacking a valid or complete chemical structure were also excluded, as this prevented proper structural standardization and downstream computational analyses. In addition, duplicate entries across multiple databases that could not be reliably reconciled using standardized identifiers were discarded to avoid redundancy and data inconsistency. Molecules with ambiguous, conflicting, or unsupported cancer indications were not retained, thereby ensuring the reliability of therapeutic annotations. Furthermore, non-therapeutic compounds, including diagnostic agents, imaging probes, and unrelated chemical entities, were excluded from the dataset. Following the rigorous application of these criteria, the final curated dataset comprised 243 ligands associated with 15 major solid tumor types, providing a robust and high-confidence resource for subsequent analyses.

2.3. Data Curation and Standardization Pipeline

Raw data collected from multiple sources were subjected to a multi-step curation and standardization pipeline to ensure data quality, consistency, and interoperability. First, ligand entries were screened for relevance based on predefined inclusion and exclusion criteria. Subsequently, duplicate records originating from different databases were identified and resolved through identifier harmonization using DrugBank and ChEMBL IDs. Next, chemical information was standardized by verifying ligand names, correcting inconsistencies, and ensuring the presence of valid SMILES representations. Cross-referencing across databases was performed to enrich each entry with consistent annotations, including cancer type, approval status, and structural information. Quality control procedures included manual validation of cancer indications, verification of structural data integrity, and removal of incomplete or ambiguous entries. The curated dataset was further checked for consistency in formatting and annotation to facilitate downstream computational analyses. This systematic curation process ensured the generation of a high-quality, non-redundant, and analysis-ready dataset suitable for structural bioinformatics and drug-discovery applications. All curation steps were performed following standardized and reproducible procedures to ensure dataset reliability.

2.4. Workflow of Database Construction

The overall workflow of OncoSolidDB construction is summarized in Figure 1. This schematic representation outlines the sequential steps of data acquisition, eligibility filtering, data curation and harmonization, structural processing, and integration into a web-accessible database. The workflow highlights the transformation of heterogeneous raw data into a curated, standardized, and computationally exploitable resource.

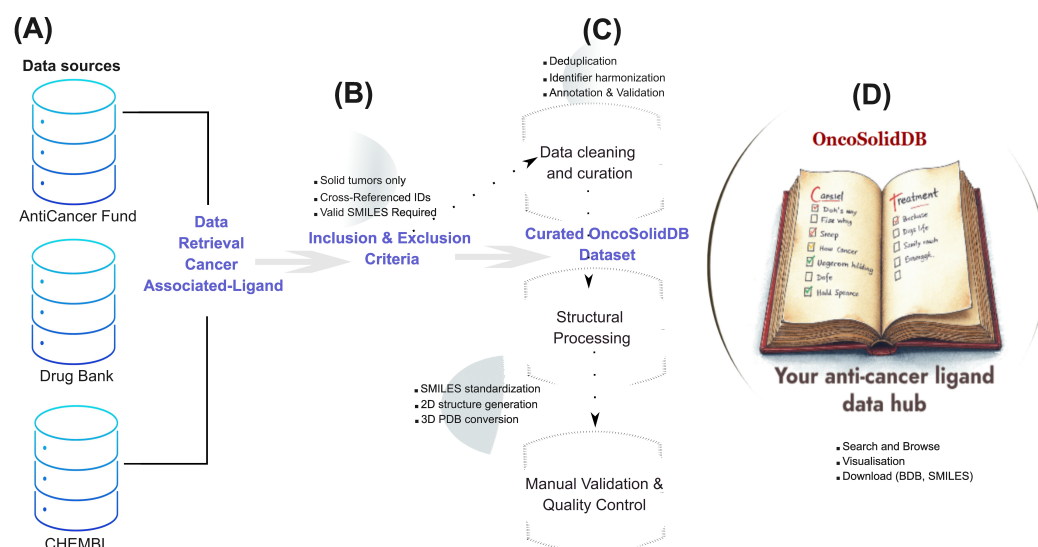


Figure 1. Schematic overview of the OncoSolidDB construction pipeline. The figure illustrates the multi-step workflow implemented to build the OncoSolidDB database, organized into four sequential panels. **(A)** Raw ligand data were retrieved from three primary repositories: DrugBank, ChEMBL, and the Anticancer Fund. Cancer-specific queries were performed using disease-related keywords to initially extract candidate ligands associated with solid tumor indications. **(B)** Retrieved ligands were subjected to stringent inclusion and exclusion criteria. Inclusion required that ligands be approved or investigational agents with validated solid tumor indications, valid chemical structures (SMILES strings), and cross-referenced identifiers (DrugBank and ChEMBL IDs). Exclusion criteria removed compounds associated exclusively with hematological malignancies, those lacking structural information, duplicate entries, molecules with ambiguous indications and non-therapeutic agents. **(C)** Following eligibility filtering, a multi-step curation pipeline was applied, including manual validation of cancer indications, structural standardization using RDKit and Open Babel, SMILES-to-PDB conversion, cross-database identifier harmonization, and quality control procedures to ensure data integrity and consistency. **(D)** The final curated and standardized ligand collection was integrated into a web-accessible platform, providing users with search, browse, and download functionalities, along with interactive visualizations, and structural data outputs. This workflow ensured the generation of a high-quality, non-redundant, and analysis-ready resource dedicated to ligands targeting solid tumors.

2.5. SMILES and Structural Processing

To ensure computational compatibility for structural bioinformatics applications, including molecular docking, virtual screening and molecular dynamics simulations, ligands were standardized into SMILES strings and converted into 2D images and 3D PDB files. SMILES extraction was performed using ChEMBL annotations [24], while structural generation was conducted using RDKit (v2023.09) and Open Babel (v3.1.1) [13,26,27]. The conversion from SMILES to PDB leveraged Open Babel's interconversion capabilities, producing downloadable PDB files for each ligand. To facilitate reproducible ligand preparation, all workflows were executed using Python 3.11.5 within a Linux Ubuntu 22.04.4 LTS (Jammy) environment, ensuring robust handling of chemical structures, and integration into computational pipelines [27].

