DOI 10.24412/1829-0450-fm-2025-2-70-81
УДК 577.29(577.322.9)

Поступила: 23.10.2025г.
Сдана на рецензию: 24.10.2025г.
Подписана к печати: 31.10.2025г.

ARTIFACTS CAUSED BY CRYSTALLOGRAPHIC NEIGHBORS DURING DOCKING. THE IMPORTANCE OF THE BIOLOGICAL UNIT FOR EVALUATING DOCKING ACCURACY

H. Grabski

L.A. Orbeli Institute of Physiology NAS RA
Hovakim_grabski@outlook.com
ORCID: https://orcid.org/0000-0001-6115-9339

ABSTRACT

Reliable structure-based virtual screening critically depends on the accuracy of experimental protein–ligand complexes. However, many crystallographic models in the Protein Data Bank (PDB) contain crystal-packing neighbors that distort the local geometry of binding sites. In this work, the impact of crystallographic neighbors on molecular docking accuracy using Molsoft ICM-Pro was systematically assessed. All ligands were docked starting from two-dimensional structures, without prior conformational information, to mimic realistic virtual screening conditions. Across three representative systems *Schistosoma mansoni* SmBRD3(2), human TIM-3, and USP5 ZnF-UBD that crystal neighbors can have a huge impact was observed. Incorporating neighboring molecules yielded dramatic improvements, reducing RMSD values below 2 Å and substantially enhancing docking and RTCNN scores. The results emphasize the necessity of careful analysis of crystallographic structures before docking to ensure correct biological unit, reproducibility, reliability, and meaningful interpretation of computational screening outcomes.

Keywords: Virtual screening, Crystallographic neighbors, Molecular docking, Protein–ligand interactions, Structure-based drug discovery.

Introduction

Virtual screening (VS) has become an indispensable tool in contemporary drug discovery, enabling the rapid identification of candidate molecules from vast chemical libraries through computational evaluation of ligand–target interactions [1]. By integrating molecular docking, pharmacophore modeling, and machine learning-based scoring, VS substantially reduces the time and cost associated with early-stage drug development compared to purely experimental approaches. The reliability of virtual screening outcomes, however, is critically dependent on the structural accuracy and biological relevance of the macromolecular models employed.

Experimental structural biology methods, most notably X-ray crystallography and cryo-electron microscopy (cryo-EM), serve as the primary sources of three-dimensional protein structures for structure-based drug design [2,3]. These techniques have collectively produced hundreds of thousands of entries in the Protein Data Bank (PDB) [4], providing an unprecedented foundation for rational ligand discovery. Nonetheless, the interpretation of these structures for computational screening requires careful contextualization. Specifically, protein crystals represent periodic arrangements of molecules stabilized by intermolecular contacts that may not reflect biologically relevant conformations or interfaces.

A frequently overlooked aspect of this crystallographic context is the presence of crystallographic neighbors, or symmetry related molecules generated by the crystal lattice. These neighbors can create artificial interfaces or occlude biologically meaningful binding sites. When such artificial surfaces are inadvertently treated as part of the functional protein surface during virtual screening, docking algorithms may identify binding pockets or predict energetically favorable yet biologically irrelevant poses.

A recent and notable example of this issue arose in the PoseBusters benchmark [5], a large-scale evaluation framework designed to assess the accuracy of pose prediction algorithms. The initial dataset comprised over 500 protein-ligand complexes; however, following peer review, it was

identified that a significant portion of these entries were influenced by crystallographic neighbors. After these cases were removed, the benchmark was reduced to 393 validated examples. This correction not only highlighted the pervasiveness of crystal packing artefacts in publicly available structural data but also emphasized the necessity of rigorous structural validation in computational benchmarking. The PoseBusters case serves as a compelling reminder that even carefully curated datasets can be compromised by unrecognized crystallographic contacts, affecting the perceived performance of docking and scoring methods.

Therefore, increasing awareness of crystallographic artefacts and incorporating validation steps that distinguish biologically meaningful assemblies from packing, induced contacts is essential for improving the fidelity of structure, based virtual screening. This study underscores the significance of crystallographic neighbors in X-ray–derived protein structures and proposes practical approaches for their identification and mitigation, aiming to enhance the biological interpretability and predictive power of virtual screening workflows.

