Supplementary Information for 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`

Every analysis in this Supplementary Information is scoped to the same snapshot 2026.07 used in the main manuscript.

List of Tables

List of Figures

S1 Data Dictionary

Full column-level definitions for the four tab-separated tables released with every DockTData snapshot. Every table carries the release `snapshot_id` as a header comment for provenance. Foreign-key relationships across the four tables are documented in the Data Records section of the main manuscript.

S1.1 docktdata_complexes.tsv

Fact table with one row per prepared complex (assembly × ligand).

Column	Description	Type	Example
`complex_id`	Composite primary key `{pdb_id}_{assembly_id}_{ligand_inchikey[:14]}`	String	`1ATP_1_ZKHQWZAMYRWXGA`
`pdb_id`	Four-character PDB identifier	String	`1ATP`
`assembly_id`	PDB biological assembly identifier	Integer	`1`
`ligand_inchikey`	27-character InChIKey of the ligand	String	`ZKHQWZAMYRWXGA-KQYNXXCUSA-N`
`target_uniprot`	UniProt accession of the protein target; empty for the small fraction of complexes whose source records did not resolve to a UniProt accession	String	`P12345`
`ligand_class`	Ligand class assigned by the pipeline	String	`small_molecule, peptide, oligosaccharide`
`experimental_method`	Method used to determine the crystal structure	String	`X-ray diffraction`
`resolution_A`	Crystallographic resolution in Å; empty for NMR	Float	`1.85`
`deposition_date`	PDB deposition date (ISO 8601)	Date	`2015-06-12`
`n_bioactivities`	Number of matching rows in `docktdata_bioactivities.tsv`	Integer	`3`
`source_coverage`	Which upstream bioactivity sources back this complex; one of `bindingdb_only`, `chembl_only`, `both`. Aggregating on this column reproduces the 3-row overlap in Supplementary Section S3 (Table S5)	String	`both`
`snapshot_id`	Snapshot version identifier	String	`2026.07`

Table S1: `docktdata_complexes.tsv` column dictionary. Each row is a unique prepared complex.

S1.2 docktdata_ligands.tsv

Dimension table with one row per distinct ligand identified by InChIKey.

Column	Description	Type	Example
`ligand_inchikey`	Primary key. 27-character InChIKey	String	`ZKHQWZAMYRWXGA-KQYNXXCUSA-N`
`preferred_name`	Ligand name from the preferred source (BindingDB when available, else the ChEMBL molecule identifier). Values longer than 512 characters are truncated	String	Palbociclib
`het_code`	Three-character PDB Chemical Component Dictionary code, rolled up per InChIKey by most-common assignment across complexes; empty when no complex records a CCD	String	`EAM`
`ligand_class`	Ligand class assigned by the pipeline	String	`small_molecule`
`formula`	Molecular formula recorded upstream	String	`C10 H16 N5 O13 P3`
`exact_mass`	Monoisotopic mass in Daltons	Float	`507.181`
`mol_weight`	RDKit-computed average molecular weight in Daltons	Float	`507.184`
`smiles_canonical`	RDKit-computed canonical SMILES; 2D topology only, conformers stripped	String	CCNC(=O)CC1…
`inchi`	RDKit-computed InChI string (the parent of `ligand_inchikey`)	String	InChI=1S/C22H22ClN5O2/c1…
`n_complexes`	Number of distinct complexes containing this ligand in the snapshot	Integer	`42`

Table S2: `docktdata_ligands.tsv` column dictionary. Each row is a unique ligand identified by InChIKey.

S1.3 docktdata_targets.tsv

Dimension table with one row per distinct protein target keyed by UniProt accession.