2.6. Web Interface Development

The OncoSolidDB web interface was developed to provide intuitive and interactive access to the curated dataset described in the construction workflow (Section 2.4). The platform translates processed data into a user-friendly environment for visualization, exploration and download. The overall technical architecture and implementation of the platform are illustrated in Figure 2.

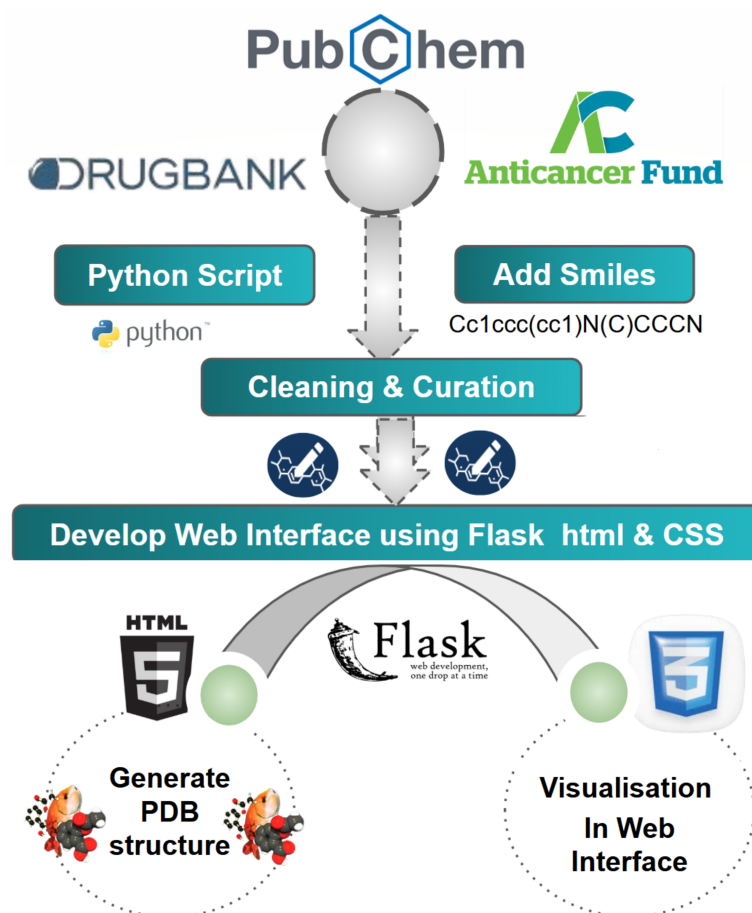


Figure 2. Technical architecture and implementation of the OncoSolidDB web platform. This figure illustrates the workflow integrating multiple data sources, including PubChem, DrugBank and the Anticancer Fund, into OncoSolidDB. Ligand data is processed using Python scripts for cleaning and curation, followed by the development of a web interface built with Flask, HTML5 and CSS3. The platform supports the generation of PDB structures and interactive visualization of chemical data within the web interface.

The application was implemented using the Flask microframework (v2.3), which handles backend logic, routing, and dynamic data rendering. The frontend was developed using HTML5 and CSS3 to ensure a responsive and accessible user experience. The system architecture integrates curated ligand data into a structured database, enabling efficient querying based on cancer type, ligand name, or standardized identifiers such as DrugBank and ChEMBL IDs. Dynamic content delivery is achieved through template-based rendering, allowing seamless interaction between the database and the user interface. The web platform provides several core functionalities, including the ability to browse ligands by solid tumor category, visualize ligand annotations and metadata, display 2D chemical structures derived from SMILES strings, access these SMILES strings for computational workflows and download ligand structures in PDB format for more advanced use.

By decoupling data processing (Sections 2.1–2.5) from visualization, the web interface ensures efficient data access while maintaining scalability and usability for bioinformatics and drug-discovery applications.

Overview of the technical components used in the development of OncoSolidDB. The figure illustrates the integration of data-processing tools (Python, RDKit, Open Babel) with the web framework (Flask) and frontend technologies (HTML/CSS), enabling interactive visualization and access to curated ligand data [28].

2.7. Data Analysis and Visualization

To explore the characteristics and potential applications of the curated dataset, several analytical and visualization approaches were implemented. Descriptive statistical analyses were performed to assess the distribution of ligands across cancer types and approval years. These analyses were used to generate summary tables and histograms presented in the Section 3. Ligand overlap analysis across solid tumor types was conducted to identify compounds associated with multiple cancer indications. In this context, an “overlap” was defined as the presence of a single ligand annotated with more than one cancer type in the curated dataset. Based on this definition, a binary association matrix was constructed, linking ligands to their corresponding cancer types. This matrix was used to generate an UpSet plot, which provides a scalable alternative to Venn diagrams for visualizing intersections among multiple sets. The analysis highlights shared ligands across different cancer types, revealing potential opportunities for drug repurposing and multi-indication therapeutic strategies. All analyses and visualizations were performed using Python (version 3.11), with relevant libraries for data manipulation and visualization [29].

2.8. Availability and Access

OncoSolidDB is freely available for the scientific community through those links: https://liganddb.pythonanywhere.com/solid_cancer (accessed on 14 January 2026) and via our Bioinformatics Research Portal “Einstein” <http://196.203.66.81/oncosoliddb/> (accessed on 14 January 2026).