Material and Methods

Molecular Docking and Visualization

To quantitatively assess the impact of crystallographic neighbors on virtual screening outcomes, molecular docking simulations using Molsoft ICM-Pro were performed (version 3.9-4a; Molsoft LLC, La Jolla, CA, USA) [6]. All protein structures were visualized, prepared, and analyzed within the ICM-Pro environment. The software's docking algorithm employs a biased-probability Monte Carlo (BPMC) sampling approach for exploring ligand conformational space [6]. This stochastic method generates three-dimensional conformations by random perturbations of internal coordinates, guided by a probability function that biases sampling toward low-energy configurations.

Docking Configuration

For each target–ligand complex, the docking grid (or "box") was centered on the position of the co-crystallized ligand, with dimensions encompassing a 5 Å radius surrounding the ligand atoms. This configuration was chosen to ensure sufficient coverage of the binding pocket while avoiding inclusion of irrelevant surface regions or symmetry related molecules. To maintain computational realism and comparability with practical virtual screening workflows, the docking effort parameter was set to 5, and the maximum number of generated ligand conformations was limited to 10 per compound.

Ligand Preparation

In all docking experiments, ligands were provided exclusively in their two-dimensional (2D) representations (SMILES format) without any pre-assigned three-dimensional coordinates or conformational information. This approach ensured that each ligand's 3D geometry was generated from scratch during the docking process, thereby emulating realistic early-stage drug discovery scenarios where only chemical structure is known. No prior assumptions regarding bioactive conformations were made. Partial charges, rotatable bonds, and atom types were automatically assigned using ICM-Pro's internal parameterization routines.

Scoring and Pose Evaluation

The algorithm for conformational sampling 3D structures of ligands is generated randomly by biased probability Monte Carlo [6]. All scoring functions and predictions were performed by the method implemented in ICM-Pro v3.9-4a [6]. For each ligand, the top-scoring pose was retained for subsequent analysis. Comparisons were made between docking runs performed on native protein structures and those influenced by crystallographic neighbors, enabling a direct quantification of the artificial effects on pose prediction and score.

All docking simulations were executed on a Linux workstation equipped with 8 CPU cores and 64 GB of RAM. Visualization of resulting poses and protein–ligand interactions was performed using ICM-Pro.

Results

Case Study 1: Schistosoma mansoni Bromodomain 3 (SmBRD3(2))

To illustrate the practical implications of neglecting crystallographic neighbors in structure based virtual screening, the X-ray structure 7AMC [7] was examined, representing the Bromodomain 2 of Bromodomain containing protein 3 (SmBRD3(2)) from *Schistosoma mansoni* in complex with the small molecule inhibitor iBET726. *S. mansoni* is a parasitic trematode responsible for intestinal schistosomiasis, a major neglected tropical disease affecting millions globally. The pathology of schistosomiasis arises primarily from the host immune reaction to parasite eggs lodged in the intestinal and hepatic vasculature. Given its significant global health burden, *S. mansoni* proteins - including bromodomains involved in epigenetic regulation - represent promising therapeutic targets.

In the PDB: 7AMC structure, the co-crystallized inhibitor iBET726 (CCD: 73B) is bound within the canonical acetyl-lysine recognition pocket of SmBRD3(2). To evaluate the docking reliability, two separate docking experiments using Molsoft ICM-Pro under identical parameters were conducted - first using the isolated asymmetric unit, and then with crystallographic neighbors explicitly included. The resulting comparisons clearly demonstrated the influence of crystal packing on docking accuracy.

When the structure was treated in isolation, docking of iBET726 yielded an ICM score of –18.00 and an RTCNN score of –28.49, with a root-mean-square deviation (RMSD) of 8.00 Å relative to the experimental pose (Figure 1A). This RMSD indicates a severe misplacement of the predicted binding pose, suggesting that the docking algorithm failed to recapitulate the experimentally observed ligand position. The electron density maps show close contacts with symmetry-related molecules in the

crystal lattice, indicating that relevant intermolecular interactions had been omitted from the docking environment (Figure 1B).

After incorporating the crystallographic neighbors into the docking setup, the predicted results improved dramatically. The ICM docking score increased to –38.00, the RTCNN score to –41.00, and the RMSD dropped to 0.83 Å, indicating excellent agreement with the experimental ligand pose (Figure 1C).

These findings emphasize that overlooking crystallographic neighbors can lead to substantial errors in virtual screening and pose validation. Even when such contacts arise from crystal packing rather than true biological assemblies, their structural and electrostatic influence can distort docking results. Therefore, researchers should carefully examine electron density maps and crystal symmetry to determine whether neighboring molecules represent biologically meaningful interfaces or artefacts before employing such structures in computational studies.