Column	Description	Type	Example
`target_uniprot`	UniProt accession from the bioactivity source; empty for the small fraction of targets covered only by sources that do not carry UniProt in the current release	String	`P12345`
`target_name`	Protein target name from the bioactivity source; empty when `target_uniprot` is empty for the same target	String	Urease subunit alpha
`organism`	Organism source of the protein target	String	Homo sapiens
`sequence`	Full canonical amino-acid sequence	Text	`MVK...`
`sequence_length`	Length of the canonical sequence in residues	Integer	`298`
`n_complexes`	Number of complexes for this target in the snapshot	Integer	`128`
`n_ligands`	Number of distinct ligands bound to this target in the snapshot	Integer	`87`

Table S3: `docktdata_targets.tsv` column dictionary. Each row is a unique protein target keyed by UniProt accession.

S1.4 docktdata_bioactivities.tsv

Fact table with one row per integrated evidence measurement: each row links one upstream source measurement to one complex in this snapshot.

Column	Description	Type	Example
`complex_id`	Foreign key to `docktdata_complexes.complex_id`	String	`1ATP_1_ZKHQWZAMYRWXGA`
`bioactivity_type`	Type of bioactivity measurement	String	K_d, K_i, IC_{50}, EC_{50}
`bioactivity_value_nm`	Numeric bioactivity value, always in nanomolar	Float	`150.0`
`bioactivity_relation`	Relation prefix from the source record	String	`=`, $<$, $>$, $\sim$
`bioactivity_source`	Primary source database of the bioactivity measurement	String	`bindingdb, chembl`
`bioactivity_id`	Original record identifier from the source database	String	`BDBM50000123 or CHEMBL1234567`
`mmseqs_identity`	MMseqs2 sequence identity between source target and PDB canonical sequence	Float	`1.00`
`mmseqs_coverage`	MMseqs2 alignment coverage on the source-target side (`alnlen` / target sequence length), clipped to $[0, 1]$; empty when the bioactivity row has no resolved target	Float	`0.97`
`pk_value`	Derived $-\log_{10}$(`bioactivity_value_nm` $\times 10^{-9}$); empty when the source value is missing or non-positive	Float	`6.82`

Table S4: `docktdata_bioactivities.tsv` column dictionary. Each row is one integrated evidence measurement (upstream source record propagated to one complex).

The `mmseqs_identity` column reflects the integration criterion of $\geq$85% identity described in the Methods section of the main manuscript. When the same triple (PDB entry, ligand InChIKey, target sequence) has bioactivity records from both BindingDB and ChEMBL, both records are retained as separate rows with distinct `bioactivity_id` values, so users can compare, average, or select records per source.

S2 Ligand Molecular Weight Distribution

Figure S1 shows the molecular weight distribution of the 11,742 distinct ligands in the snapshot, split by `ligand_class` (small-molecule, peptide, oligosaccharide). Molecular weights were computed by RDKit from each ligand's resolved SMILES and stored in the `ligands.mol_weight` column of the database. The histogram overlay uses semi-transparent fills so overlapping regions remain visible; per-class medians are marked with dashed vertical rules.

The legend reports sample size (N), mean, and median for each class: small-molecule ligands (N = 11,631) cluster around drug-like molecular weights, peptide ligands extend to higher masses, and the 11 oligosaccharide entries fall in the intermediate range.

