Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Automated dataset integrating structural and bioactivity data for structure-based drug discovery

José Renato Duarte Fajardo
Laboratório Nacional de Computação Científica
Matheus Müller Pereira da Silva
Laboratório Nacional de Computação Científica
Leon Sulfierry Corrêa Costa
Laboratório Nacional de Computação Científica
Fabio Lima Custódio
Laboratório Nacional de Computação Científica
Isabella Alvim Guedes
isabella@posgrad.lncc.br

Laboratório Nacional de Computação Científica
Laurent Emmanuel Dardenne
Laboratório Nacional de Computação Científica

data-descriptor

Keywords:

Posted Date: August 4th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10449747/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Automated dataset integrating structural and bioactivity data for structure-based drug discovery

José Renato Duarte Fajardo[1,†], Matheus Müller Pereira da Silva[1], Leon Sulfierry Corrêa Costa[1], Fabio Lima Custódio[1], Isabella Alvim Guedes[1,†], Laurent Emmanuel Dardenne[1,†]

[1]Laboratório Nacional de Computação Científica (LNCC), Brazil

[†]Corresponding authors: joserdf@posgrad.lncc.br, isabella@posgrad.lncc.br, dardenne@lncc.br

Abstract

Machine learning for structure-based drug design needs large datasets linking protein-ligand structures to quantitative bioactivity values. DockTData is an openly licensed structural bioactivity dataset that links experimentally determined protein-ligand complexes from the Protein Data Bank (PDB) to bioactivity measurements from BindingDB and ChEMBL. The described snapshot comprises 24,719 protein-ligand complexes, 11,742 distinct ligands identified by InChIKey, 17,732 unique PDB entries, and 2,163 distinct protein targets, aggregated from 45,331 upstream bioactivity measurements. Measurements include K_d, K_i, IC_{50}, and EC_{50} values standardized to nanomolar units. Protein records are matched with an 85% sequence-identity threshold and ligand equivalence is established by exact InChIKey matching. Each bioactivity record stays linked to its three-dimensional PDB complex, enabling structure-based analyses at the atomic level. This small-molecule subset (24,378 complexes) contains 46% more entries than the small-molecule portion of the PDBbind version 2020 general set (16,744). Generated by a reproducible, automated pipeline and released under the CC-BY-SA 4.0 license, DockTData provides a transparent, reusable and continually expanding resource for training and evaluating machine-learning scoring functions, developing structure-based affinity-prediction models, and conducting structure-activity relationship analyses.

1 Background & Summary

Machine learning has advanced protein-ligand bioactivity prediction in recent years, but the problem remains challenging due to the limited availability of experimental data[1,2,3,4]. The relevant chemical space is vast and only sparsely sampled by experiment, so the quantitative measurements that models depend on are scarce, heterogeneous, and noisy. Binding is also shaped by solvent, protonation, conformational entropy, and receptor flexibility that cannot be fully captured by a single static structure. In this context, experimental structures and the associated bioactivity data provide the foundation for understanding structure-activity relationships and for developing and evaluating computational approaches, including molecular docking, virtual screening, and machine-learning-based scoring functions[5,6]. Recent co-folding and foundation models such as AlphaFold 3[7] and Boltz-2[8] now train bioactivity heads on ad-hoc unions of Protein Data Bank (PDB), BindingDB[9,10,11], ChEMBL[12,13], and PubChem records. This pattern

shows the need for a shared, openly accessible, versioned training corpus for bioactivity data. Yet bioactivity data remain fragmented across multiple specialized resources that differ in scope, data representation, curation standards, licensing terms, and update frequency.

The quality, coverage, and consistency of training data ultimately constrain what machine learning models can learn. Data accessibility and reusability are also important, affecting how easily results can be reproduced, evaluated, and extended[14]. Several resources link protein structures to ligand-binding measurements. Among them, PDBbind[15,16,17] is one of the most widely used collections of curated protein–ligand complexes paired with experimental bioactivity measurements. As of July 2026, v2020 was the latest PDBbind version available without a paid subscription[18]. Its v2020 general set contains 19,443 complexes, including 16,744 with small-molecule ligands and 2,699 with peptidic ligands. Beyond PDBbind, other databases also combine protein–ligand structures with binding annotations. BindingNet[19] provides a web-accessible database of protein–ligand interactions, particularly for less-studied targets. The final release of Binding MOAD[20,21], published in 2023, contains 41,409 high-quality crystal structures, including 15,223 complexes with quantitative bioactivity data. BioLiP2[22], by contrast, was updated weekly and provided functional annotations (EC numbers, GO terms, binding sites, catalytic sites) for PDB entries, with bioactivity data for some complexes sourced from Binding MOAD and BindingDB.

More recent resources have expanded the scope of structural datasets toward machine learning, generative modelling, and molecular dynamics. CrossDocked2020[23] contains approximately 22.5 million computationally generated ligand poses, whereas PLINDER[24] provides 449,383 protein-ligand systems organized into leakage-controlled data splits. MISATO[25] includes approximately 20,000 complexes with quantum-mechanically refined geometries and molecular-dynamics trajectories. LP-PDBbind[26] reorganizes PDBbind into training, validation, and test sets designed to reduce information leakage. HiQBind[18] integrates PDB structures with bioactivity information from BioLiP, Binding MOAD, and BindingDB through an automated workflow, producing 32,275 refined complexes derived from 18,160 unique PDB entries. SAIR[27] follows a different route, pairing large-scale co-folded conformations of ChEMBL and BindingDB records with their activity measurements, so its scale comes from in silico folding rather than experimental deposition.

ActivityFinder[28] takes a related automated approach, linking protein–ligand complex structures in the PDB to ChEMBL bioactivity records using protein sequence alignment and chemical structure matching. Its initial release covered 20,197 PDB structures linked to bioactivity data points from ChEMBL 35, representing 13,734 PDB ligands. The development illustrates the broader movement toward scalable and automated integration of structural and bioactivity data. Although ActivityFinder and DockTData address a similar data-integration problem, they adopt different source-coverage, matching, chemical-equivalence, and dissemination strategies.

DockTData was developed independently and contemporaneously with ActivityFinder and adopts four methodological and dissemination choices that define its distinct contribution. First, it integrates bioactivity records from both BindingDB[9,10,11] and ChEMBL[12,13], expanding source coverage and incorporating additional BindingDB-curated measurements, including patent-derived records that may not be represented in ChEMBL. Second, protein matching is performed with MMseqs2[29] using a single, predefined sequence-identity threshold of 85%, providing a transparent and reproducible criterion for associating bioactivity records with PDB structures. Third, ligand equivalence is established through exact InChIKey matching after local structure processing and InChIKey recalculation with RDKit[30]. This deliberately stringent criterion prioritizes unambiguous associations between experimental measurements and specific ligand, which is particularly important for scoring-function development, and other structure-based bioactivity analyses. The same chemical-identity framework is extended to peptidic ligands through CCD-supported

sequence-to-molecule conversion followed by InChIKey generation. Fourth, DockTData was designed as an openly reusable resource: the integrated dataset is distributed under CC BY-SA 4.0 with record-level source provenance retained, supporting transparent reconstruction, updating, reuse, and redistribution.

DockTData integration workflow runs end-to-end from raw source pulls, through sequence-identity matching with MMseqs2[29] and InChIKey-based chemical normalization[31,30], followed by explicit filtering and quality-control procedures. The resulting collection comprises 24,719 molecular complexes distributed across 17,732 PDB entries and 2,163 distinct protein targets. This small-molecule subset (24,378 complexes) contains 46% more entries than the small-molecule portion of the PDBbind version 2020 general set (16,744). The dataset also includes 271 complexes with 100 distinct peptide ligands, and 70 oligosaccharide complexes with 11 distinct oligosaccharide ligands. All matching, normalization, and filtering criteria are defined as configurable parameters, so the dataset can be reconstructed and updated from source releases.

Bioactivities in DockTData are reported as dissociation constants (K_d), inhibition constants (K_i), half-maximal inhibitory concentrations (IC_{50}), and half-maximal effective concentrations (EC_{50}), all normalized to nanomolar units. Each bioactivity record is linked to its corresponding three-dimensional PDB structure, enabling users to examine ligand-binding mode and molecular interactions at the atomic level.

Machine-learning approaches to protein-ligand bioactivity prediction, interaction prediction, and structure-based molecular design are strongly influenced by the scale, diversity, and curation quality of their training data. Structure-based machine-learning scoring functions, including RF-Score[32], Pafnucy[33], AK-Score[34], InteractionGraphNet[35], and GIGN[36], as well as more recent data-augmented training methods[37,4,38], have relied extensively on experimentally determined protein-ligand complexes from PDBbind. Sequence-based drug-target models such as DeepDTA[39] and TransformerCPI[40] have been developed and evaluated using large interaction matrices, including the Davis[41] and KIBA[42] datasets and ChEMBL-derived data. The demand for large and diverse protein-ligand datasets has grown with the emergence of co-folding architectures for interaction prediction. DockTData strengthens this evolving data ecosystem by providing a large, diverse and continuously expanding collection of experimentally determined PDB complexes systematically linked to bioactivity measurements from BindingDB and ChEMBL through a transparent and reproducible integration workflow.

