

# DockTData: Automated Integration of Binding Affinity and Molecular Structure Data for Receptor–Ligand Interaction Modeling

José Renato D Fajardo,<sup>1\*</sup> Matheus M P da Silva,<sup>1</sup> Leon S C Costa,<sup>1</sup> Isabella A Guedes,<sup>1</sup> Laurent E Dardenne<sup>1\*</sup>

<sup>1</sup> National Laboratory of Scientific Computation  
joserdf@posgrad.lncc.br; dardenne@lncc.br



## Background

### Receptor–ligand affinity prediction

Predicting receptor–ligand affinity is central to rational drug discovery and molecular modeling, guiding compound design and reducing experimental costs.

### Challenges in data integration

Integrating information from multiple databases is hindered by structural inconsistencies, lack of standardization, and weak linkage between activity values and three-dimensional structures.

### The DockTData project

DockTData introduces an automated pipeline to extract, transform, and integrate receptor–ligand data from public sources such as BindingDB, ChEMBL, and the Protein Data Bank (PDB).

### Open and free access

The dataset produced by DockTData is openly available and free of charge, supporting the development of machine learning models and scoring functions for affinity prediction.

## The DockTData Project

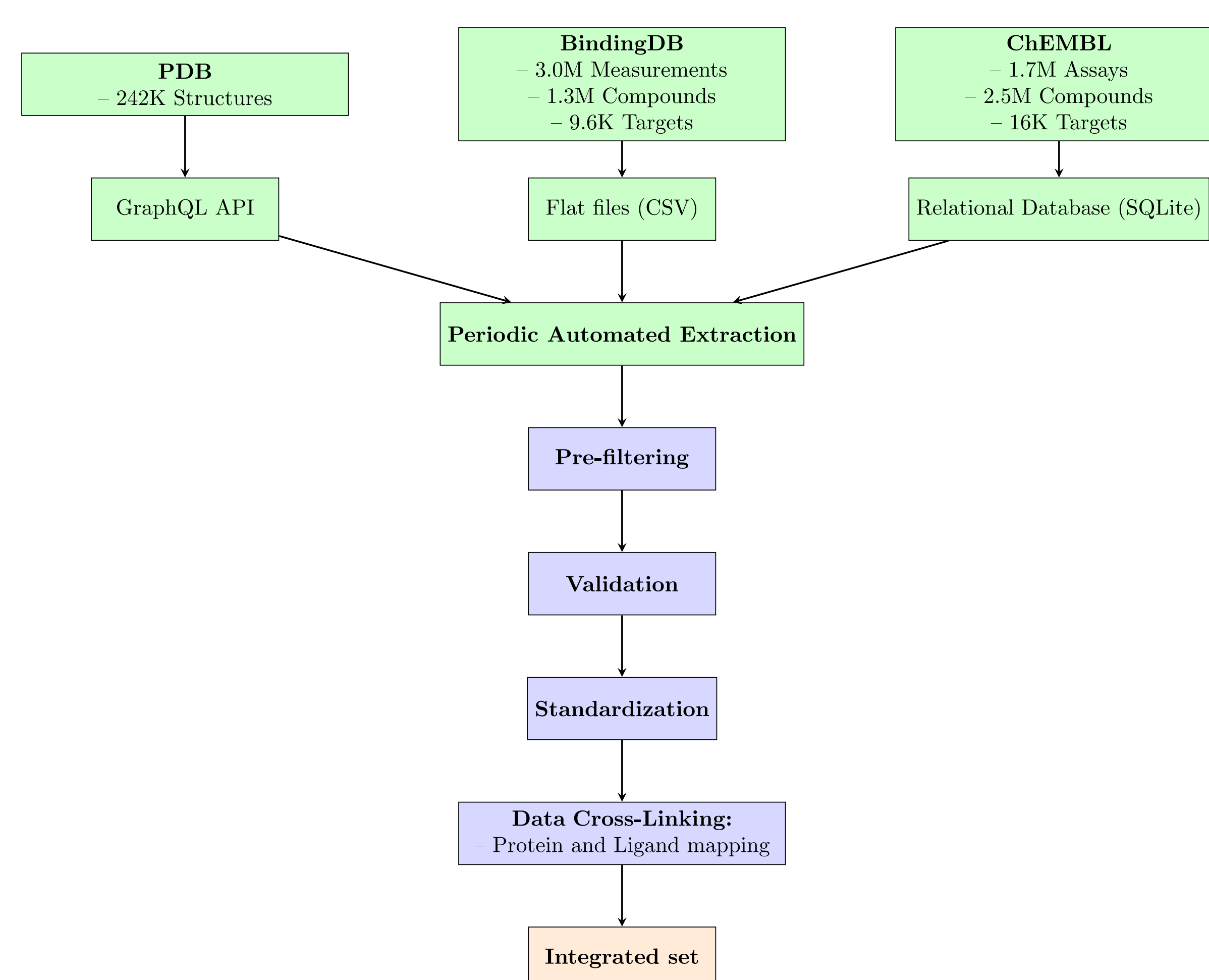


Fig 1 - DockTData Workflow

- **37,0K** Unique PDB structures
- **13,8K** Unique Ligands
- **Binding Affinity Types:** Ki, Kd, IC50 & EC50
- **Protein- & Nucleic acid- Ligand Complexes**