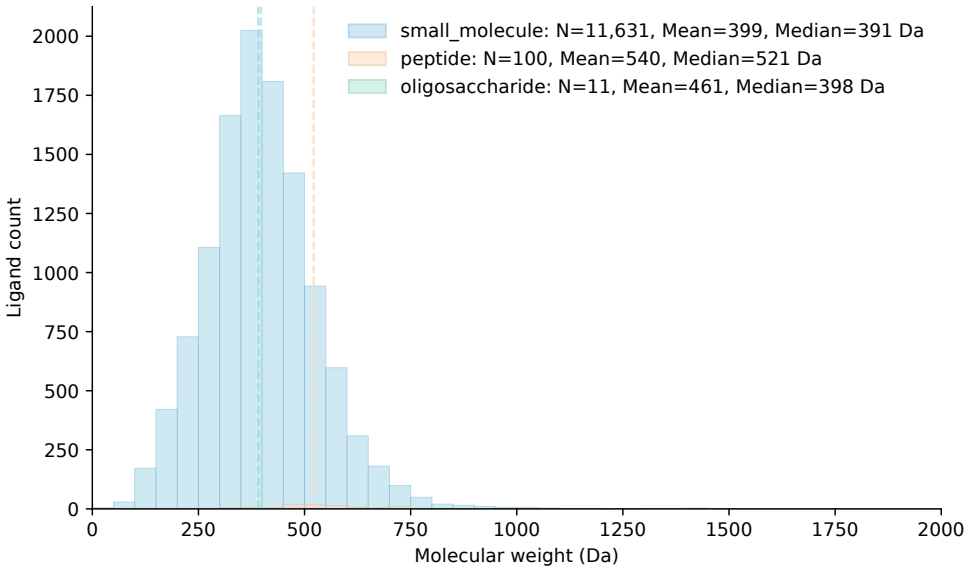

Figure S1: Molecular weight distribution of DockTData ligands by class. Small-molecule ligands cluster around drug-like values; peptide ligands extend to higher masses. Histogram bins span 0–2000 Da; masses above 2000 Da are clamped into the final bin. Per-class medians are marked on the figure.

This distribution helps users decide whether the ligand pool matches their target chemical domain (drug-like, macrocyclic, glycan) before subsetting.

S3 Per-Complex Source Overlap

Table S5 reports the split of the 24,719 complexes by which bioactivity sources provide at least one measurement. BindingDB-only and ChEMBL-only counts show how much each source contributes independently; the intersection quantifies the redundancy that supports inter-source consistency checks.

The breakdown is computed directly from the `source_coverage` column of `docktdata_complexes.tsv` in the released snapshot.

Source coverage	**Complexes**	**% of total**
BindingDB only	2,593	10.49%
ChEMBL only	9,350	37.83%
Both sources	12,776	51.68%
Total	24,719	100.00%

Table S5: Per-complex breakdown by contributing bioactivity source. Complexes with both BindingDB and ChEMBL measurements support cross-source consistency analyses.

The intersection is the auditable substrate for consistency checks between BindingDB and ChEMBL for the same (target, ligand) pair. It also characterises the value of the integration beyond the row-count sum reported in the main text.

S4 Cumulative Temporal Coverage

Figure S2 presents the cumulative distribution of the 24,719 complexes by PDB deposition year. This companion view of the per-year histogram in the main text (Figure 3 in the Technical Validation section) shows when the dataset reaches given coverage fractions, which is the natural reading for time-based holdout splits. The step plot includes horizontal percentile marks at 25%, 50%, 75%, and 90% of the cumulative total; vertical drop lines from the curve to each intersection indicate the corresponding deposition year threshold.