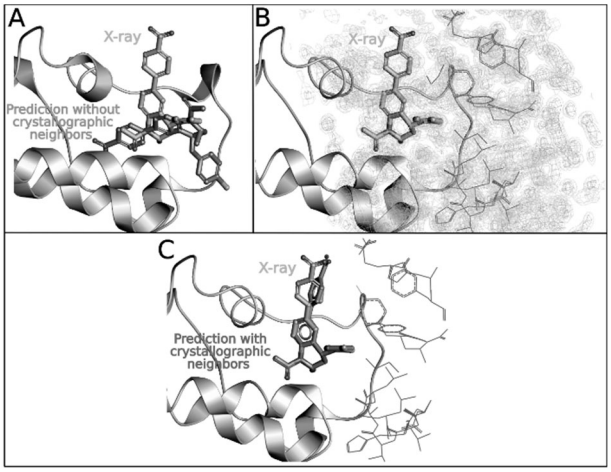

Fig. 1. Influence of crystallographic neighbors on docking accuracy (PDB ID: 7AMC). (A) Docking of the ligand into the crystallographic structure as deposited, showing suboptimal alignment with the experimental pose.
(B) Examination of the electron density map reveals the presence of a crystallographic neighbor interacting within the binding site, suggesting its potential influence on ligand positioning.
(C) Redocking performed with the crystallographic neighbor demonstrates a nearly perfect overlap with the experimental ligand conformation.

Case Study 2: Human TIM-3 Immune Checkpoint

A second example demonstrating the impact of crystallographic neighbors involves the X-ray structure 7M41, which captures the T-cell immunoglobulin and mucin domain-containing molecule 3 (TIM-3; HAVCR2) bound to the small molecule inhibitor compound **38**, formally described as N-(4-(8-chloro-2-methyl-5-oxo-5,6-dihydro-[1,2,4]triazolo [1,5-c]quinazolin-9-yl)-3-methylphenyl)-1H-imidazole-2-sulfonamide [8]. TIM-3 has emerged as an important immune checkpoint target in oncology, acting as a negative regulator of T-cell activation and contributing to immune exhaustion.

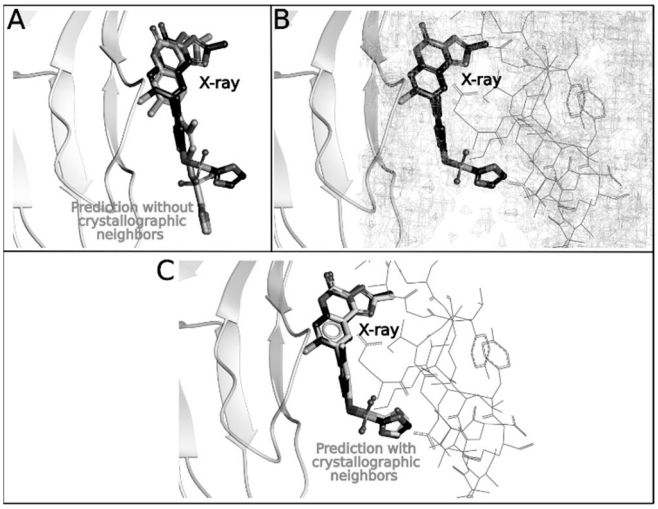

Fig. 2 Influence of crystallographic neighbors on docking accuracy (PDB ID: 7M41).
(A) Docking of the ligand into the crystallographic structure as deposited, showing suboptimal alignment with the experimental pose.
(B) Examination of the electron density map reveals the presence of a crystallographic neighbor interacting within the binding site, suggesting its potential influence on ligand positioning.
(C) Redocking performed with the crystallographic neighbor demonstrates a nearly perfect overlap with the experimental ligand conformation.

Using the 7M41 structure in its deposited form (without considering symmetry-related molecules), docking of compound **38** yielded an ICM

score of –15.00, an RTCNN score of –27.00, and an RMSD of 3.50 Å relative to the crystallographic ligand pose (Figure 2A). The moderate RMSD suggested a partially correct orientation but significant deviations from the experimentally observed geometry. Upon inspecting the electron density, the neighboring asymmetric units formed close packing interactions around the ligand pocket, suggesting that the absence of these contacts distorted the electrostatic and steric environment during docking (Figure 2B). Upon redocking with the inclusion of crystallographic neighbors, the docking score improved markedly to –43.74, the RTCNN score to –64.18, and the RMSD decreased to 0.27 Å, indicating near perfect agreement with the experimental pose (Figure 2C).