Benchmark datasets are commonly used to assess scoring, ranking, docking, and virtual-screening performance[43,6,44]. However, systematic biases, structural artifacts, and similarities between training and test proteins or ligands have been identified in PDBbind-derived benchmarks[45,46], potentially resulting in overly optimistic estimates of model generalization. Broader resources with transparent construction criteria, such as DockTData, can facilitate the development of more representative and leakage-aware training and evaluation protocols. In addition, the integration of experimentally observed binding modes with bioactivity measurements across multiple ligands and protein targets enables large-scale structure-activity relationship analyses[47]. Thus, by expanding the structural and chemical space available for model development, DockTData emerges as a potential resource to support both the training and evaluation of machine-learning scoring functions and other structure-based bioactivity-prediction methods.

In this work, we present the construction, organization, and characterization of DockTData and its automated workflow for integrating structural and bioactivity records from the PDB, BindingDB, and ChEMBL databases. We detail the protein- and ligand-matching procedures, data normalization, filtering criteria, and quality-control steps, and characterize the resulting dataset in terms of its structural, chemical, target, and bioactivity coverage. The openly available dataset provides a transparent and extensible foundation for training and benchmarking machine-learning scoring functions, structure-based affinity-prediction models, and other computational

drug-discovery approaches. DockTData is freely available for the scientific community at `https://docktdata.lncc.br`.

2 Methods

The DockTData dataset was generated through an integrated ETL (Extract, Transform, Load) pipeline that combines structural data from the PDB with bioactivity measurements from BindingDB and ChEMBL. The Extract stage fetches raw data from PDB, BindingDB, and ChEMBL (see the Data Sources and Data Acquisition subsections below). The Transform stage runs InChIKey normalisation, sequence deduplication, and MMseqs2-based sequence integration (see the Data Cleaning, Normalisation and Provenance subsection). The Load stage merges the matched sequences with the bioactivity records, applies the assembly-level ligand disambiguation, and writes the versioned tabular export (see the Integration Criteria subsection).

This section documents the data sources, the acquisition procedure for each source, the cleaning and integration criteria applied, the peptide-ligand subset, and the structure preparation stage.

2.1 Data Sources

The dataset integrates data from three primary sources:

- **PDB.** The PDB provides experimentally solved three-dimensional structural data for protein-ligand complexes determined by X-ray diffraction, NMR spectroscopy, and cryo-electron microscopy[48,49,50]. The Structure Integration with Function, Taxonomy and Sequences resource (SIFTS) provides residue-level mappings between PDB chains and UniProt sequences, updated weekly alongside each PDB release. Through these mappings, SIFTS transfers functional annotations including Gene Ontology terms, Pfam and InterPro domain assignments, and enzyme classification codes to structural entries[51].

- **BindingDB.** BindingDB is a database of experimentally determined protein-ligand bioactivities, primarily focused on drug-like small molecules[9,10,11]. It reports distinct bioactivity data such as K_d, K_i, IC_{50}, and $EC_{50}$with associated protein sequence information and literature references. This work uses BindingDB release 202607 (July 2026 monthly snapshot in the `YYYYMM` release scheme), which contains approximately 3 million single-chain and 0.18 million multi-chain bioactivity records.

- **ChEMBL.** ChEMBL is a manually curated database of bioactive molecules with drug-like properties[12,13]. It provides bioactivity data extracted from scientific literature, along with target information and compound metadata. Release chembl_37 was used.

BindingDB and ChEMBL operate under Creative Commons licenses: ChEMBL is distributed under CC-BY-SA 3.0, while BindingDB is dual-licensed (CC-BY 3.0 for staff-curated records and CC-BY-SA 3.0 for ChEMBL-imported records). Redistribution of the integrated dataset under CC-BY-SA 4.0 is permitted via the adapter's-choice provision of CC-BY §5, ensuring license compatibility across the merged corpus[52,53].

2.2 Data Acquisition

Each source is fetched by an independent automated extractor. Every record is tagged with an upstream version identifier so that snapshots remain reproducible from the source releases.

2.2.1 PDB Extraction

The list of released entries for a given weekly cutoff is retrieved from the RCSB holdings REST endpoint. Per-entry metadata is then fetched from the RCSB Data GraphQL endpoint. The metadata query pulls the accession block (deposit, initial release, and revision dates), the entry information block (experimental method, structure determination methodology, and combined resolution), the assembly definition (polymer, branched, and non-polymer entity instances plus polymer instance validation scores), and the entity definitions (polymer canonical and authoritative one-letter sequences, non-polymer chemical component identifiers, branched entity references). Only entries not already present from a previous snapshot are refetched, so incremental updates track the weekly PDB release without redownloading the archive.

2.2.2 BindingDB Extraction

The complete monthly release is downloaded as a ZIP archive from the BindingDB download server, following the URL pattern `BindingDB_All_YYYYMM_tsv.zip`. The companion 3D SDF release (`BindingDB_All_3D_YYYYMM_sdf.zip`) is downloaded alongside. The extractor reads the ligand InChIKey, the BindingDB monomer identifier, the curation source, the four bioactivity columns in nM (K_i, IC_{50}, K_d, EC_{50}), the chain count, the source assay identifier, and, for every protein chain, the sequence and associated chain metadata. Sequences are uppercased. Single-chain and multi-chain records are kept separately, with multi-chain records preserving the full chain list. The SDF stream produces a monomer-to-InChIKey lookup computed with RDKit; dative bonds are converted to single bonds before InChI generation. Rows missing an InChIKey or all four bioactivity values are dropped.

2.2.3 ChEMBL Extraction

ChEMBL is distributed as a full SQLite database. The extractor downloads the release matching a pinned version string (`chembl_37` for this dataset) from the EBI ChEMBLdb download area. Bioactivity records, target sequences, and compound structures are read from the standard ChEMBL tables, with the same InChIKey, bioactivity type, bioactivity value, and unit fields captured as for BindingDB.

2.3 Data Cleaning, Normalisation and Provenance

The following operations are applied to ensure a consistent data format:

- Bioactivity cells are parsed with a regular expression that separates a relation prefix (<, >, ∼, or the implicit equality) from the numeric value. Cells whose numeric part is missing or malformed are dropped. Non-quantitative records are excluded at this stage. The relation prefix is preserved in a dedicated column so that downstream users can distinguish exact measurements from bounded or approximate ones.
- All bioactivity values are stored in nanomolar (nM) units. BindingDB already exposes its four bioactivity columns in nM, and ChEMBL values are converted to nM at read time.
- All ligands are processed with RDKit before InChIKey generation, ensuring that the same version of the InChI algorithm is used across all sources[30,31]. This absorbs canonicalisation differences between source databases and reduces the impact of documented bioactivity-database error rates[54]. The 27-character InChIKey serves as the ligand primary key throughout the pipeline.

- Protein sequences are retrieved directly from each source and deduplicated on the canonical amino-acid string before being passed to the alignment stage, so identical sequences from many source records collapse to a single alignment target.
- When the same triple (PDB entry, ligand InChIKey, target sequence) has bioactivity records from both BindingDB and ChEMBL, both values are retained. Each row in the final table carries an `bioactivity_source` column that identifies the origin database and an `bioactivity_id` column that points back to the original record, so users can compare, average, or select records per source.

The released TSV files include target and ligand metadata in addition to the primary identifiers and bioactivity measurements. The target fields (`target_uniprot`, `target_name`, and `target_organism`) are copied from the corresponding target entry in the bioactivity source, either BindingDB or ChEMBL. The MMseqs2 procedure described in the Integration Criteria subsection links each source target to the aligned PDB polymer entity. The ligand table, `docktdata_ligands.tsv`, contains the IUPAC name, molecular formula, molecular weight, canonical SMILES, and InChI for each ligand. The name, formula, and molecular weight are copied from the designated source record, whereas the canonical SMILES and InChI are recomputed with RDKit[30] alongside the InChIKey. This provides consistent chemical representations across data sources. The `bioactivity_id` column allows each entry to be traced back to its original source record.

2.4 Integration Criteria

The integration pipeline applies two primary matching criteria to establish links between structural and bioactivity data:

2.4.1 Protein Matching

Protein sequences from BindingDB and ChEMBL were matched to PDB canonical polymer sequences using MMseqs2[29] in the `easy-search` module with a minimum sequence identity of 85%. The exact command-line flags are reported in Supplementary Section S6 (Table S6).

The query set contains the PDB canonical polymer sequences from all released entries; the target set contains the union of BindingDB and ChEMBL target sequences. FASTA identifiers encode all bioactivity records that share the same sequence, so identical sequences from many source rows collapse to a single MMseqs2 target and are re-expanded after the join. Coverage filtering is disabled because the sequence-similarity join operates on already-curated canonical sequences from either side, and coverage cutoffs would eliminate legitimate short-domain matches.