3. Results

3.1. Database Content

The solid tumor dataset of OncoSolidDB includes 243 ligands across 15 cancer types (Figure 3). The distribution of ligands reflects the relative research focus and clinical relevance of each solid tumor type, with lung and breast cancers accounting for the highest proportion of ligands. This distribution reflects both the intensity of research efforts and the clinical burden associated with different cancers. Notably, lung cancer (54 ligands, 18.9%) and breast cancer (48 ligands, 16.8%) account for the highest proportions, consistent with their high global incidence, extensive molecular characterization, and the availability of targeted therapies. Intermediate representation is observed for ovarian cancer and melanoma (26 ligands each, 9.1%), followed by prostate cancer (21 ligands, 7.4%) and pancreatic cancer (17 ligands, 6%) (Table S1).

These cancer types benefit from increasing interest in targeted treatment strategies and biomarker discovery. In contrast, cancers such as gastric cancer and head and neck cancer (3 ligands each, 1.1%) remain underrepresented, which may reflect limited availability of validated ligands–receptor data or fewer structurally characterized therapeutic compounds. Other cancer types, including cervical cancer (16 ligands, 5.6%), renal cancer (15 ligands, 5.3%), colorectal cancer (14 ligands, 4.9%), hepatitis-related cancers (13 ligands, 4.6%), esophageal cancer (11 ligands, 3.9%), thyroid cancer (10 ligands, 3.5%), and bladder

cancer (8 ligands, 2.8%), show moderate representation within the dataset. This variability highlights disparities in research focus and therapeutic development across tumor types.

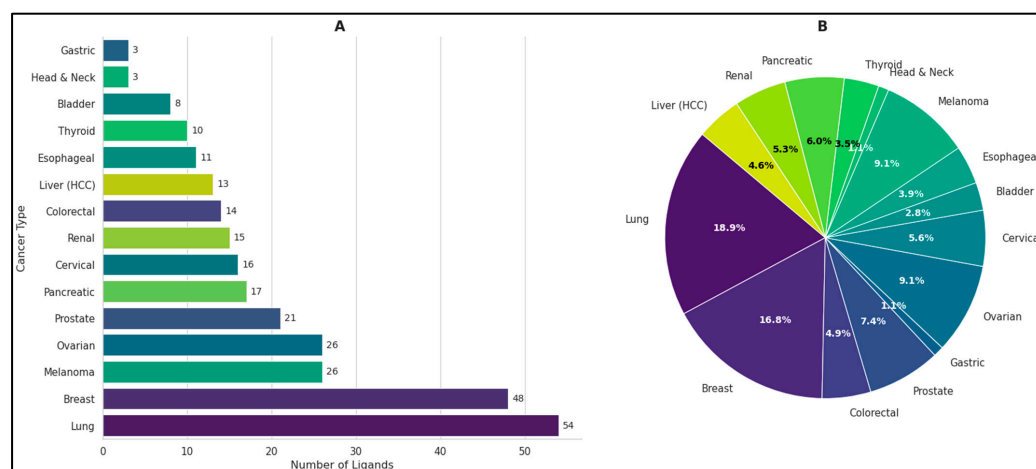


Figure 3. Distribution of ligands across solid tumor types. **(A)** Number of ligands associated with each solid tumor type in OncoSolidDB. Lung cancer and breast cancer represent the highest proportions, reflecting their high incidence and research focus. Intermediate representation is seen for ovarian cancer and melanoma, as well as prostate cancer and pancreatic cancer. Gastric and head and neck cancers are underrepresented, possibly due to fewer validated or structurally characterized ligands. **(B)** Relative proportions of ligands by cancer type, highlighting variability in dataset coverage across tumor types.

This curated dataset provides a comprehensive resource for researchers to explore ligand–receptor interactions, enabling structural bioinformatics analyses, virtual screening and drug-repurposing studies across a wide range of clinically important solid tumors [30].

3.2. Temporal Trends and Chemical Complexity

The distribution of ligand approval years in OncoSolidDB shows a heterogeneous temporal pattern across the period 1953–2025 (Figure 4A). Only a limited number of ligands are associated with approval dates prior to the late 1980s, corresponding to early anticancer agents. From the 1990s onward, a progressive increase in the number of approved ligands is observed, with a pronounced accumulation after the year 2000. The highest density of approvals occurred during the 2015–2025 period, indicating that a substantial proportion of ligands included in OncoSolidDB are associated with recent regulatory approvals.

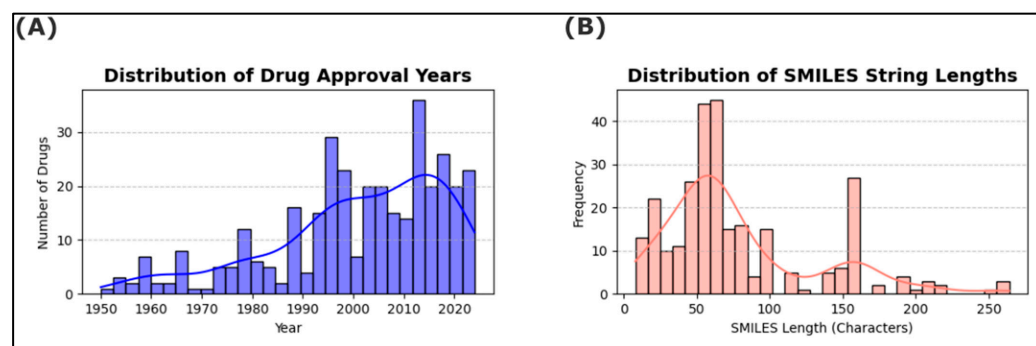


Figure 4. Temporal Trends and Chemical Complexity of Ligands in OncoSolidDB. **(A)** Histogram showing the distribution of ligand approval years in OncoSolidDB spanning 1953 to 2025. **(B)** Histogram of the SMILES string lengths illustrating the molecular complexity of ligands included in the database.

corresponding to the number of ligands shared among the specific combination of cancer types indicated in the matrix below. The bars are ordered in decreasing intersection size, with the tallest bars representing the most frequent intersection patterns. The bottom panel (B) displays a dot matrix in which each row corresponds to one of the 15 cancer types, labeled along the left side together with their total number of associated ligands. Each column represents a specific intersection. Filled colored dots indicate the cancer types involved in a given intersection and vertical connecting lines highlight ligands shared across multiple cancer types. Single filled dots without connecting lines correspond to ligands unique to a single cancer type.