Case Study 3: USP5 Zinc-Finger Ubiquitin-Binding Domain (ZnF-UBD)

A third illustrative case is the X-ray structure of the USP5 ZnF-UBD co-crystallized with (5-((4-(4-chlorophenyl)piperidin-1-yl)sulfonyl)picolinoyl) glycine (PDB ID: 7MS7 [9]), a member of a novel chemical series that targets the C-terminal ubiquitin-binding site of USP5 [7]. USP5 (ubiquitin-specific protease 5) is a deubiquitinase implicated in several diseases, including cancer, through its role in ubiquitin recycling and proteostasis regulation. Despite its biological importance, no selective USP5-targeting chemical probe has yet been reported. The ZnF-UBD domain represents a secondary, poorly characterized binding region that can be exploited to allosterically inhibit catalytic activity.

When the default structure was used without considering symmetry related contacts, docking yielded a score of –31.00, an RTCNN score of –23.00, and an RMSD of 4.40 Å, indicating suboptimal pose prediction (Figure 3A). The electron density revealed proximity of the ligand pocket to a crystallographic neighbor, suggesting potential lattice mediated stabilization (Figure 3B). Upon repeating the docking with the neighbor included, the docking score improved to –36.00, the RTCNN score to –33.00, and the RMSD dropped dramatically to 0.54 Å (Figure 3B). This substantial improvement further demonstrates how crystallographic neighbors affect computational modeling.

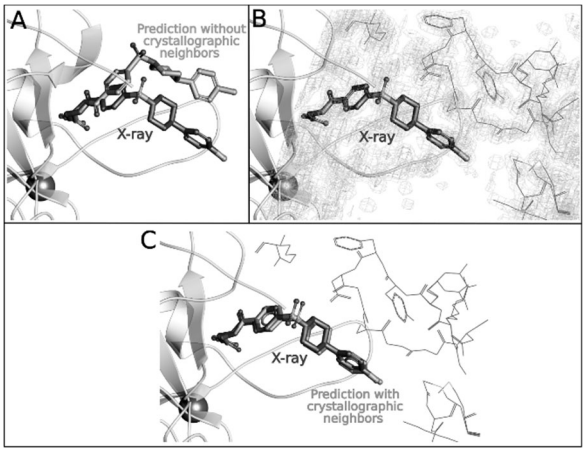

Fig. 3 Influence of crystallographic neighbors on docking accuracy (PDB ID: 7MS7). (A) Docking of the ligand into the crystallographic structure as deposited, showing suboptimal alignment with the experimental pose.
(B) Examination of the electron density map reveals the presence of a crystallographic neighbor interacting within the binding site, suggesting its potential influence on ligand positioning.
(C) Redocking performed with the crystallographic neighbor demonstrates a nearly perfect overlap with the experimental ligand conformation.

Conclusions

This study demonstrates that neglecting crystallographic neighbors in X-ray structures can have a profound and often underestimated impact on the accuracy of structure-based virtual screening. Through three representative case studies, *Schistosoma mansoni* SmBRD3(2), human TIM-3, and USP5 ZnF-UBD, showed that excluding neighboring molecules during docking consistently led to substantial deviations in ligand pose prediction and energy scoring. Inclusion of crystallographic neighbors, by contrast, restored agreement with experimental data and dramatically improved RMSD values, highlighting that even non-biological packing interactions can alter the physicochemical landscape of a binding site.

These findings underscore that many protein–ligand complexes deposited in the Protein Data Bank (PDB) contain symmetry related

contacts that can distort the perceived accessibility or geometry of a binding pocket. Consequently, automated docking workflows that ignore this context risk producing misleading or irreproducible results. This issue has direct implications not only for individual virtual screening projects.

By integrating awareness of crystallographic context and performing targeted validation, researchers can improve both the reliability and biological relevance of computational screening. As virtual screening continues to expand through automation and machine learning, the importance of structural correctness cannot be overstated.

