DockTData

Automated Integration of Binding Affinity and Molecular Structure Data for Receptor—Ligand Interaction Modeling

José Renato D. Fajardo, Matheus M. P. da Silva, Leon S. C. Costa

Isabella A. Guedes, Laurent E. Dardenne

Molecular Modeling of Biological Systems Group (GMMSB)

GMMSB

GRUPO DE MODELAGEM MOLECULAR DE SISTEMAS BIOLÓGICOS

LNCC

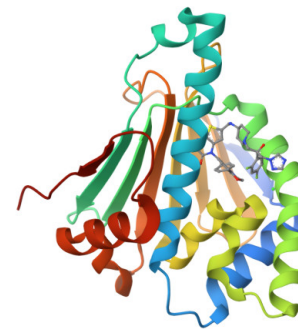
Laboratório Nacional de Computação Científica

Index

1 Background

The Central Role of Data

- **Binding affinity** as central task in **drug design**
- **Structural data** as a **rich information** reference
- **Structure-based affinity prediction models** are as key components for **Virtual Screening**, ***De Novo*** **Drug Design**
- Need for **open** and FAIR-compliant resources



Current Data Landscape

1 Background

- Data availability is the **bottleneck**: *no data = no AI/ML*
- **PDBbind**[1] was valuable but constrained by **restrictive licensing**
- **Boltz-2**[2] recent success highlights how **large-scale, curated datasets can unlock affinity prediction models**
  - but the pipelines for obtaining the data are **not readily available**
- **DockTData** contributes to this ecosystem as a **free and open resource**

[1] S. Passaro et al., *bioRxiv*, **2025**, 10.1101/2025.06.14.659707.

[2] Z. Liu et al., *Acc. Chem. Res.*, **2017**, 50.

Bridging Disciplines Through Data

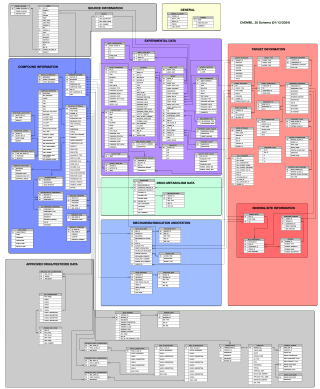
1 Background

- **Medicinal Chemists:** a hub for deposition, extraction and curation of experimental binding data
- **Machine Learning Community:** comprehensive training material with multiple representations of proteins and ligands
- **Bioinformatics & Molecular Modeling:** consolidated datasets for validation of docking, virtual screening & QSAR simulations

Data Integration Challenges

1 Background

- Missing **structural–functional links**
- Lack of supporting information and **documentation**
- **Data quality** and consistency issues
- **Scalability** for large datasets
- **Standardization** of data formats and identifiers
- Understanding different **data source structures**



Index

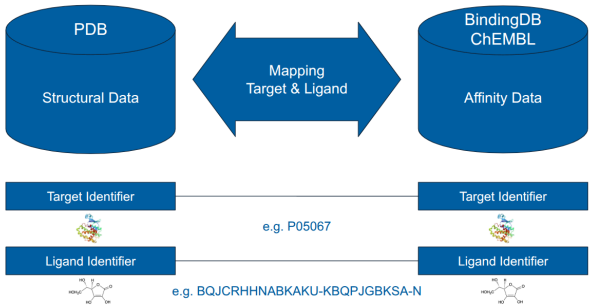
ETL Architecture

- **Extract:** PDB[3] (API), BindingDB[4] (flat files), ChEMBL[5] (relational DB)
- **Transform:** Validation, Filtering, Processing, Cross-linking
- **Load:** Structure the integrated set

[1] Berman, H. M. et al *Nucleic Acids Res.* **2000**, 28.
[2] Z. Liu, T. et al. *Nucleic Acids Res.*, **2025**, 53.
[3] Mendez, D. et al. *Nucleic Acids Res.* **2019**, 47.