The 85% threshold matches BindingDB's own choice for its native PDB-linking layer[9] and is consistent with the 80–100% range used by the ActivityFinder pipeline[28] and the 90% threshold used by SIFTS for its expanded UniProt-to-PDB mapping[51]. MMseqs2 was chosen as the alignment engine because the RCSB PDB itself adopted MMseqs2 as the underlying tool for its precomputed sequence-similarity clusters and its sequence search service[50,29], so DockTData's protein matching stays consistent with the sequence-space partitioning that PDB consumers are already exposed to. The chosen identity sits at the upper end of the 50–90% identity range commonly used for redundancy filtering in structural bioinformatics workflows. It is above the 40–50% range used by AlphaFold[55] for training-set decorrelation and the approximate 30% sequence-identity level below which homology detection becomes increasingly uncertain[56]. The rationale is that DockTData aims to group, rather than separate, experimentally equivalent receptor variants such as point mutants, truncations, and close orthologs.

2.4.2 Ligand Matching

Ligand chemical structures were matched using InChIKey identifiers, a canonical 27-character representation of molecular structure[31]. On the PDB side, ligand InChIKeys are resolved from the PDB Chemical Component Dictionary (CCD) distributed by wwPDB. Ligands not covered by the CCD are resolved from the entity-level structure and normalised through RDKit. Water molecules, common buffer ions, and other non-informative chemical components listed in a curated exclusion set are removed before matching.

2.4.3 Assembly-level Pairing

MMseqs2 hits are joined to ligand matches at the level of the PDB biological assembly. An assembly is retained only if it associates with exactly one small-molecule InChIKey, which keeps the receptor-ligand pairing unambiguous. Assemblies that carry more than one distinct co-crystallised ligand are excluded from the final table.

2.5 Peptide Handling

Polymer entities with a canonical sequence length of 40 residues or fewer and an RCSB polymer type of "Protein" are treated as peptide ligands. Their InChIKeys are generated by an ad hoc CCD-backed monomer library that parses three sequence formats: three-letter hyphen-separated codes, the RCSB one-letter format with parentheses around non-standard residues, and HELM-style bracketed monomers. Non-standard residues (for example MSE, HYP, and D-amino acids) are resolved against the PDB Chemical Component Dictionary and assembled into an RDKit Mol so that a canonical InChIKey can be computed[31,30]. Canonical amino-acid sequences take a fast path through RDKit's `Chem.MolFromSequence`, while sequences containing non-standard residues go through the CCD monomer builder.

InChIKey resolution for peptide ligands follows a three-step order. The first step uses the authoritative one-letter sequence (with parenthesised non-standard residues) fed through the CCD-backed builder. The second step falls back to the canonical sequence, which drops non-standard residue detail. The third step is used only when both sequence paths fail and the peptide is fully modelled in the structure (modelled residue count equal to sequence length); it computes the InChIKey from the mmCIF-derived entity structure. The sequence-first order was adopted because mmCIF-derived InChIKeys carry stereo, protonation, and three-dimensional artefacts that are absent from ChEMBL and BindingDB records, which reduces the match rate against those sources.

2.6 Structure Preparation

Each protein-ligand complex is processed through the Schrödinger Protein Preparation Workflow[57,58]. Preprocessing deletes and re-adds all hydrogens to ensure atom names and bond geometry are internally consistent, converts selenomethionine (MSE) residues to methionines, and creates covalent bonds that a structure may require but that PDB CONNECT records do not always carry: disulfide bridges between proximal sulfurs, N-linked and O-linked bonds to glycosylation sites, and palmitoylation bonds. Structural waters distant by more than 4 Å from any hetero group are removed at this stage rather than during the later cleanup.

The hydrogen-bond network is then optimised with water orientations sampled as free rotors, so buried waters can be flipped to their most consistent state. Residue ionisation states are assigned by PROPKA3[59]. The pH is read from the PDB header when the deposit carries an experimental pH annotation, so each complex is protonated at its own provided pH; when the

annotation is missing, PrepWizard's own default pH applies. Predicted pK_a values are labelled onto the corresponding residues in the output for downstream inspection. Ligand ionisation states are predicted with Ionizer to assign the most probable protonation state at the given pH. No geometry optimizations were made in the protein-ligand complex structures.

The full list of flags used is reported in Supplementary Section S6 (Table S7), and the per-assembly success rate for the released snapshot is reported in Supplementary Section S7 (Table S8). The prepared mmCIF and PDB files populate the archive described in the Data Records section. Assemblies for which PrepWizard failed ship only the raw mmCIF, so the presence of a `_prepwizard.pdb` file inside `docktdata_structures_v{VER}.zip` is the authoritative success marker. The integrated tabular data and the prepared structures are openly redistributed under CC-BY-SA 4.0.

2.7 Snapshot Versioning

Every complete pipeline run produces a versioned snapshot identified by the triple (PDB week, BindingDB month, ChEMBL release). PDB snapshots use the YYYYNN format, where YYYY is the ISO year and NN is the two-digit ISO week number, matching the RCSB weekly release cadence. BindingDB snapshots use the YYYYMM format matching the monthly download filename. ChEMBL snapshots record the upstream release string (for example chembl_37). Every published version of DockTData is therefore reproducible from the same three upstream release identifiers. DockTData releases use calendar versioning in the form `YYYY-0M-PATCH`, hyphen-separated in ISO-like fashion.

The snapshot described in this article is 2026.07, built from PDB week 202628, BindingDB 202607, and chembl_37.

3 Data Records

3.1 Repository Location

DockTData is openly available at `https://docktdata.lncc.br` under the CC-BY-SA 4.0 license, with terms compatible with those of the source databases, BindingDB and ChEMBL.

- **Repository:** Laboratório Nacional de Computação Científica (LNCC)
- **Dataset URL:** `https://docktdata.lncc.br`
- **License:** Creative Commons Attribution-ShareAlike 4.0 International (CC-BY-SA 4.0)

The dataset uses a calendar versioning scheme in the form `YYYY-0M-PATCH` (ISO-like, hyphen-separated). The version identifier is embedded in every artefact filename, so a specific snapshot can be reproduced from the release string alone. Each release is identified by its `snapshot_id`, described in the Snapshot Versioning subsection of the Methods.

3.2 File Structure

Each release contains four tab-separated tables, one archive of prepared structures, and two documentation files. Table 1 lists the contents.

The four TSV files use tab separators, LF line endings, UTF-8 encoding without BOM, and stable column order across releases. They are compatible with standard spreadsheet software and load directly with `pandas.read_csv(..., sep='\t')`, where the \t may need to

File	Description
docktdata_complexes.tsv	Fact table. One row per prepared complex (assembly × ligand). Primary artefact for structure-based workflows.
docktdata_ligands.tsv	Dimension table. One row per distinct ligand identified by InChIKey, with chemical descriptors recomputed by RDKit.
docktdata_targets.tsv	Dimension table. One row per distinct protein target keyed by UniProt accession, with the full canonical sequence.
docktdata_bioactivities.tsv	Fact table. One row per integrated upstream bioactivity measurement, linked back to complex_id and to the original source record.
docktdata_structures_v{VER}.zip	Archive of structural files, one per PDB biological assembly, with a manifest and per-file SHA-256 checksums.
README.md	Dataset quick-start, dictionary summary, load examples, citation guidance.
CHANGELOG.md	Version history and upstream-release provenance for each snapshot.

Table 1: File manifest of a DockTData release. Every filename embeds the release string YYYY-OM-PATCH.

be replaced with a literal tab character in some environments. Three foreign-key relationships cross-reference the four tables. The primary key docktdata_complexes.complex_id is referenced by docktdata_bioactivities.complex_id. The column docktdata_complexes.ligand_inchikey references docktdata_ligands.ligand_inchikey. The column docktdata_complexes.target_uniprot references docktdata_targets.target_uniprot.

3.3 Data Dictionary

Column-level definitions for each of the four TSV files (column name, data type, description, and example value) are provided in Supplementary Section S1 (Tables S1–S4). That section documents every column in docktdata_complexes.tsv, docktdata_ligands.tsv, docktdata_targets.tsv, and docktdata_bioactivities.tsv, along with the MMseqs2 identity and coverage columns used for auditing.

3.4 Prepared Structures Archive

The archive docktdata_structures_v{VER}.zip contains the structural files produced by the Structure Preparation stage (Section 2.6), one set per PDB biological assembly. Files are organised by PDB identifier and assembly number:

- {pdb_id}/{assembly_id}/{pdb_id}-assembly{assembly_id}.cif. The raw mmCIF file downloaded from the PDB.
- {pdb_id}/{assembly_id}/{pdb_id}-assembly{assembly_id}_prepwizard.cif. The mmCIF file processed by PrepWizard (hydrogens added, ionisation states assigned, hydrogen-bond network optimised).

- `{pdb_id}/{assembly_id}/{pdb_id}-assembly{assembly_id}\_prepwizard.pdb`. The same prepared structure in PDB format.
- `manifest.tsv`: one row per file with columns `pdb_id`, `assembly_id`, `file_role` (`raw_cif`, `prepared_cif`, or `prepared_pdb`), `path`, `sha256`, and `n_bytes`. Enables integrity checks without unzipping.

Assemblies for which PrepWizard failed ship only the raw mmCIF; the presence of a `_prepwizard.pdb` file is the authoritative success marker.