In contrast, certain ligands are uniquely associated with specific cancer types, such as gastric, head and neck, or thyroid cancers, indicating highly specialized therapeutic roles and potential relevance for precision medicine. Each intersection corresponds to a specific combination of cancer types, defined by the pattern of connected dots in Figure 5B, while the corresponding bar in Figure 5A represents the number of ligands associated with that combination. Overall, this analysis demonstrates that while some ligands act on universal cancer mechanisms, others exhibit high specificity, underscoring the importance of integrating both broad-spectrum and targeted therapeutic approaches in oncology.

The analysis reveals that the majority of ligands in OncoSolidDB are cancer type-specific, as evidenced by the predominance of single-dot intersections. This observation highlights the strong molecular specificity underlying targeted therapies in precision oncology. Nevertheless, a subset of ligands is shared across multiple tumor types, reflecting common oncogenic mechanisms.

3.4. Target Family Distribution

Functional classification revealed that kinase-associated ligands represent the dominant category within OncoSolidDB, reflecting the central role of kinase dysregulation in solid tumor pathogenesis. Receptor tyrosine kinases and intracellular signaling enzymes constitute the majority of annotated targets, followed by nuclear hormone receptors and proteasome-associated proteins. This distribution highlights the strong therapeutic focus on signaling-pathway modulation in precision oncology. Several ligands in OncoSolidDB are specifically kinase-targeting molecules, including inhibitors of EGFR, VEGFR, HER2, BRAF, and CDK family kinases. These ligands represent the dominant pharmacological category in the database, reflecting the central role of kinase dysregulation in solid tumor pathogenesis.

3.5. Temporal Approval Trends

Temporal analysis demonstrated a marked increase in solid tumor-targeting ligands following the year 2000, corresponding to the emergence of targeted therapies. A second acceleration phase was observed after 2010, coinciding with advancements in molecular diagnostics and personalized oncology. These findings underscore the growing complexity and specificity of anticancer drug development.

3.6. Database Overview and Interrogation

OncoSolidDB provides an interactive web-based interface that enables systematic interrogation of ligands associated with solid tumors (Figure 6). Users can explore the database through multiple entry points, including browsing by cancer type (Figure 6A) and searching by ligand name or standardized identifiers.

For each ligand, the database displays curated information such as cancer indication, approval year, DrugBank (Figure 6C), ChEMBL (Figure 6D) identifiers, SMILES representation, and two-dimensional (2D) chemical structure, as represented in Figure 6B. In addition, downloadable three-dimensional (3D) ligand structures in PDB format are provided to support structural bioinformatics analyses.