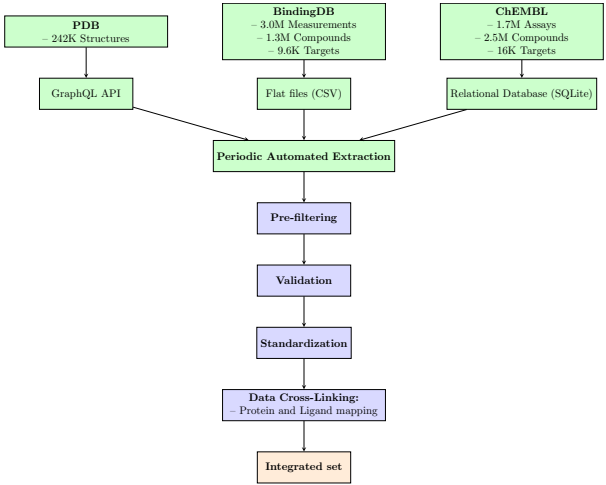
Data Cross-Linking

- **Protein mapping:** UniProt IDs (ChEMBL), sequence identity $\geq 85\%$ (BindingDB)
- **Ligand mapping:** InChIKey and CCD



Pipeline

2 The DockTData Project



Index

Dataset Composition

3 What We Have

- **37.0K** unique PDB structures
- **13.8K** unique ligands
- Binding affinity types: **Ki, Kd, IC50 & EC50**
- Protein- & Nucleic acid- ligand complexes

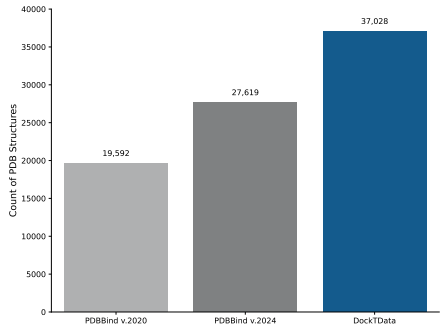
Democratizing data

3 What We Have

- **DockTData** (Last update 17-Set-25)
  Unique PDB structures: **37.0K**

- **PDBbind*** v.2020 (Free)
  Unique PDB structures: **19.5K**

- **PDBbind*** v.2024 (Paid**)
  Unique PDB structures: **27.6K**



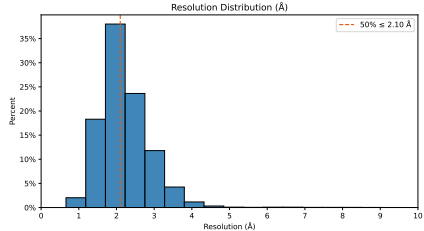
---

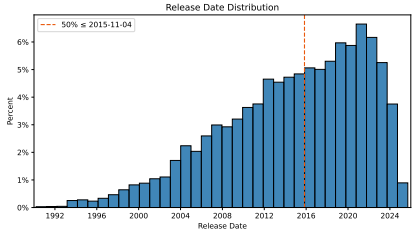
* Subject to a highly restrictive license

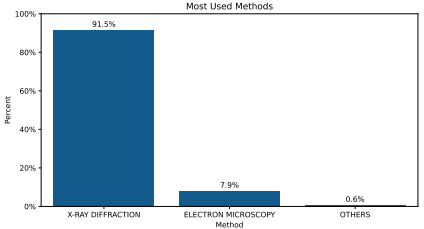
** Cost for academic users: USD 2,000

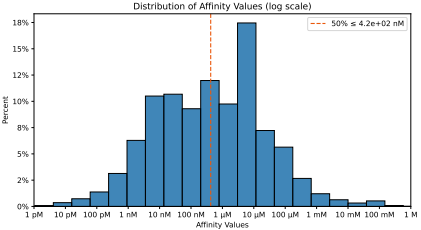
Dataset Characterization

3 What We Have









Index

4 Perspectives

Conclusions

4 Perspectives

- **Scalable, reproducible, automatic pipeline** integrating structural–functional data
- A resource for many purposes:
  - ML-based **Affinity Prediction**
  - **Virtual Screening** Validation
  - **Generative Models** for **Drug Design**
  - and **more**

Work-In-Progress

4 Perspectives

- **Peptide** and **Oligosaccharides** as ligands
- Target mapping by **sequence alignment** (MMseqs2)
- Subsets with **refined** filters
- Availability of **prepared structures** (e.g. protonation, cofactors)

Future Perspectives

4 Perspectives

- **Expansion:** Multi-ligand Complexes, Nucleotide Receptors
- Public **web portal** and **collaborative curation**
- **Call-to-action:** *from the community to the community*

Acknowledgement

4 Perspectives

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).

Obrigado!