3.5 External Dataset Citations

The DockTData dataset integrates data from three external databases. The Protein Data Bank provides structural data[48,49]. BindingDB and ChEMBL provide bioactivity measurements[9,12]. All source databases should be cited in any publication that uses DockTData. Traceability to primary sources is preserved through the `bioactivity_id` column, so users can verify and cite original records.

4 Technical Validation

The technical quality of the DockTData dataset was checked through a multi-stage validation process built into the ETL pipeline. This section describes the quality assurance procedures, validation methods, and supporting analyses. All numbers are scoped to the snapshot 2026.07, which aggregates 124,836 integrated evidence rows from 45,331 distinct upstream bioactivity measurements.

4.1 Snapshot Composition

Table 2 summarises the headline composition of the snapshot. The dataset covers 17,732 unique PDB entries, 2,163 distinct protein targets, and 11,742 distinct ligands. Each upstream bioactivity measurement is propagated on average to 2.75 complexes via MMseqs2 sequence homology at $\geq$85% identity, and the 95.0% equality-relation rate confirms that almost every retained record is an exact experimental measurement rather than a bounded or approximate one.

Metric	**Value**
Integrated evidence rows	124,836
Unique complexes (assembly $\times$ ligand)	24,719
Unique PDB entries	17,732
Unique ligands (InChIKey)	11,742
Unique protein targets	2,163
Unique upstream bioactivity measurements	45,331
Small-molecule complexes / ligands	24,378 / 11,631
Peptide complexes / ligands	271 / 100
Oligosaccharide complexes / ligands	70 / 11

Table 2: Headline composition of the DockTData snapshot 2026.07.

The source-database split is dominated by ChEMBL (86.9% of rows) with BindingDB contributing the remaining 13.1%. This ratio reflects the relative sizes of the two upstream databases in the bioactivity types kept here.

4.2 Data Quality Assurance

The data quality assurance procedures address three key dimensions of the integrated dataset: protein sequence verification, ligand chemical validation, and bioactivity value integrity.

4.2.1 Sequence Identity Verification

The integration rule of at least 85% sequence identity between bioactivity database protein sequences and PDB canonical sequences was applied using MMseqs2 alignment [29]. The threshold was chosen to keep sequence variants and experimental truncations while cutting false positive matches. The alignment step ran sequence-to-sequence comparisons between millions of bioactivity database sequences and PDB canonical sequences. Each successful alignment was checked to confirm that the aligned sequence belonged to the same protein family as declared in the bioactivity source. This step blocks cross-family false positives that can appear when proteins are homologous but distinct.

Table 3 reports the identity-bucket distribution across the snapshot. Most rows (65.6%) fall in the 0.99–1.00 bucket, indicating an exact PDB-to-bioactivity-DB match, and 37.4% of rows have identity exactly equal to 1.0. The lower buckets (0.85–0.90 and 0.90–0.95) together account for $9.1\% + 10.9\% \approx 20\%$ of rows and correspond to point mutants, truncations, and close orthologs. This tail is the natural source of a target-holdout evaluation set with a controlled covariate shift. The finer-grained shape of the 0.85–0.99 tail is reported in Supplementary Section S5 (Figure S3).

Identity bucket	**% of rows**
0.99–1.00	65.6%
0.95–0.99	14.4%
0.90–0.95	9.1%
0.85–0.90	10.9%

Table 3: MMseqs2 sequence-identity distribution across integrated evidence rows.

4.2.2 Ligand Chemical Validation

Ligand matching between structural databases and bioactivity databases relies on InChIKey identifiers. The pipeline recomputes each InChIKey locally with RDKit [30] to keep the same algorithm version across records. Outdated identifiers in source databases, produced by older versions of the InChI algorithm, are therefore replaced by a consistent value. Molecular structures are read from MOLBlock records (ChEMBL) and SDF records (BindingDB), converted to InChIKey using the locally-installed RDKit version, and used to replace the source-provided identifier. This procedure covers known inconsistencies in older database records where different InChI versions produced different canonical representations of the same chemical structure.

4.2.3 Bioactivity Value Validation

All bioactivity values were checked for numerical consistency and unit standardization. The pipeline converts every value to nanomolar (nM) units, so records from different source databases can be compared directly. Values outside physiologically plausible ranges were flagged for review. This includes negative values, zero values, and values above millimolar concentrations, which usually indicate data entry errors. The step also checks that each bioactivity type label (K_d, K_i, IC_{50}, EC_{50}) matches an appropriate value range for that measurement type.

Table 4 reports the bioactivity-type mix. IC_{50} dominates the dataset (58.8%), followed by K_i (20.3%), K_d (11.3%), and EC_{50} (9.7%). This ordering is characteristic of literature-derived bioactivity databases and reflects the relative cost of the underlying assays: IC_{50} is the natural output of cell-based screens, whereas K_d and K_i require more expensive biophysical measurements. For scoring-function training, K_d and K_i generalise better on thermodynamic grounds, while IC_{50} and EC_{50} are more system-dependent and can be filtered by users when required.

Affinity type	**Rows**	**% of rows**
IC_{50}	73,344	58.8%
K_i	25,314	20.3%
K_d	14,041	11.3%
EC_{50}	12,137	9.7%

Table 4: Affinity-type mix in DockTData, normalized to nanomolar units.

4.2.4 Duplicate Detection and Handling

The pipeline detects duplicates at more than one level. Within each source database, records with identical protein-ligand combinations were merged, keeping the record with the most complete metadata. Cross-database duplicates, where the same protein-ligand pair appears in both BindingDB and ChEMBL, are kept as separate records with distinct `bioactivity_id` values so the origin of each measurement remains traceable. The pipeline also drops ambiguous cases where a PDB assembly is associated with more than one InChIKey, so each complex in the final dataset has a single protein-ligand association.

The affinities-per-complex distribution has a median of 2.0 and a mean of 5.1, so most complexes carry several independent measurements. This redundancy is useful for spotting inter-source inconsistencies and for computing measurement-averaged targets during model training. The per-complex breakdown of contributing bioactivity sources (BindingDB only, ChEMBL only, both) is reported in Supplementary Section S3 (Table S5).

4.3 Integration Validation

The integration validation procedures verify the correctness of cross-references between data sources and ensure consistent handling of edge cases.

4.3.1 Cross-reference Verification

The pipeline runs an alignment-based check that does not rely only on pre-existing mappings in bioactivity databases. For each candidate protein-ligand pair, MMseqs2[29] aligns the bioactivity database sequence to the PDB canonical sequence extracted directly from the structure file. This procedure contributes to the integrated dataset in two ways. The first contribution is resolution of isolated discrepancies against the pre-existing BindingDB mapping. Cases were observed where a BindingDB PDB claim referenced a protein with a different biological function, a different source organism, or a different protein family from the sequence recorded in the bioactivity entry. The independent alignment replaces the mapping in those cases.

The second contribution is recovery of protein-ligand pairs that BindingDB does not tag. Restricted to BindingDB's curated PDB ID(s) for Ligand-Target Complex field, DockTData maps 11,636 (row, PDB) pairs against BindingDB's 56,130. Of these, 5,900 pairs agree, and 5,736

pairs (49.3% of the DockTData set) are recovered by MMseqs2 alignment but were not tagged by BindingDB. Of the recovered pairs, 75.0% sit at $\geq$0.99 identity, so most are annotation-gap fills against near-identical target sequences in other PDB entries. The remaining 89.5% of BindingDB pair claims not present in DockTData are BindingDB entries where the target protein appears in a PDB structure but the ligand is not co-crystallised, so DockTData deliberately excludes those.

4.3.2 Missing Data Handling

Missing data are handled explicitly at several levels. Protein sequences that could not be aligned to any PDB structure at the 85% identity threshold are excluded from the final dataset, and the exclusion is written to the pipeline logs. Ligands without InChIKey representations in all three source databases are also excluded. Bioactivity records without quantitative values, or with ambiguous measurement types, are flagged and excluded from the primary dataset. The data dictionary lists every optional field that may contain missing values, so users can decide how to handle them for their analysis.

4.3.3 Outlier Detection in Affinity Values

A statistical analysis of the bioactivity value distributions is used to find potential outliers. The pipeline computes median, mean, standard deviation, and interquartile range for each bioactivity type (K_d, K_i, IC_{50}, EC_{50}). Values more than three standard deviations from the mean for their bioactivity type are flagged as potential outliers. Flagged values were reviewed manually to separate genuine biological outliers (for example, extremely weak binders with millimolar bioactivity values) from data entry errors. ChEMBL reports a small number of IC_{50} entries with nominal values as high as 10^{18} nM, which are "no activity above assay threshold" markers. These are already removed by keeping only the equality-relation subset described above. Most other flagged values were legitimate extreme cases, not errors. Users should still be careful when analyzing the tails of the bioactivity distribution.

4.4 Ingest Funnel

The pipeline ingests every released PDB entry and progressively narrows the set to complexes that carry a matched, quantitative bioactivity record. Table 5 reports the row count at each stage. Reduction from raw PDB to covered entries is 14.5×, corresponding to approximately 6.9% of ingested PDB entries that carry a ligand-target pair with an experimentally measured bioactivity in the sources. The step-by-step reporting supports auditing of every filter that shapes the released dataset.