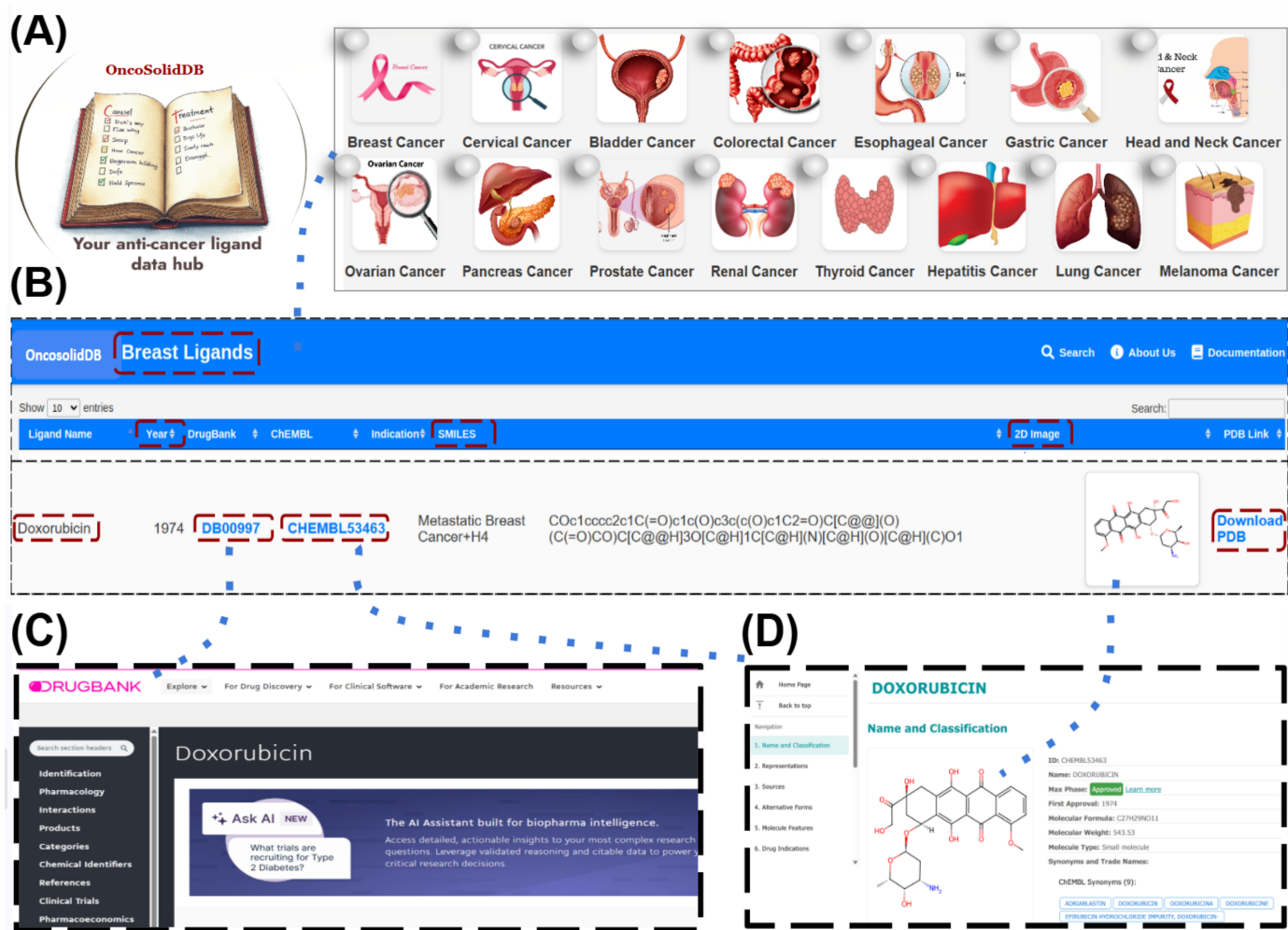


Figure 6. Overview of the OncoSolidDB web interface. The figure highlights the main database pages, including cancer-type browsing, ligand detail views, and structural data visualization. (A) browsing ligands by cancer type, (B) viewing ligand 2D chemical structures, (C) accessing DrugBank identifiers, and (D) accessing ChEMBL identifiers. Dashed arrows indicate the trend direction. Users can explore ligand-specific data, download 3D PDB structures and navigate between global overviews and detailed molecular-level information.

The database interface includes dedicated pages for cancer-specific ligand listings, individual ligand detail views, and summary visualizations illustrating ligand distribution across tumor types https://liganddb.pythonanywhere.com/cancer_type/breast or http://196.203.66.81/oncosolidb/cancer_type/breast (accessed on 14 January 2026). These features allow users to navigate between global overviews and detailed molecular-level information, facilitating efficient retrieval and inspection of solid tumor-associated ligands [33,34].

4. Discussion

OncoSolidDB provides a specialized, oncology-focused view on ligands targeting solid tumors, addressing a gap left by general-purpose chemical databases. Unlike broader repositories such as DrugBank and ChEMBL, which cover a wide range of bioactive molecules across multiple therapeutic areas, OncoSolidDB contextualizes ligands by cancer type, enabling translational and computational oncology research [35,36]. This specificity is particularly valuable in oncology, where therapeutic efficacy depends on tumor biology, molecular targets, and clinical context [37,38]. In this way, OncoSolidDB complements

but does not duplicate the scope of larger repositories. Furthermore, OncoSolidDB addresses the limitations of current oncology resources by bridging the gap between general drug-target mapping and tumor-specific application. While the Anticancer Fund focuses on clinical outcomes and canSAR on drug discovery informatics, OncoSolidDB provides a specialized focus on solid tumors, a level of detail often missing in broader repositories like the Therapeutic Target Database [39]. The development of OncoSolidDB addresses a critical gap in oncology bioinformatics by centralizing ligand–target information specific to solid malignancies. While several general drug repositories exist, none are specifically curated for solid tumor-focused ligand exploration with integrated structural and regulatory annotations [40]. In its current version, OncoSolidDB comprises 243 curated ligands associated with 15 major solid tumor types. While this number may appear limited compared to large-scale chemical repositories, the dataset was intentionally curated to prioritize data quality, structural consistency, and cross-database validation. Only well-characterized ligands with reliable annotations and complete structural information were included, ensuring high confidence in downstream computational applications. The database architecture has been designed to support continuous updates, and future releases will incorporate additional investigational compounds, ligands in clinical trials and newly approved therapeutics, thereby progressively expanding the dataset. The predominance of kinase-targeting ligands within the database reflects the central role of aberrant signaling pathways in tumor progression [41]. The observed temporal acceleration in drug approvals aligns with the paradigm shift toward molecularly targeted therapies and precision oncology strategies. Beyond data aggregation, OncoSolidDB provides a structured framework suitable for computational modeling, network pharmacology and AI-driven drug repurposing.

By enabling standardized ligand retrieval and cross-cancer comparison, this platform may facilitate hypothesis generation and translational research. In addition to data curation, preliminary analytical approaches were implemented to explore the structure and utility of the dataset. For instance, ligand overlap analysis across cancer types enables the identification of compounds associated with multiple tumor indications, highlighting potential opportunities for drug repurposing. Although the current analyses remain primarily descriptive, they demonstrate the potential of OncoSolidDB as a foundation for more advanced computational studies, including network pharmacology, predictive modeling, and integrative multi-omics analyses. Future expansions may include integration of genomic mutation profiles, resistance-associated variants, and multi-omics annotations to further enhance its precision oncology applications. An important extension of the database will involve the integration of corresponding target proteins and ligand–protein interaction data. Future developments will incorporate protein annotations from resources such as UniProt and the Protein Data Bank (PDB), including experimentally resolved ligand–protein complexes when available. In cases where structural data are lacking, predicted protein models may be integrated to support structure-based analyses [42]. This extension will enable a more comprehensive representation of ligand–target interactions and further enhance the utility of OncoSolidDB for structural bioinformatics and drug discovery applications. Similar domain-specific bioinformatics initiatives, such as DisintegrinDB (<http://196.203.66.81/>, accessed on 14 January 2026), have demonstrated the value of curated, disease- or molecule-focused databases in accelerating biomedical discovery [43]. While OncoSolidDB is primarily a curated bioinformatics resource, its cancer-type-specific ligand annotations can facilitate the design of targeted therapies and support precision medicine initiatives by highlighting potential drug candidates for specific tumor profiles [44,45].