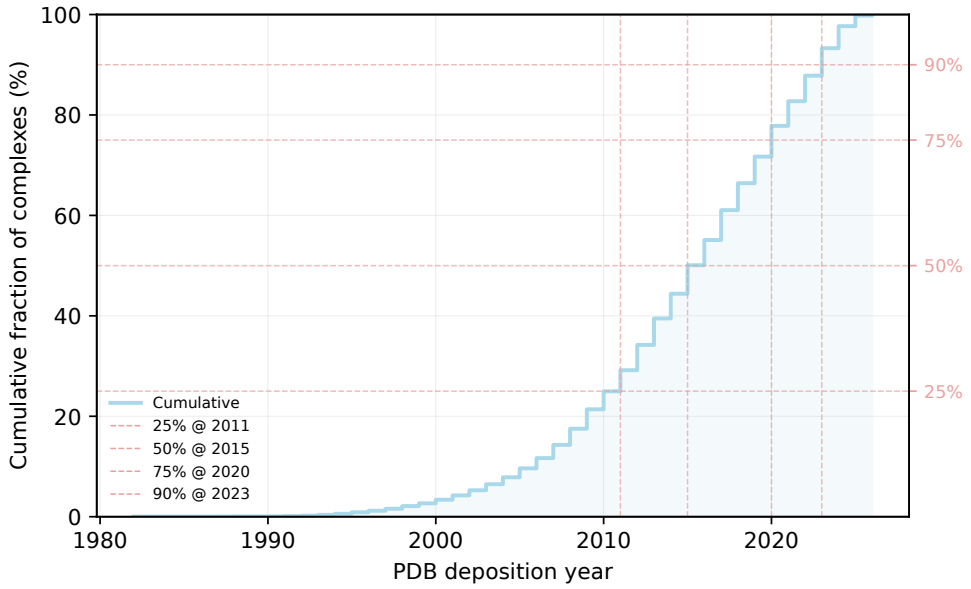

Figure S2: Cumulative fraction of DockTData complexes by PDB deposition year. Percentile marks (25%, 50%, 75%, 90%) support the design of time-based holdout splits: reading the year at the target coverage gives the split threshold.

Users constructing temporal benchmarks (train on deposits before year T, test on later deposits) can read the year for any target coverage fraction directly off this curve.

S5 Identity Tail 0.85 to 0.99

The MMseqs2 identity buckets summarised in the main text (Table 3 in the Technical Validation section) are dominated by the 0.99–1.00 bucket (65.6% of rows). This section zooms into the 0.85–0.99 tail, which is the natural source of a target-holdout evaluation set with a controlled covariate shift.

Figure S3 shows the fine-grained identity histogram in bins of 0.01 across the 0.85–0.99 range. The legend reports the total number of alignments in the tail (sum of bin counts). Bin frequencies drop sharply as identity increases, confirming that most integrated alignments either meet the inclusion threshold by a narrow margin or match at near-perfect identity, with relatively few sequences in the intermediate 0.90–0.98 bands.

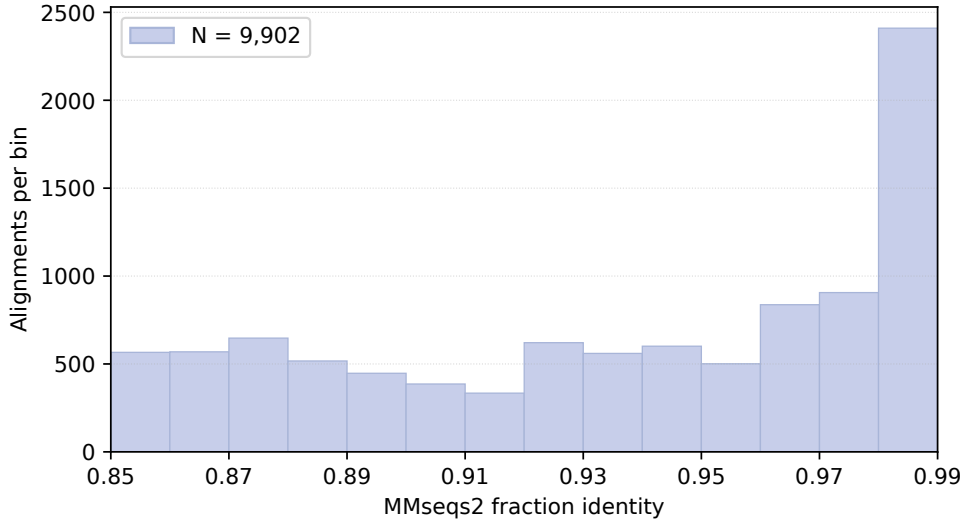

Figure S3: Fine-grained MMseqs2 sequence-identity distribution in the 0.85–0.99 tail (bin width 0.01, alignment counts on the y-axis). This range covers point mutants, truncations, and close orthologs, and is the natural source of held-out target splits.

Users can select a lower identity floor for stricter target holdout, or a tighter upper bound for near-identical targets, by filtering `docktdata_affinities.mmseqs_identity` against the reported bin boundaries.

S6 Pipeline Parameters

Exact command-line and API parameters used by the two external tools that shape the released dataset: MMseqs2 for sequence-based integration and the Schrödinger Protein Preparation Workflow (PrepWizard) for structure preparation. Reported verbatim from the pipeline source so a reader can reproduce a snapshot without inspecting the code.