Stage	**Rows**
PDB entries ingested	256,448
`complex_ligands` rows persisted	2,652,013
`complex_targets` rows persisted	787,582
Assemblies with bioactivity after MMseqs2	24,719

Table 5: Ingest funnel from the raw PDB entries to the final assemblies with a matched bioactivity record.

4.5 Distribution of Bioactivity Values

The bioactivity value distribution provides insight into the composition of the dataset and supports validation of data quality.

Figure 1 presents the distribution of bioactivity values across the full dataset, expressed as pK. The pK transform is defined as the negative decimal logarithm of the bioactivity in molar units. The distribution is centered between low and high nanomolar bioactivity (pK values approximately 6–9), with an extended tail toward weaker binders, consistent with the pharmacological focus of the source databases on active compounds.

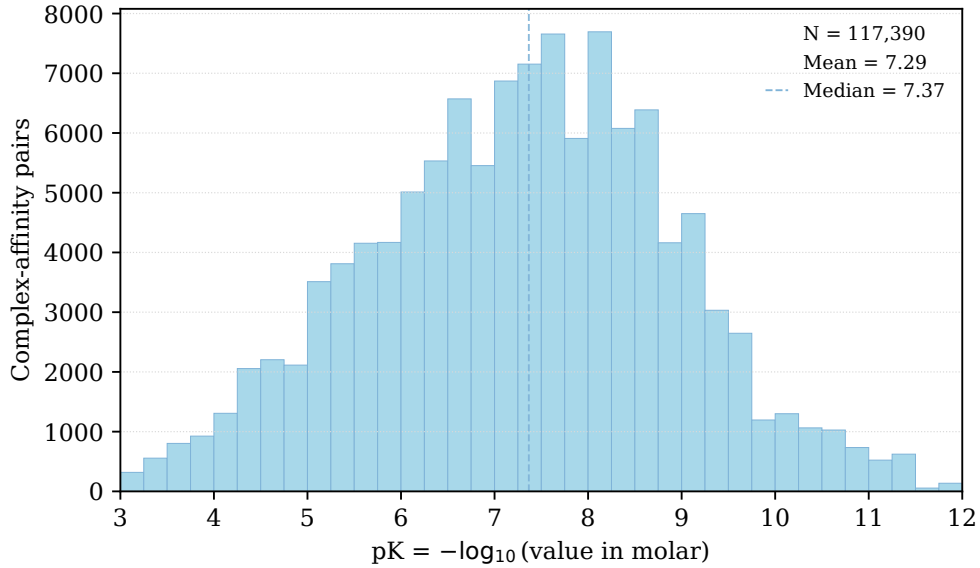


Figure 1: Distribution of bioactivity values in DockTData. Histogram of pK values ($-\log_{10}$ of K_d, K_i, IC_{50}, or EC_{50}in molar units) for all records with an equality relation, showing the range and central tendency of measurements.

Figure 2 shows the relative proportions of bioactivity measurement types (K_d, K_i, IC_{50}, EC_{50}) in the dataset. DockTData contains a broad mix of IC_{50}, K_i, K_d, and EC_{50} values (Table 4), in contrast to PDBBind[17], which predominantly contains K_d and K_i measurements. This diversity allows users to select subsets appropriate for their specific analytical requirements.

Figure 3 presents the temporal distribution of PDB structure deposition dates for complexes in the dataset. DockTData includes substantial representation of structures deposited after 2015, providing more recent coverage than annual-update datasets such as PDBBind v2020[17]. The cumulative growth curve of the covered PDB set through the same window is reported in Supplementary Section S4 (Figure S2).

Figure 4 presents the crystallographic resolution distribution for structures in the dataset. The X-ray subset has a median resolution of 2.05 Å and a 95th-percentile of 3.19 Å. The distribution is tighter than the overall RCSB distribution, which is expected because requiring a bound ligand biases the set towards actively refined active-site X-ray crystals rather than the result of a deliberate quality filter.

Table 6 summarises the experimental-method breakdown across covered PDB entries. X-ray diffraction dominates (93.3%), followed by cryo-electron microscopy (6.3%) and NMR (0.3%).

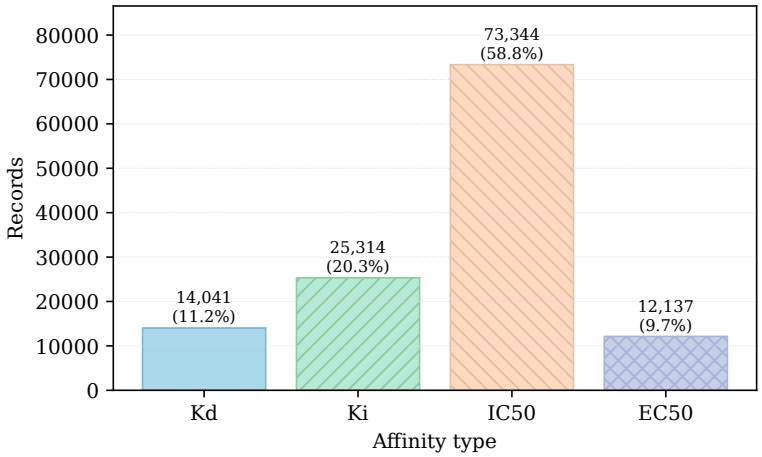


Figure 2: Distribution of bioactivity measurement types. Counts and proportions of IC_{50}, K_i, K_d, and EC_{50} records in DockTData.

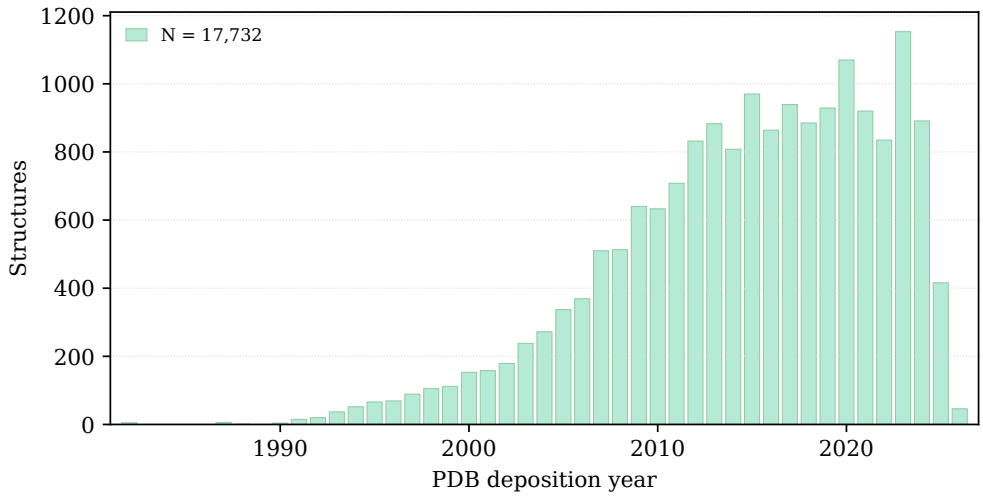


Figure 3: Temporal distribution of PDB structure deposition dates. Number of DockTData complexes by year of PDB deposition.

Experimental method	**% of PDB entries**
X-ray diffraction	93.3%
Cryo-electron microscopy	6.3%
NMR spectroscopy	0.3%

Table 6: Experimental-method breakdown for the 17,732 PDB entries covered by DockTData. Percentages are computed over unique PDB entries.

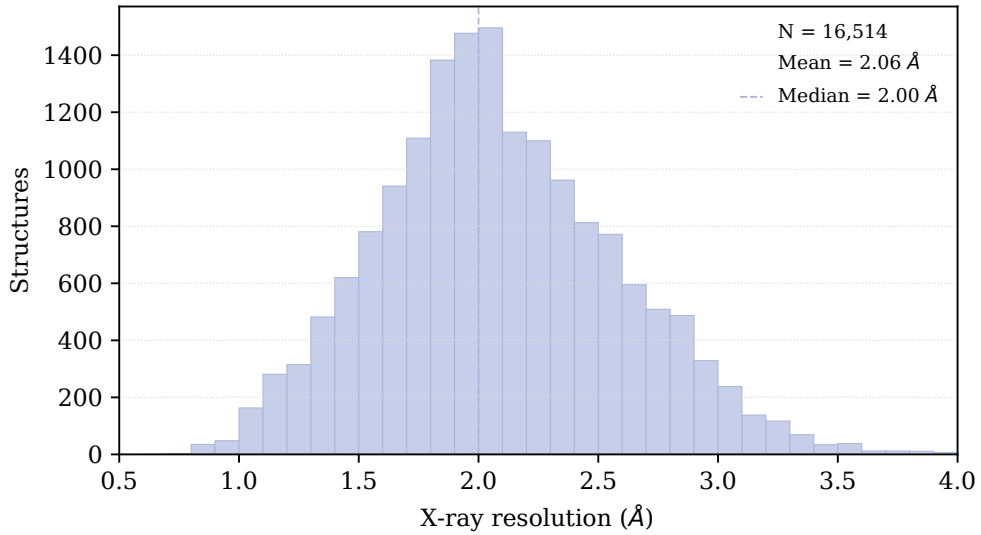


Figure 4: Crystallographic resolution distribution. Histogram of resolution values (Å) for X-ray crystallography structures in DockTData, with the median marked by the dashed line.