While several large chemical and pharmacological databases exist, including DrugBank, ChEMBL and the Anticancer Fund, these resources were designed as general-purpose repositories rather than oncology-specific platforms. In these databases, cancer-related ligands are typically embedded within very large collections of bioactive molecules covering diverse therapeutic areas, meaning that extracting a curated dataset specifically associated with solid tumors requires extensive filtering and manual preprocessing. OncoSolidDB addresses this limitation by providing a domain-specific database dedicated exclusively to ligands targeting solid tumors, distinguished by several key features: a cancer-type specific organization enabling rapid identification of ligands associated with particular tumor types; integrated structural information including SMILES representations, 2D chemical structures and downloadable 3D PDB files; curated cross-references linking DrugBank and ChEMBL identifiers and data formats directly usable for structural bioinformatics workflows such as molecular docking; virtual screening; and molecular dynamics simulations. To our knowledge, no existing resource simultaneously provides curated ligand annotations, structural representations (SMILES and PDB), and cancer-type associations within a single unified platform specifically focused on solid tumors.

5. Conclusions

OncoSolidDB consolidates curated data on ligands targeting solid tumors into a single, freely accessible resource. Although the current release includes a focused set of 243 ligands, emphasis was placed on data quality, curation rigor and structural completeness to ensure reliability for computational applications. By combining structural, chemical and pharmacological annotations with interactive visualization, it bridges the gap between bench science and computational oncology.

Looking ahead, several perspectives are envisioned to extend the scope and impact of OncoSolidDB. These include the expansion of the dataset through the integration of additional investigational compounds, clinical trial candidates, and newly approved therapeutics, as well as the incorporation of ligand–target interaction data and protein structural information. Future developments will also focus on generating AI-ready datasets to support the application of high-performance computing HPC approaches in drug discovery [46], machine learning and deep learning approaches for predicting ligand–receptor interactions, optimizing drug design and advancing precision oncology [47–49]. Collectively, these future developments will strengthen OncoSolidDB as a dynamic and evolving resource, ultimately empowering the cancer research community in the pursuit of more effective targeted therapies. Furthermore, the integration of analytical and visualization approaches demonstrates the potential of the database not only as a curated repository but also as a resource for generating biologically and pharmacologically relevant insights. OncoSolidDB provides a structured platform not only for computational analyses but also to guide precision oncology research, supporting rational drug design and potential patient-specific therapeutic strategies.

Supplementary Materials: The following supporting information can be downloaded at: <https://www.mdpi.com/article/10.3390/cancers18101559/s1>, Table S1: Distribution of ligands across solid tumor types.

Author Contributions: Conceptualization, O.K. and K.G.; methodology, O.K. and K.G.; software, O.K.; validation, O.K., K.G., R.M. and G.M.; formal analysis, O.K. and R.M.; investigation, O.K.; resources, K.G.; data curation, O.K. and G.M.; writing—original draft preparation, O.K., K.G., R.M. and G.M.; writing—review and editing, O.K., K.G., R.M. and G.M.; visualization, O.K.; supervision, K.G.; project administration, O.K. and K.G.; funding acquisition, K.G. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by grants from the Tunisian Ministry of Higher Education and Scientific Research (LR16IPT09) and the Pasteur Institute of Tunisia.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The database is freely available at: https://liganddb.pythonanywhere.com/solid_cancer and <http://196.203.66.81/oncosoliddb/> (accessed on 14 January 2026). All curated datasets used in this study are accessible through the web interface.

Acknowledgments: Thanks are addressed to Samia Menif, General Director of Institut Pasteur de Tunis for constant encouragement.

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

Abbreviations

AI	Artificial Intelligence
ChEMBL	Chemical database of bioactive molecules with drug-like properties
CSS	Cascading Style Sheets
DDR1	Discoidin Domain Receptor 1
FAIR	Findable, Accessible, Interoperable, Reusable
HCC	Hepatocellular Carcinoma
HER2	Human Epidermal Growth Factor Receptor 2
HTML	HyperText Markup Language
HPC	High-performance computing
LRI	Ligand–Receptor Interactions
ML	Machine Learning
PDB	Protein Data Bank
QSAR	Quantitative Structure–Activity Relationship
SMILES	Simplified Molecular Input Line Entry System
TTD	Therapeutic Target Database
UniProt	Universal Protein Resource
VEGF	Vascular Endothelial Growth Factor

References

- Bray, F.; Laversanne, M.; Sung, H.; Ferlay, J.; Siegel, R.L.; Soerjomataram, I.; Jemal, A. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J. Clin.* **2024**, *74*, 229–263. [\[CrossRef\]](#)
- World Health Organization. Global Cancer Burden Growing, Amidst Mounting Need for Services. WHO News Release. 2024. Available online: <https://www.who.int/news/item/01-02-2024-global-cancer-burden-growing--amidst-mounting-need-for-services> (accessed on 24 December 2025).
- Siegel, R.L.; Miller, K.D.; Fuchs, H.E.; Jemal, A. Cancer statistics, 2022. *CA Cancer J. Clin.* **2022**, *72*, 7–33. [\[CrossRef\]](#)
- Alotaibi, H.; Anis, A.M.; Alloghbi, A.; Alshammari, K. Oncology Early-Phase Clinical Trials in the Middle East and North Africa: A Review of the Current Status, Challenges, Opportunities, and Future Directions. *J. Immunother. Precis. Oncol.* **2024**, *7*, 178–189. [\[CrossRef\]](#) [\[PubMed\]](#)
- Hanahan, D.; Weinberg, R.A. Hallmarks of cancer: The next generation. *Cell* **2011**, *144*, 646–674. [\[CrossRef\]](#)
- Lawler, M.; Keeling, P.; Kholmanskikh, O.; Minnaard, W.; Moehlig-Zuttermeister, H.; Normanno, N.; Philip, R.; Popp, C.; Salgado, R.; Santiago-Walker, A.E.; et al. Empowering effective biomarker-driven precision oncology: A call to action. *Eur. J. Cancer* **2024**, *209*, 114225. [\[CrossRef\]](#)
- Wishart, D.S.; Feunang, Y.D.; Guo, A.C.; Lo, E.J.; Marcu, A.; Grant, J.R.; Sajed, T.; Johnson, D.; Li, C.; Sayeeda, Z.; et al. DrugBank 5.0: A major update to the DrugBank database for 2018. *Nucleic Acids Res.* **2018**, *46*, D1074–D1082. [\[CrossRef\]](#)
- Mendez, D.; Gaulton, A.; Bento, A.P.; Chambers, J.; De Veij, M.; Félix, E.; Magarinos, M.P.; Mosquera, J.F.; Mutowo, P.; Nowotka, M.; et al. ChEMBL: Towards direct deposition of bioassay data. *Nucleic Acids Res.* **2019**, *47*, D930–D940. [\[CrossRef\]](#)
- Gaulton, A.; Bellis, L.J.; Bento, A.P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; et al. ChEMBL: A large-scale bioactivity database for drug discovery. *Nucleic Acids Res.* **2012**, *40*, D1100–D1107. [\[CrossRef\]](#) [\[PubMed\]](#)