S6.1 MMseqs2 (integration)

Table S6 lists the flags passed to `mmseqs easy-search`. The invocation is:

```
mmseqs easy-search <query.fasta> <target.fasta> <output.tsv> <tmp_dir> \
    --min-seq-id 0.85 \
    -c 0.0 --cov-mode 0 \
    -a 1 --alignment-mode 3 \
    --prefilter-mode 2 --rescore-mode 2 \
    --max-seqs 2147483647 \
    --remove-tmp-files 1
```

Flag	Rationale
`-min-seq-id 0.85`	Integration threshold declared in the Methods section (85%). Matches BindingDB's own PDB-linking layer.
`-c 0.0 -cov-mode 0`	Coverage filtering disabled. Query and target come from already-curated canonical sequences, so short-domain matches must not be pruned.
`-a 1 -alignment-mode 3`	Full alignment result emitted, including sequence identity and alignment length; alignment mode 3 aligns the full query against the full target.
`-prefilter-mode 2` `-rescore-mode 2`	Ungapped prefilter with the standard scoring matrix; ungapped rescore. Chosen because inputs are already high-identity within a family.
`-max-seqs 2147483647`	No cap on hits per query. The join is one-to-many by construction and every homologous target must be kept.
`-remove-tmp-files 1`	Deletes MMseqs2 scratch after the run so the pipeline's tmp footprint stays bounded.

Table S6: MMseqs2 `easy-search` parameters used by the integration stage. GPU acceleration is enabled by default when a supported device is available; no other flags are set.

S6.2 Schrödinger PrepWizard (structure preparation)

Table S7 lists the flags passed to `$SCHRODINGER/utilities/prepwizard`. Full invocation:

```
$SCHRODINGER/utilities/prepwizard <input> <output> \
    -WAIT -preserve_st_titles \
    -rehtreat \
    -disulfides -glycosylation -palmitoylation \
    -mse \
    -preprocess_watdist 4 \
    -samplewater \
    -use_PDB_pH \
    -label_pkas \
    -noepik \
    -noimpref
```

The trailing refinement-skip flag is emitted as `-noimpref` on Schrödinger releases up to 2025-3 and as `-norefine` on 2026-2 and later. The pipeline probes the binary's help text once per path and picks the accepted spelling; both encode the same behaviour.

Flag	Rationale
`-WAIT`	Blocks until the job finishes so per-complex outcomes can be captured directly, without the Schrödinger job daemon.
`-preserve_st_titles`	Keeps the input structure titles on the output, so downstream tooling can join back on the original identifier.
`-rehtreat`	Deletes and re-adds all hydrogens so atom names and bond geometry are internally consistent regardless of the input's hydrogen state.
`-disulfides`	Rebuilds disulfide bridges that PDB CONNECT records do not always carry.
`-glycosylation`	Adds N-linked and O-linked bonds to glycosylation sites.
`-palmitoylation`	Adds palmitoylation bonds when applicable.
`-mse`	Converts selenomethionine (MSE) residues to methionines.
`-preprocess_watdist 4`	Removes crystallographic waters more than 4 Å from any hetero group before hydrogen-bond optimisation.
`-samplewater`	Samples water orientations as free rotors so buried waters can be flipped during the optimisation stage.
`-use_PDB_pH`	Reads the pH from the PDB header (`r_pdb_PDB_EXPDTA_PH`) and runs PROPKA3 at that pH; when the annotation is missing, PrepWizard's built-in default applies.
`-label_pkas`	Labels predicted PROPKA pK_a values onto the corresponding residues in the output.
`-noepik`	Disables Epik so ligand ionisation states are not re-enumerated and the run does not require the Epik module licence.
`-noimpref (or -norefine)`	Skips the restrained impref minimisation, so the prepared structure preserves the crystallographic heavy-atom positions. The token is chosen at invocation time by probing the binary's help output; `-norefine` is the current name on Schrödinger 2026-2 onwards.

Table S7: PrepWizard flags used by the Structure Preparation stage. Every other flag stays at its Schrödinger default.