4.6 Peptide subset

Peptide ligands were identified from PDB polymer metadata by scanning peptide entities with a canonical sequence length of at most 40 residues and an RCSB polymer type of "Protein" (see Methods). Each peptide entity derives its chemical structure from a CCD-backed monomer library, and its InChIKey is generated from the sequence-ordered monomer list. This sequence-first InChIKey generation captures peptide ligands that standard small-molecule InChIKey pipelines would miss.

The released snapshot contains 271 distinct peptide complexes and 100 distinct peptide ligands. A small oligosaccharide branch adds 70 complexes and 11 ligands. Small-molecule complexes account for the remaining 24,378 entries of the 24,719 total, so peptides and oligosaccharides are minority but non-trivial fractions of the release.

Small molecules dominate the dataset by bioactivity pair count. The peptide subset adds coverage for peptidic ligands that standard InChIKey-only pipelines would miss. This separation by ligand class is a differentiator described in the Background section.

4.7 Sequence Coverage Statistics

The sequence coverage statistics report the diversity of protein targets in the dataset. The dataset spans several pharmacologically relevant protein families, including kinases, proteases, nuclear receptors, and metabolic enzymes. MMseqs2[29] produced alignments for protein sequences from both BindingDB and ChEMBL. The union of the sequence dictionaries from both sources covers the bioactivity data used here.

The 2,163 distinct protein targets range from 488-residue median length to a maximum of 7,096 residues. The long tail includes multidomain complexes exceeding 7,000 residues that were kept in full because DockTData stores the target sequence as unbounded text rather than truncating to a fixed cap.

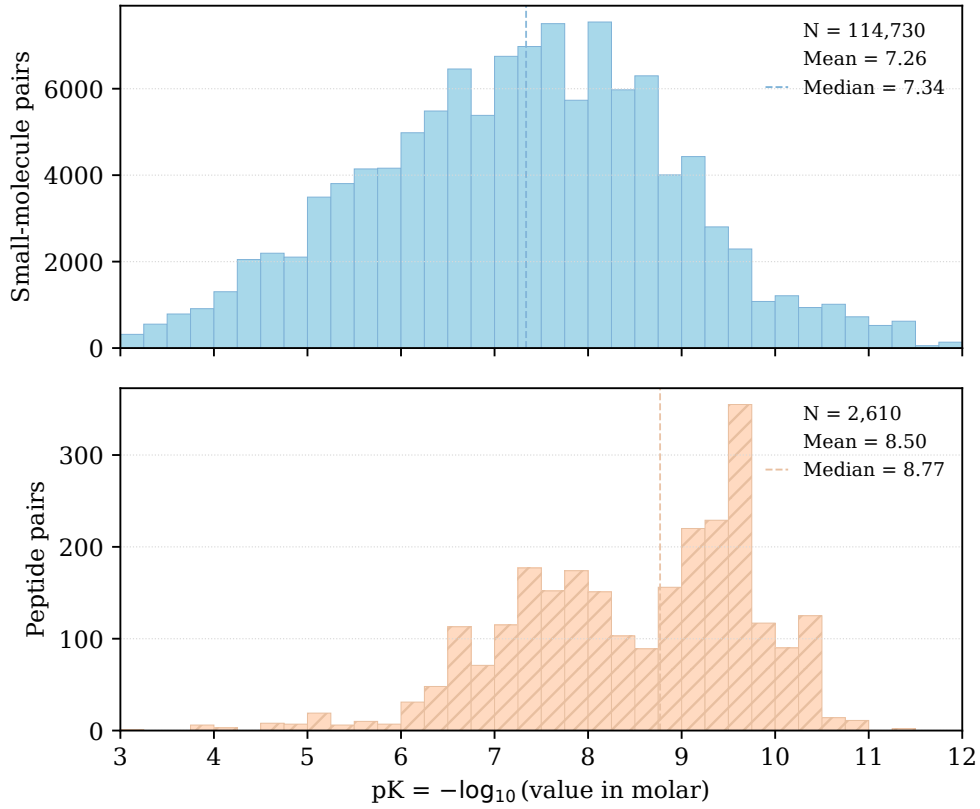


Figure 5: Bioactivity distribution split by ligand class. Top: small-molecule complex-bioactivity pairs. Bottom: peptide complex-bioactivity pairs, identified as polymer chains of at most 40 residues matched via CCD-derived InChIKeys.

The 85% sequence identity threshold keeps sequence variants and experimental truncations while preserving biological relevance. Most complexes have sequence identity at or near 100%, which corresponds to exact matches between bioactivity database sequences and PDB structures. A smaller share of complexes have sequence identity between 85% and 99%, and they cover variants and closely related proteins.

Each record traces back to its primary source through the `bioactivity_id` column, so sequence identity calculations and cross-reference checks can be reproduced.

4.8 Ligand Class Distribution

The ligand chemical diversity is described by the 11,742 distinct InChIKeys in the dataset. The InChIKey intersection filtering step removes compounds without a crystal structure in the PDB. Every ligand in the dataset therefore appears in at least one three-dimensional structure. This differs from datasets that also include predicted or homology-modeled structures, and it keeps the focus on experimentally determined structural data.

The ligand class distribution reflects the pharmacological focus of the source databases (BindingDB and ChEMBL), which concentrate on drug-like small molecules rather than peptides or oligonucleotides. This composition fits small-molecule drug discovery and machine learning model training, where chemical diversity and drug-like properties matter. The molecular-weight distribution of the small-molecule subset is reported in Supplementary Section S2 (Figure S1).

The RDKit-based normalization produces a consistent chemical representation across all ligands. It handles known issues with alternative naming conventions, protonation states, and structural representations that can create duplicate entries in less rigorously normalized datasets.

5 Data Availability

The DockTData dataset is publicly available at `https://docktdata.lncc.br`, hosted by the Laboratório Nacional de Computação Científica (LNCC). The release described in this article includes 24,719 receptor-ligand complexes with bioactivity measurements, 11,742 distinct ligands, per-record metadata, and three-dimensional structure files. Snapshots are calendar-versioned following the YYYY.0M.PATCH scheme, and every snapshot remains individually retrievable. The snapshot described here is 2026.07, built from PDB week 202628, BindingDB 202607, and chembl_37. The dataset is distributed under the Creative Commons Attribution-ShareAlike 4.0 International (CC-BY-SA 4.0) license. Users should cite the dataset using the citation format and version identifier provided on the portal landing page.

6 Code Availability

The ETL pipeline used to generate DockTData is not distributed as a code release. The pipeline is described in full in the Methods section, including data sources, integration criteria, peptide handling, and the Schrödinger PrepWizard structure preparation stage.

The generated artefacts, that is the integrated tabular index, per-record metadata, and the three-dimensional structure archive, are released through the DockTData portal at `https://docktdata.lncc.br` under the CC-BY-SA 4.0 license.

The pipeline source code is available from the corresponding author upon reasonable request.

References

[1] Jiang, J., Li, D., Wang, G. & Wei, G.-W. Recent advances in machine learning predictions of protein-ligand binding affinities. *Curr. Opin. Struct. Biol.* **96**, 103193 (2026).

[2] Harren, T., Gutermuth, T., Grebner, C., Hessler, G. & Rarey, M. Modern machine-learning for binding affinity estimation of protein-ligand complexes: Progress, opportunities, and challenges. *WIREs Comput. Mol. Sci.* **14**, e1716 (2024).

[3] Wei, J., Zhang, Y., Ramdhan, P. A., Huang, Z., Seabra, G., Jiang, Z., Li, C. & Li, Y. GatorAffinity: boosting protein-ligand binding affinity prediction with large-scale synthetic structural data. *bioRxiv* https://doi.org/10.1101/2025.09.29.679384 (2025).

[4] da Silva, M. M. P., Vidal, L., Guedes, I., de Magalhães, C., Custódio, F. & Dardenne, L. Data-centric training enables meaningful interaction learning in protein-ligand binding affinity prediction. *ChemRxiv* https://doi.org/10.26434/chemrxiv-2025-khbrl/v3 (2026).

[5] Sindt, F. & Rognan, D. Structure-based virtual screening of ultra-large chemical spaces: Advances and pitfalls. *Eur. J. Med. Chem.* **305**, 118576 (2026).

[6] Guedes, I. A., Pereira, F. S. S. & Dardenne, L. E. Empirical scoring functions for structure-based virtual screening: applications, critical aspects, and challenges. *Front. Pharmacol.* **9**, 1089 (2018).