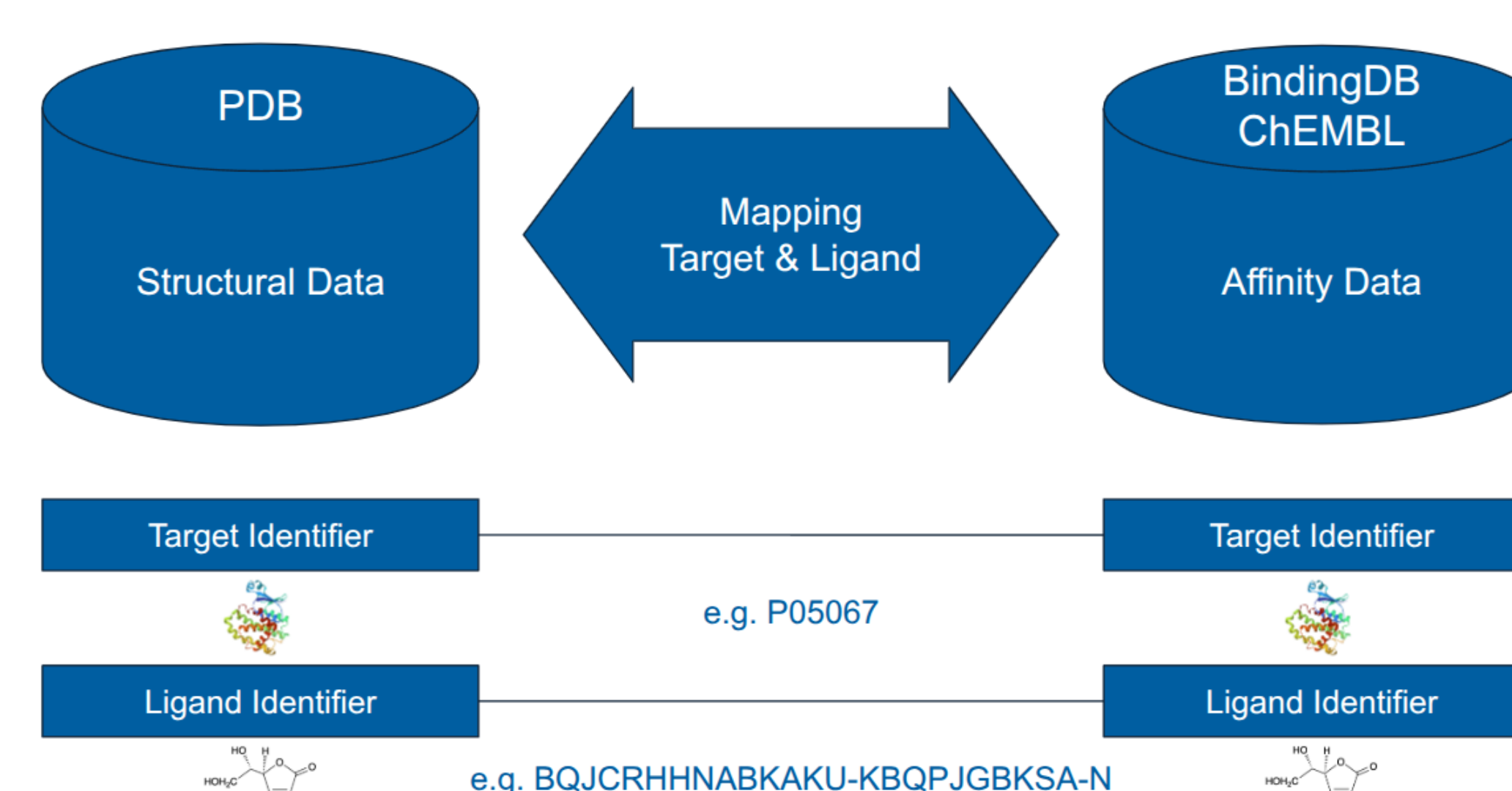


Fig 2 - Integration of Structural and Affinity Data

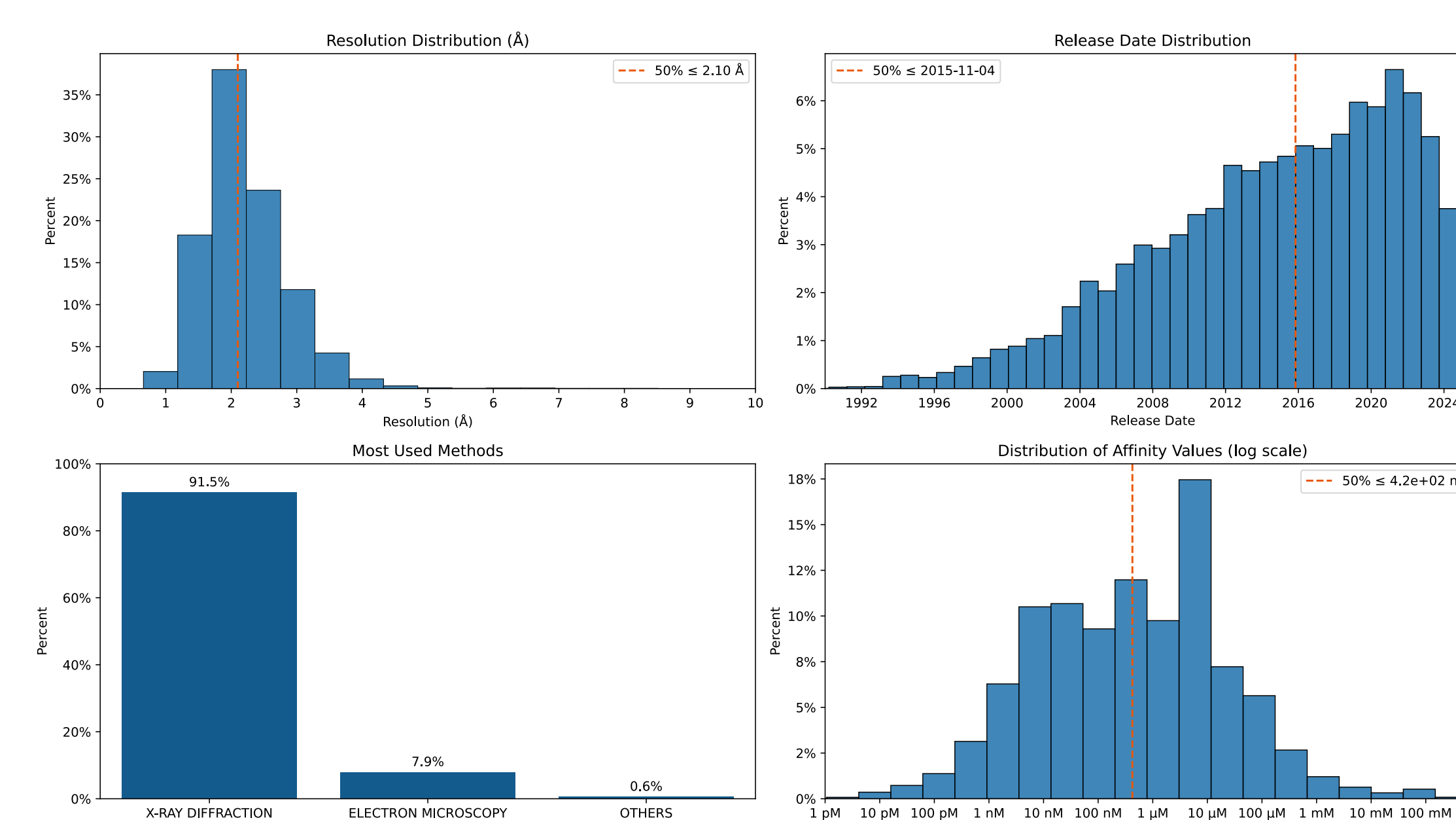


Fig 3 - Dataset Characterization

The figure summarizes deposited structural data, including resolution (median 2.10 Å), release date (median November 2015), experimental methods (dominated by X-ray diffraction, 91.5%), and affinity values (median  $4.2 \times 10^2$  nM, log scale).