10. Spjuth, O. Novel applications of Machine Learning in cheminformatics. *J. Cheminform.* **2018**, *10*, 46. [CrossRef]
11. Wang, Y.; Zhang, S.; Li, F.; Zhou, Y.; Zhang, Y.; Wang, Z.; Zhang, R.; Zhu, J.; Ren, Y.; Tan, Y.; et al. Therapeutic target database 2020: Enriched resource for facilitating research and early development of targeted therapeutics. *Nucleic Acids Res.* **2020**, *48*, D1031–D1041. [CrossRef] [PubMed]
12. Bento, A.P.; Hersey, A.; Félix, E.; Landrum, G.; Gaulton, A.; Atkinson, F.; Bellis, L.J.; De Veij, M.; Leach, A.R. An open source chemical structure curation pipeline using RDKit. *J. Cheminform.* **2020**, *12*, 51. [CrossRef]
13. O'Boyle, N.M.; Banck, M.; James, C.A.; Morley, C.; Vandermeersch, T.; Hutchison, G.R. Open Babel: An open chemical toolbox. *J. Cheminform.* **2011**, *3*, 33. [CrossRef]
14. Anticancer Fund. Anticancer Database. 2025. Available online: <https://www.anticancerfund.org/en/trials> (accessed on 24 December 2025).
15. Python Software Foundation. Python Programming Language. Available online: <https://www.python.org/> (accessed on 24 December 2025).
16. Grinberg, M.; Ronacher, A. Contributors, Flask: A Microframework for Python. Available online: <http://flask.pocoo.org> (accessed on 24 December 2025).
17. W3C. HTML5 Specification & CSS: Cascading Style Sheets. Available online: <https://www.w3.org/> (accessed on 24 December 2025).
18. Chen, H.; Engkvist, O.; Wang, Y.; Olivecrona, M.; Blaschke, T. The rise of deep learning in drug discovery. *Drug Discov. Today* **2018**, *23*, 1241–1250. [CrossRef]
19. Gasteiger, J. Chemoinformatics: A new field with a long tradition. *Anal. Bioanal. Chem.* **2006**, *384*, 57–64. [CrossRef] [PubMed]
20. Lavecchia, A. Machine-learning approaches in drug discovery: Methods and applications. *Drug Discov. Today* **2015**, *20*, 318–331. [CrossRef]
21. Berman, H.M.; Battistuz, T.; Bhat, T.N.; Bluhm, W.F.; Bourne, P.E.; Burkhardt, K.; Feng, Z.; Gilliland, G.L.; Iype, L.; Jain, S.; et al. The Protein Data Bank. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **2002**, *58*, 899–907. [CrossRef]
22. Kitchen, D.B.; Decornez, H.; Furr, J.R.; Bajorath, J. Docking and scoring in virtual screening for drug discovery: Methods and applications. *Nat. Rev. Drug Discov.* **2004**, *3*, 935–949. [CrossRef] [PubMed]
23. Fourches, D.; Muratov, E.; Tropsha, A. Trust, but verify: On the importance of chemical structure curation in cheminformatics and QSAR modeling research. *J. Chem. Inf. Model.* **2010**, *50*, 1189–1204. [CrossRef]
24. Wilkinson, M.D.; Dumontier, M.; Aalbersberg, I.J.; Appleton, G.; Axton, M.; Baak, A.; Blomberg, N.; Boiten, J.W.; da Silva Santos, L.B.; Bourne, P.E.; et al. The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* **2016**, *3*, 160018. [CrossRef] [PubMed]
25. RDKit: Open-Source Cheminformatics Software. Available online: <https://www.rdkit.org> (accessed on 17 December 2025).
26. Van Rossum, G.; Drake, F.L. *Python 3 Reference Manual*; Python Software Foundation: Beaverton, OR, USA, 2009.
27. Trott, O.; Olson, A.J. AutoDock Vina: Improving speed and accuracy. *J. Comput. Chem.* **2010**, *31*, 455–461. [CrossRef]
28. Grinberg, M. *Flask Web Development*; O'Reilly Media: Sebastopol, CA, USA, 2018.
29. Fielding, R.T.; Taylor, R.N. Principled design of the modern Web architecture. *ACM Trans. Internet Technol.* **2002**, *2*, 115–150. [CrossRef]
30. Santos, R.; Ursu, O.; Gaulton, A.; Bento, A.P.; Donadi, R.S.; Bologa, C.G.; Karlsson, A.; Al-Lazikani, B.; Hersey, A.; Oprea, T.I.; et al. A comprehensive map of molecular drug targets. *Nat. Rev. Drug Discov.* **2017**, *16*, 19–34. [CrossRef]
31. Bemis, G.W.; Murcko, M.A. The properties of known drugs. 1. Molecular frameworks. *J. Med. Chem.* **1996**, *39*, 2887–2893. [CrossRef]
32. Todeschini, R.; Consonni, V. *Molecular Descriptors for Chemoinformatics*; Wiley-VCH: Weinheim, Germany, 2009.
33. Slamon, D.J.; Clark, G.M.; Wong, S.G.; Levin, W.J.; Ullrich, A.; McGuire, W.L. Human breast cancer: Correlation of relapse and survival with amplification of the HER-2/neu oncogene. *Science* **1987**, *235*, 177–182. [CrossRef]
34. Ferrara, N.; Hillan, K.J.; Gerber, H.P.; Novotny, W. Discovery and development of bevacizumab, an anti-VEGF antibody for treating cancer. *Nat. Rev. Drug Discov.* **2004**, *3*, 391–400. [CrossRef]
35. Barretina, J.; Caponigro, G.; Stransky, N.; Venkatesan, K.; Margolin, A.A.; Kim, S.; Wilson, C.J.; Lehár, J.; Kryukov, G.V.; Sonkin, D.; et al. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* **2012**, *483*, 603–607. [CrossRef] [PubMed]
36. Subramanian, A.; Tamayo, P.; Mootha, V.K.; Mukherjee, S.; Ebert, B.L.; Gillette, M.A.; Paulovich, A.; Pomeroy, S.L.; Golub, T.R.; Lander, E.S.; et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. USA* **2005**, *102*, 15545–15550. [CrossRef] [PubMed]
37. Hyman, D.M.; Taylor, B.S.; Baselga, J. Implementing Genome-Driven Oncology. *Cell* **2017**, *168*, 584–599. [CrossRef]
38. Garraway, L.A.; Verweij, J.; Ballman, K.V. Precision oncology: An overview. *J. Clin. Oncol.* **2013**, *31*, 1803–1805. [CrossRef] [PubMed]