[7] Abramson, J., Adler, J., Dunger, J., Evans, R., Green, T., Pritzel, A., Ronneberger, O., Willmore, L., Ballard, A. J., Bambrick, J., Bodenstein, S. W., Evans, D. A., Hung, C.-C., O'Neill, M., Reiman, D., Tunyasuvunakool, K., Wu, Z., Žemgulytė, A., Arvaniti, E., Beattie, C., Bertolli, O., Bridgland, A., Cherepanov, A., Congreve, M., Cowen-Rivers, A. I., Cowie, A., Figurnov, M., Fuchs, F. B., Gladman, H., Jain, R., Khan, Y. A., Low, C. M. R., Perlin, K., Potapenko, A., Savy, P., Singh, S., Stecula, A., Thillaisundaram, A., Tong, C., Yakneen, S., Zhong, E. D., Zielinski, M., Žídek, A., Bapst, V., Kohli, P., Jaderberg, M., Hassabis, D. & Jumper, J. M. Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* **630**, 493–500 (2024).

[8] Passaro, S., Corso, G., Wohlwend, J., Reveiz, M., Thaler, S., Somnath, V. R., Getz, N., Portnoi, T., Roy, J., Stark, H., Kwabi-Addo, D., Beaini, D., Jaakkola, T. & Barzilay, R. Boltz-2: towards accurate and efficient binding affinity prediction. *bioRxiv* https://doi.org/10.1101/2025.06.14.659707 (2025).

[9] Liu, T., Hwang, L., Burley, S. K., Nitsche, C. I., Southan, C., Walters, W. P. & Gilson, M. K. BindingDB in 2024: a FAIR knowledgebase of protein-small molecule binding data. *Nucleic Acids Res.* **53**, D1633–D1644 (2025).

[10] Gilson, M. K., Liu, T., Baitaluk, M., Nicola, G., Hwang, L. & Chong, J. BindingDB in 2015: a public database for medicinal chemistry, computational chemistry and systems pharmacology. *Nucleic Acids Res.* **44**, D1045–D1053 (2016).

[11] Liu, T., Lin, Y., Wen, X., Jorissen, R. N. & Gilson, M. K. BindingDB: a web-accessible database of experimentally determined protein–ligand binding affinities. *Nucleic Acids Res.* **35**, D198–D201 (2007).

[12] Zdrazil, B., Felix, E., Hunter, F., Manners, E. J., Blackshaw, J., Corbett, S., de Veij, M., Ioannidis, H., Lopez, D. M., Mosquera, J. F., Magariños, M. P., Bosc, N., Arcila, R.,

Kizilören, T., Gaulton, A., Bento, A. P., Adasme, M. F., Monecke, P., Landrum, G. A. & Leach, A. R. The ChEMBL Database in 2023: a drug discovery platform spanning multiple bioactivity data types and time periods. *Nucleic Acids Res.* **52**, D1180–D1192 (2024).

[13] Gaulton, A., Bellis, L. J., Chambers, J., Davies, M., Hersey, A., Light, Y., McGlinchey, S., Michalovich, D., Al-Lazikani, B. & Overington, J. P. The ChEMBL database: an open-access database for drug discovery. *Nucleic Acids Res.* **40**, D1100–D1107 (2012).

[14] Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., Gonzalez-Beltran, A., Gray, A. J. G., Groth, P., Goble, C., Grethe, J. S., Heringa, J., 't Hoen, P. A. C., Hooft, R., Kuhn, T., Kok, R., Kok, J., Lusher, S. J., Martone, M. E., Mons, A., Packer, A. L., Persson, B., Rocca-Serra, P., Roos, M., van Schaik, R., Sansone, S.-A., Schultes, E., Sengstag, T., Slater, T., Strawn, G., Swertz, M. A., Thompson, M., van der Lei, J., van Mulligen, E., Velterop, J., Waagmeester, A., Wittenburg, P., Wolstencroft, K., Zhao, J. & Mons, B. The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* **3**, 160018 (2016).

[15] Wang, R., Fang, X., Lu, Y. & Wang, S. The PDBbind database: collection of binding affinities for protein–ligand complexes with known three-dimensional structures. *J. Med. Chem.* **47**, 2977–2980 (2004).

[16] Wang, R., Fang, X., Lu, Y., Yang, C.-Y. & Wang, S. The PDBbind database: methodologies and updates. *J. Med. Chem.* **48**, 4111–4119 (2005).

[17] Liu, Z., Li, Y., Han, L., Li, J., Liu, J., Zhao, Z., Nie, W., Liu, Y. & Wang, R. PDB-wide collection of binding data: current status of the PDBbind database. *Bioinformatics* **31**, 405–412 (2015).

[18] Wang, Y., Sun, K., Li, J., Guan, X., Zhang, O., Bagni, D., Zhang, Y., Carlson, H. A. & Head-Gordon, T. A workflow to create a high-quality protein-ligand binding dataset for training, validation, and prediction tasks. *Digital Discovery* **4**, 1209–1220 (2025).

[19] Zhang, X., Gao, H., Wang, H., Chen, Z., Zhang, Z., Chen, X., Li, Y., Qi, Y. & Wang, R. BindingNet: a web-accessible database revealing the protein–ligand interactions for less-studied targets. *Front. Pharmacol.* **14**, 1112471 (2023).

[20] Hu, L., Benson, M. L., Smith, R. D., Lerner, M. G. & Carlson, H. A. Binding MOAD (Mother Of All Databases). *Proteins* **60**, 333–340 (2005).

[21] Wagle, S., Smith, R. D., Dominic, A. J., DasGupta, D., Tripathi, S. K. & Carlson, H. A. Sunsetting Binding MOAD with its last data update and the addition of 3D-ligand polypharmacology tools. *Sci. Rep.* **13**, 3008 (2023).

[22] Zhang, C., Zhang, X., Freddolino, P. L. & Zhang, Y. BioLiP2: an updated structure database for biologically relevant ligand-protein interactions. *Nucleic Acids Res.* **52**, D404–D412 (2024).

[23] Francoeur, P. G., Masuda, T., Sunseri, J., Jia, A., Iovanisci, R. B., Snyder, I. & Koes, D. R. Three-dimensional convolutional neural networks and a cross-docked data set for structure-based drug design. *J. Chem. Inf. Model.* **60**, 4200–4215 (2020).

[24] Durairaj, J., Adeshina, Y., Cao, Z., Zhang, X., Oleinikovas, V., Duignan, T., McClure, Z., Robin, X., Studer, G., Kovtun, D., Rossi, E., Zhou, G., Veccham, S., Isert, C., Peng, Y., Sundareson, P., Akdel, M., Corso, G., Stärk, H., Tauriello, G., Carpenter, Z., Bronstein, M., Kucukbenli, E., Schwede, T. & Naef, L. PLINDER: the protein-ligand interactions dataset and evaluation resource. *bioRxiv* https://doi.org/10.1101/2024.07.17.603955 (2024).

[25] Siebenmorgen, T., Menezes, F., Benassou, S., Merdivan, E., Didi, K., Mourão, A. S. D., Kitel, R., Liò, P., Kesselheim, S., Piraud, M., Theis, F. J., Sattler, M. & Popowicz, G. M. MISATO: machine learning dataset of protein-ligand complexes for structure-based drug discovery. *Nat. Comput. Sci.* **4**, 367–378 (2024).

[26] Li, J., Guan, X., Zhang, O., Sun, K., Wang, Y., Bagni, D. & Head-Gordon, T. Leak Proof PDBBind: a reorganized dataset of protein-ligand complexes for more generalizable binding affinity prediction. *J. Phys. Chem. B* **130**, 730–740 (2026).

[27] Lemos, P., Beckwith, Z., Bandi, S., Van Damme, M., Crivelli-Decker, J., Shields, B. J., Merth, T., Jha, P. K., De Mitri, N., Callahan, T. J., Nish, A., Abruzzo, P., Salomon-Ferrer, R. & Ganahl, M. SAIR: enabling deep learning for protein-ligand interactions with a synthetic structural dataset. *bioRxiv* `https://doi.org/10.1101/2025.06.17.660168` (2025).

[28] Ehmki, E. S. R., Gutermuth, T., Harren, T., Kurtz, S. & Rarey, M. ActivityFinder: a fully automated workflow for the integration of PDB structures and ChEMBL bioactivity data. *J. Chem. Inf. Model.* **66**, 1013–1034 (2026).

[29] Steinegger, M. & Söding, J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat. Biotechnol.* **35**, 1026–1028 (2017).

[30] RDKit Community & Landrum, G. A. RDKit: Open-source cheminformatics. *Zenodo* https://doi.org/10.5281/zenodo.21291217 (2024).

[31] Heller, S. R., McNaught, A., Pletnev, I., Stein, S. & Tchekhovskoi, D. InChI, the IUPAC International Chemical Identifier. *J. Cheminform.* **7**, 23 (2015).

[32] Ballester, P. J. & Mitchell, J. B. O. A machine learning approach to predicting protein-ligand binding affinity with applications to molecular docking. *Bioinformatics* **26**, 1169–1175 (2010).

[33] Stępniewska-Dziubińska, M. M., Zielenkiewicz, P. & Siedlecki, P. Development and evaluation of a deep learning model for protein–ligand binding affinity prediction. *Bioinformatics* **34**, 3666–3674 (2018).

[34] Kwon, Y., Shin, W.-H., Ko, J. & Lee, J. AK-Score: accurate protein-ligand binding affinity prediction using an ensemble of 3D-convolutional neural networks. *Int. J. Mol. Sci.* **21**, 8424 (2020).