## Conclusions

**DockTData** establishes a **scalable and reproducible** pipeline for integrating **structural and functional molecular data** into an **open and freely accessible resource** that enables **predictive modeling, virtual screening, and generative drug design**.

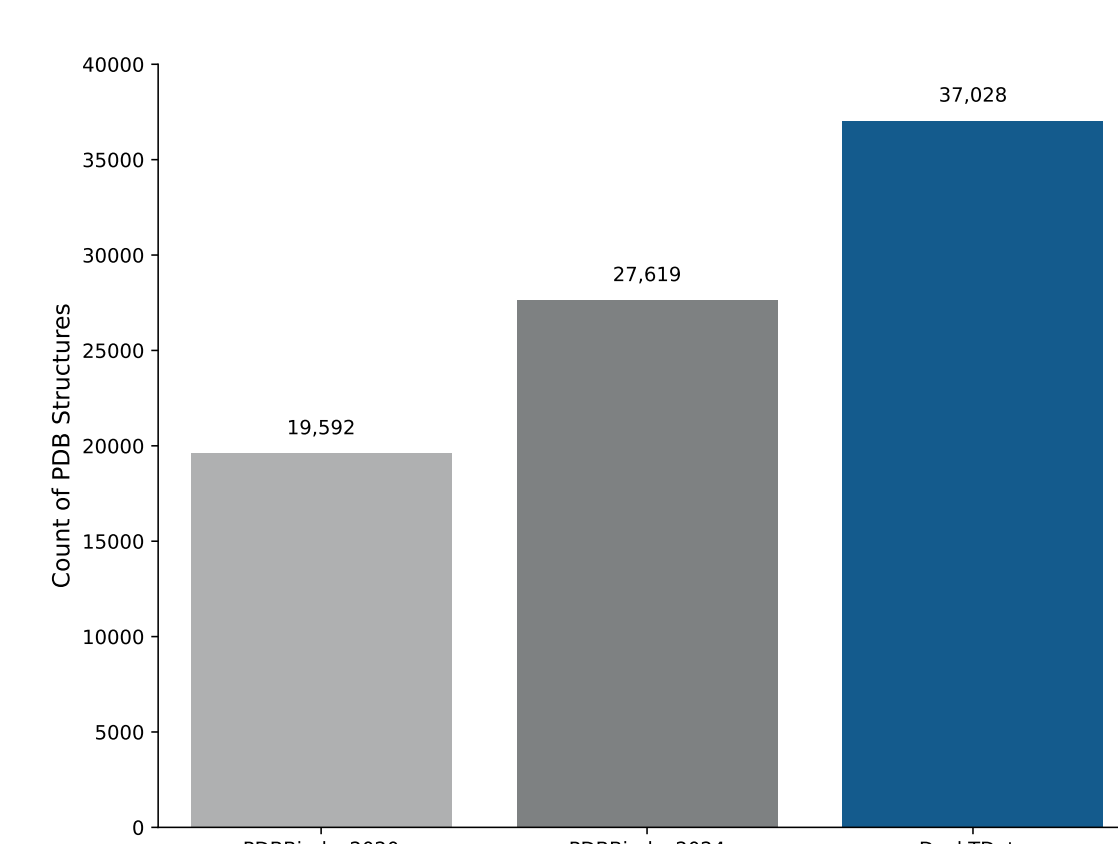


Fig 3 - Dataset Characterization  
PDBBind v.2020 is freely available, whereas PDBBind v.2024 requires a paid license (USD 2,000 for academic users), however both versions are distributed under a highly restrictive license.

## Further Information

For further information, please attend the Oral Communications in Session 2 on September 24 (Wednesday) at 16:55, or access the QR code provided.



This work was supported by CAPES, CNPq (grant number 309744/2022-9) and FAPERJ (grant numbers E-26/010.001415/2019, E-26/211.357/2021, E-26/200.393/2023).

[**BindingDB**] Liu, T. et al. Nucleic Acids Res. 2025, 53.  
[**PDB**] Berman, H. M. et al Nucleic Acids Res. 2000, 28.  
[**ChEMBL**] Mendez, D. et al. Nucleic Acids Res. 2019, 47.