To ensure robustness and reproducibility in structure based virtual screening, the following best practices are recommended:

1. Inspect crystallographic neighbors using molecular visualization tools (e.g., ICM-Pro, PyMOL, etc.) before initiating docking or virtual screening.
2. Verify biological assemblies through the PDB's BIOUNIT entries to differentiate biologically relevant interfaces from crystal packing artefacts.
3. Evaluate electron density maps (via Electron Density Server or PDB-REDO) to confirm the integrity and completeness of the ligand and nearby residues.
4. Avoid blind automation – manual validation of key structures remains essential, especially when preparing benchmark datasets or training data for AI-driven docking and scoring models.

Acknowledgements. *The author is thankful to Prof. Ruben Abagyan for fruitful and helpful discussion and access to computational infrastructure and for providing access to the MolSoft ICM-Pro software package, which was instrumental in conducting this study. The author is also thankful to Siranuysh Grabska for valuable feedback. The research was supported by the Science Committee of MESCS RA, in the frames of the research projects № 25FAST-1F002.*

REFERENCES

1. *Sabe V.T. et al.* Current trends in computer aided drug design and a highlight of drugs discovered via computational techniques: A review // European Journal of Medicinal Chemistry. 2021. Vol. 224. P. 113705.

2. *Renaud J.-P. et al.* Cryo-EM in drug discovery: achievements, limitations and prospects // Nat Rev Drug Discov. 2018. Vol. 17, № 7. PP. 471–492.
3. *Maveyraud L., Mourey L.* Protein X-ray Crystallography and Drug Discovery // Molecules. 2020. Vol. 25, № 5. P. 1030.
4. *Burley S.K. et al.* Protein Data Bank (PDB): The Single Global Macromolecular Structure Archive // Methods Mol Biol. 2017. Vol. 1607. PP. 627–641.
5. *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. 2024. Vol. 15, № 9. PP. 3130–3139.
6. *Abagyan R., Totrov M., Kuznetsov D.* ICM-A new method for protein modeling and design: Applications to docking and structure prediction from the distorted native conformation // J. Comput. Chem. 1994. Vol. 15, № 5. PP. 488–506.
7. *Schiedel M., McDonough M.A., Conway S.J.* SmBRD3(2), Bromodomain 2 of the Bromodomain 3 protein from Schistosoma mansoni in complex with iBET726: 7amc. 2021.
8. *Rietz T.A. et al.* Fragment-Based Discovery of Small Molecules Bound to T-Cell Immunoglobulin and Mucin Domain-Containing Molecule 3 (TIM-3) // J. Med. Chem. 2021. Vol. 64, № 19. PP. 14757–14772.
9. *Mann M.K. et al.* Structure–Activity Relationship of USP5 Inhibitors // J. Med. Chem. 2021. Vol. 64, № 20. PP. 15017–15036.

АРТЕФАКТЫ, ВЫЗВАННЫЕ КРИСТАЛЛОГРАФИЧЕСКИМИ СОСЕДЯМИ ПРИ ДОКИНГЕ. ВАЖНОСТЬ БИОЛОГИЧЕСКОЙ ЕДИНИЦЫ ДЛЯ ОЦЕНКИ ТОЧНОСТИ ДОКИНГА

О.В. Грабский

Институт Физиологии им. Л.А.Орбели НАН РА

АННОТАЦИЯ

Надежность виртуального скрининга, основанного на структурных данных, критически зависит от точности экспериментальных комплексов белок–лиганд. Однако многие кристаллографические модели, представленные в Банке данных белковых структур (PDB), содержат кристаллографические соседи, которые искажают локальную геометрию сайтов связывания. В данной работе систематически оценивалось влияние кристаллографических соседей на точность молекулярного докинга с использованием программы Molsoft ICM-Pro. Все ли-

ганды докировались, начиная с двумерных структур, без предварительной информации о конформации, чтобы имитировать реалистичные условия виртуального скрининга. Для трех репрезентативных систем – *Schistosoma mansoni* SmBRD3(2), человеческого TIM-3 и USP5 ZnF-UBD – было показано, что кристаллографические соседи могут оказывать значительное влияние. Включение кристаллографических соседей привело к резкому улучшению результатов, снижению RMSD ниже 2 Å и существенному повышению оценок докинга. Результаты подчеркивают необходимость тщательного анализа кристаллографических структур перед докингом для обеспечения корректного выбора биологической единицы, воспроизводимости, надежности и осмысленной интерпретации результатов вычислительного скрининга.

Ключевые слова: Виртуальный скрининг, кристаллографические соседи, молекулярный докинг, взаимодействия белок–лиганд, структурно-ориентированный поиск лекарств.