[35] Jiang, D., Hsieh, C.-Y., Wu, Z., Kang, Y., Wang, J., Wang, E., Liao, B., Shen, C., Xu, L., Wu, J., Cao, D. & Hou, T. InteractionGraphNet: a novel and efficient deep graph representation learning framework for accurate protein-ligand interaction predictions. *J. Med. Chem.* **64**, 18209–18232 (2021).

[36] Yang, Z., Zhong, W., Lv, Q., Dong, T. & Yu-Chian Chen, C. Geometric interaction graph neural network for predicting protein-ligand binding affinities from 3D structures (GIGN). *J. Phys. Chem. Lett.* **14**, 2020–2033 (2023).

[37] Cao, D., Chen, G., Jiang, J., Yu, J., Zhang, R., Chen, M., Zhang, W., Chen, L., Zhong, F., Zhang, Y., Lu, C., Li, X., Luo, X., Zhang, S. & Zheng, M. Generic protein-ligand interaction scoring by integrating physical prior knowledge and data augmentation modelling. *Nat. Mach. Intell.* **6**, 688–700 (2024).

[38] da Silva, M. M. P., Guedes, I. A., Custódio, F. L., Krempser, E. & Dardenne, L. E. Deep learning strategies for enhanced molecular docking and virtual screening, in *Structure-Based Drug Design* (eds Marti, M. A., Turjanski, A. G. & Fernández Do Porto, D.) 177–221 (Springer International Publishing, 2024).

[39] Öztürk, H., Özgür, A. & Ozkirimli, E. DeepDTA: deep drug-target binding affinity prediction. *Bioinformatics* **34**, i821–i829 (2018).

[40] Chen, L., Tan, X., Wang, D., Zhong, F., Liu, X., Yang, T., Luo, X., Chen, K., Jiang, H. & Zheng, M. TransformerCPI: improving compound–protein interaction prediction by sequence-based deep learning with self-attention mechanism and label reversal experiments. *Bioinformatics* **36**, 4406–4414 (2020).

[41] Davis, M. I., Hunt, J. P., Herrgard, S., Ciceri, P., Wodicka, L. M., Pallares, G., Hocker, M., Treiber, D. K. & Zarrinkar, P. P. Comprehensive analysis of kinase inhibitor selectivity. *Nat. Biotechnol.* **29**, 1046–1051 (2011).

[42] Tang, J., Szwajda, A., Shakyawar, S., Xu, T., Hintsanen, P., Wennerberg, K. & Aittokallio, T. Making sense of large-scale kinase inhibitor bioactivity data sets: a comparative and integrative analysis. *J. Chem. Inf. Model.* **54**, 735–743 (2014).

[43] Su, M., Yang, Q., Du, Y., Feng, G., Liu, Z., Li, Y. & Wang, R. Comparative assessment of scoring functions: the CASF-2016 update. *J. Chem. Inf. Model.* **59**, 895–913 (2019).

[44] Buttenschoen, M., Morris, G. M. & Deane, C. M. PoseBusters: AI-based docking methods fail to generate physically valid poses or generalise to novel sequences. *Chem. Sci.* **15**, 3130–3139 (2024).

[45] Volkov, M., Turk, J.-A., Drizard, N., Martin, N., Hoffmann, B., Gaston-Mathé, Y. & Rognan, D. On the frustration to predict binding affinities from protein-ligand structures with deep neural networks. *J. Med. Chem.* **65**, 7946–7958 (2022).

[46] Graber, D., Stocker, R., Meißner, F., Sluka, T., Vetterli, M., Werlen, Y., Dey, A., Hilvo, M., Sander, O., Aebersold, D. M. & Schneider, G. Resolving data bias improves generalization in binding affinity prediction. *Nat. Mach. Intell.* **7**, 1713–1725 (2025).

[47] Guedes, I. A., Pereira da Silva, M. M., Galheigo, M., Krempser, E., de Magalhães, C. S., Barbosa, H. J. C. & Dardenne, L. E. DockThor-VS: a free platform for receptor-ligand virtual screening. *J. Mol. Biol.* **436**, 168548 (2024).

[48] Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. & Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **28**, 235–242 (2000).

[49] Rose, P. W., Prlić, A., Bi, C., Bluhm, W. F., Christie, C. H., Dutta, S., Green, R. K., Goodsell, D. S., Westbrook, J. D., Woo, J., Young, J., Zardecki, C., Berman, H. M., Bourne, P. E. & Burley, S. K. The RCSB Protein Data Bank: redesigned visualization and analysis tools. *Nucleic Acids Res.* **43**, D345–D356 (2015).

[50] Burley, S. K., Bhatt, R., Bhikadiya, C., Bi, C., Biester, A., Biswas, P., Bittrich, S., Blaumann, S., Brown, R., Chao, H., Chithari, V. R., Craig, P. A., Crichlow, G. V., Duarte, J. M., Dutta, S., Feng, Z., Flatt, J. W., Ghosh, S., Goodsell, D. S., Green, R. K., Guranovic, V., Henry, J., Hudson, B. P., Joy, M., Kaelber, J. T., Khokhriakov, I., Lai, J.-S., Lawson, C. L., Liang, Y., Myers-Turnbull, D., Peisach, E., Persikova, I., Piehl, D. W., Pingale, A., Rose, Y., Sagendorf, J., Sali, A., Segura, J., Sekharan, M., Shao, C., Smith, J., Trumbull, M., Vallat, B., Voigt, M., Webb, B., Whetstone, S., Wu-Wu, A., Xing, T., Young, J. Y., Zalevsky, A. & Zardecki, C. Updated resources for exploring experimentally-determined PDB structures and Computed Structure Models at the RCSB Protein Data Bank. *Nucleic Acids Res.* **53**, D564–D574 (2025).

[51] Dana, J. M., Gutmanas, A., Tyagi, N., Qi, G., O'Donovan, C., Martin, M. & Velankar, S. SIFTS: updated Structure Integration with Function, Taxonomy and Sequences resource allows 40-fold increase in coverage of structure-based annotations for proteins. *Nucleic Acids Res.* **47**, D482–D489 (2019).

[52] BindingDB. BindingDB – General Information and Licensing. `https://www.bindingdb.org/rwd/bind/info.jsp` (2026).

[53] ChEMBL. ChEMBL Interface Documentation – About and Licensing. `https://chembl.gitbook.io/chembl-interface-documentation/about` (2026).

[54] Tiikkainen, P., Bellis, L., Light, Y. & Franke, L. Estimating error rates in bioactivity databases. *J. Chem. Inf. Model.* **53**, 2499–2505 (2013).

[55] Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., Back, T., Petersen, S., Reiman, D., Clancy, E., Zielinski, M., Steinegger, M., Pacholska, M., Berghammer, T., Bodenstein, S., Silver, D., Vinyals, O., Senior, A. W., Kavukcuoglu, K., Kohli, P. & Hassabis, D. Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).

[56] Rost, B. Twilight zone of protein sequence alignments. *Protein Eng.* **12**, 85–94 (1999).

[57] Sastry, G. M., Adzhigirey, M., Day, T., Annabhimoju, R. & Sherman, W. Protein and ligand preparation: parameters, protocols, and influence on virtual screening enrichments. *J. Comput. Aid. Mol. Des.* **27**, 221–234 (2013).

[58] Schrödinger, LLC. Schrödinger Release 2026-2: Protein Preparation Workflow. (Schrödinger, LLC, New York, NY, 2026).

[59] Olsson, M. H. M., Søndergaard, C. R., Rostkowski, M. & Jensen, J. H. PROPKA3: consistent treatment of internal and surface residues in empirical pK_a predictions. *J. Chem. Theory Comput.* **7**, 525–537 (2011).

[60] O'Boyle, N. M., Banck, M., James, C. A., Morley, C., Vandermeersch, T. & Hutchison, G. R. Open Babel: an open chemical toolbox. *J. Cheminform.* **3**, 33 (2011).

7 Author Contributions

J. R. Fajardo: Conceptualization, Methodology, Software, Validation, Writing - Original Draft, Writing - Review & Editing, Data Curation. M. M. Pereira da Silva: Conceptualization, Methodology, Writing - Review & Editing. L. S. C. Costa: Conceptualization, Methodology, Writing - Review & Editing. F. Lima Custódio: Methodology, Writing - Review & Editing. I. Alvim Guedes: Conceptualization, Methodology, Writing - Review & Editing, Supervision, Project Administration. L. E. Dardenne: Conceptualization, Writing - Review & Editing, Supervision, Project Administration.

8 Competing Interests

The authors declare no competing interests.

9 Funding

This work was supported by the Brazilian Federal Agency for Support and Evaluation of Graduate Education (CAPES); and the Brazilian agencies Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), grant number 309744/2022-9; and the Fundação Carlos Chagas Filho de Apoio à Ciência (FAPERJ), grant numbers E-26/010.001415/2019, E-26/211.357/2021, E-26/200.608/2022, E-26/210.372/2022 (SEI-260003/001599/2022), E-26/200.393/2023 and E-26/204.413/2025 (SEI-260003/020539/2025).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- supplementaryinformationscientificdata.pdf