39. Vogelstein, B.; Papadopoulos, N.; Velculescu, V.E.; Zhou, S.; Diaz, L.A., Jr.; Kinzler, K.W. Cancer genome landscapes. *Science* **2013**, *339*, 1546–1558. [CrossRef]
40. Guinney, J.; Dienstmann, R.; Wang, X.; de Reyniès, A.; Schlicker, A.; Soneson, C.; Marisa, L.; Roepman, P.; Nyamundanda, G.; Angelino, P.; et al. The consensus molecular subtypes of colorectal cancer. *Nat. Med.* **2015**, *21*, 1350–1356. [CrossRef] [PubMed]
41. Bian, Y.; Xie, X.Q. Generative chemistry: Drug discovery with deep learning generative models. *J. Mol. Model.* **2021**, *27*, 71. [CrossRef]
42. Khamessi, O.; Ben Mabrouk, H.; ElFessi-Magouri, R.; Kharrat, R. RK1, the first very short peptide from *Buthus occitanus tunetanus* inhibits tumor cell migration, proliferation and angiogenesis. *Biochem. Biophys. Res. Commun.* **2018**, *499*, 1–7. [CrossRef]
43. Khamessi, O.; Ben Mabrouk, H.; Hkimi, C.; Rtima, R.; Kamoun, S.; Kharrat, R.; Ghedira, K. DisintegrinDB: The first integrated database resource of disintegrins from snake venoms. *Biochem. Biophys. Res. Commun.* **2022**, *597*, 77–82. [CrossRef]
44. Khamessi, O.; Ben Mabrouk, H.; Othman, H.; ElFessi-Magouri, R.; De Waard, M.; Hafedh, M.; Marrakchi, N.; Srairi-Abid, N.; Kharrat, R. RK, the first scorpion peptide with dual disintegrin activity on $\alpha 1\beta 1$ and $\alpha v\beta 3$ integrins. *Int. J. Biol. Macromol.* **2018**, *120*, 1777–1788. [CrossRef]
45. Khamessi, O. *Method and System for Data Integration and Analysis in Bioinformatics (WO2019186344A1)*; World Intellectual Property Organization: Geneva, Switzerland, 2019. Available online: <https://patentscope.wipo.int/search/en/WO2019186344> (accessed on 24 December 2025).
46. Ghedira, K.; Khamessi, O.; Hkimi, C.; Kamoun, S.; Dhamer, N.; Daassi, K.; Ben Salah, W.; Othman, H.; Belhadj, W.; Ghorbal, Y. Design and implementation of a scalable high-performance computing (HPC) cluster for omics data analysis: Achievements, challenges and recommendations in LMICs. *GigaScience* **2024**, *13*, giae060. [CrossRef]
47. Brown, A.S.; Patel, C.J. A standard database for drug repositioning. *Sci. Data* **2017**, *4*, 170029. [CrossRef] [PubMed]
48. Zhavoronkov, A.; Ivanenkov, Y.A.; Aliper, A.; Veselov, M.S.; Aladinskiy, V.A.; Aladinskaya, A.V.; Terentiev, V.A.; Polykovskiy, D.A.; Kuznetsov, M.D.; Asadulaev, A.; et al. Deep learning enables rapid identification of potent DDR1 kinase inhibitors. *Nat. Biotechnol.* **2019**, *37*, 1038–1040. [CrossRef] [PubMed]
49. Stokes, J.M.; Yang, K.; Swanson, K.; Jin, W.; Cubillos-Ruiz, A.; Donghia, N.M.; MacNair, C.R.; French, S.; Carfrae, L.A.; Bloom-Ackermann, Z.; et al. A Deep Learning Approach to Antibiotic Discovery. *Cell* **2020**, *180*, 688–702.e13. [CrossRef]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.