



**INTERNATIONAL JOURNAL OF
PHARMACEUTICAL SCIENCES**
[ISSN: 0975-4725; CODEN(USA): IJPS00]
Journal Homepage: <https://www.ijpsjournal.com>



Research Paper

Structure-Based Virtual Screening and AI-Assisted Prioritization of Small-Molecule RANKL-Binding Candidates for Rheumatoid Arthritis-Associated Bone Erosion

Chinmaya Chidananda Behera, Rohit Dash, Satya Prakash Panda, Sagar Kumar Mishra*

University Department of Pharmaceutical Sciences, Utkal University, Vani Vihar, Bhubaneswar, Odisha, India 751004.

ARTICLE INFO

Published: 01 Oct 2026

Keywords:

Rheumatoid arthritis;
RANKL; Virtual screening;
Random Forest; Machine
learning; ADMET;
Structural alerts

DOI:

10.5281/zenodo.23074259

ABSTRACT

Receptor activator of nuclear factor- κ B ligand (RANKL) is a key mediator of osteoclast differentiation and bone erosion in rheumatoid arthritis (RA). An integrated structure-based and machine-learning workflow was developed to prioritize small-molecule RANKL-binding candidates. Approximately 10,000 ZINC-derived compounds were docked against murine RANKL (PDB ID: 1S55) using AutoDock Vina, and the top 1,000 poses were curated, yielding 671 unique chemical structures after duplicate removal. RDKit generated 1,407 molecular features comprising 217 descriptors, 166 MACCS keys and 1,024 ECFP4 fingerprint bits. The top 25% of the docking-score distribution (Vina ≤ -7.7815 kcal/mol; $n = 168$) was defined as the docking-derived elite class. A Random Forest classifier was trained using an 80:20 stratified train/test split, training-set-only feature filtering and 5-fold stratified cross-validation. The classifier achieved a cross-validated ROC-AUC of 0.648 ± 0.062 and a held-out test ROC-AUC of 0.617. For downstream prioritization, a model refitted to all 671 compounds ranked candidates according to their predicted probability of belonging to the elite class. ZINC70673359 ranked first, with a predicted elite probability of 97.6% and Vina score of -8.130 kcal/mol. Its computational profile indicated molecular weight of 549.413 Da, LogP of -1.138 , TPSA of $136.99 \text{ \AA}^2$, low predicted gastrointestinal absorption, no predicted BBB permeability and no PAINS alerts, whereas three Brenk structural-alert patterns were identified. ZINC70673359 therefore represents a computationally prioritized candidate for experimental evaluation. However, interpretation is limited by the docking-derived class definition, modest classifier performance and rule-based ADMET assessment

***Corresponding Author:** Sagar Kumar Mishra

Address: University Department of Pharmaceutical Sciences, Utkal University, Vani Vihar, Bhubaneswar, Odisha, India 751004

Email ✉: sagar.cognosy@gmail.com

Relevant conflicts of interest/financial disclosures: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.



INTRODUCTION

Rheumatoid arthritis (RA) is a chronic autoimmune polyarthritis characterised by synovial inflammation and progressive, often irreversible, periarticular bone erosion. A substantial part of this bone destruction is mediated through the receptor activator of nuclear factor- κ B ligand (RANKL)–RANK–osteoprotegerin (OPG) axis: RANKL, produced by activated T-cells, synovial fibroblasts and osteoblasts within the inflamed joint, binds RANK on osteoclast precursors and drives their differentiation and resorptive activity [1, 2]. The clinical relevance of this pathway is well established: the anti-RANKL monoclonal antibody denosumab measurably slows radiographic joint damage in RA, independent of its effect on synovitis. However, denosumab remains the only RANKL-targeted agent in clinical use, and as a parenterally administered biologic it carries the cost, injection burden and accessibility limitations characteristic of that drug class – motivating continued interest in an orally deliverable small-molecule alternative acting on the same pathway [3, 4].

RANKL nevertheless presents a non-trivial medicinal chemistry target: as a member of the tumour necrosis factor superfamily, it functions as a homotrimer and engages RANK across a comparatively large, shallow protein–protein interaction (PPI) interface rather than a classical, deep enzymatic pocket. The availability of high-resolution structural data - including the murine RANKL extracellular-domain crystal structure deposited under PDB accession 1S55 [5, 6] - together with the maturation of computer-aided drug design methodology, has made structure-based, in silico-guided discovery of RANKL-binding small molecules an increasingly tractable proposition.

Computational drug-discovery pipelines increasingly combine structure-based docking with machine-learning-based descriptor analysis, quantitative structure–activity relationship (QSAR) modelling and in silico ADMET/toxicity-alert screening, rather than relying on docking score alone [4, 5]. This hybrid approach allows a screened library to be triaged not only by predicted binding affinity but also by drug-likeness and the presence of known problematic substructures (pan-assay-interference and reactive/toxic structural alerts), which is particularly important for PPI targets such as RANKL, where docking scores alone are a comparatively weak proxy for true binding.

The present study reports an auditable computational workflow that combines structure-based virtual screening, chemical-structure curation, RDKit descriptor/fingerprint profiling, docking-derived Random-Forest classification and validation, RF-based candidate prioritization, and rule-based ADMET/structural-alert screening. Rather than treating the docking score as an experimentally validated activity label, the analysis explicitly defines a docking-derived elite class and evaluates the classifier using both cross-validation and a held-out test set. The final model is then used to prioritize chemically distinct candidates for experimental follow-up.

2. MATERIALS AND METHODS

2.1. Overall Workflow

The study was implemented as an integrated computational pipeline comprising: (i) RANKL structure retrieval and receptor preparation; (ii) ZINC-derived ligand preparation and AutoDock Vina docking; (iii) selection of the top 1,000 docking poses; (iv) canonical-SMILES curation and removal of duplicate chemical structures; (v) RDKit descriptor and fingerprint calculation; (vi) definition of a docking-derived elite class and



Random-Forest model development/validation; (vii) RF-based prioritization of the complete curated library; and (viii) rule-based ADMET and PAINS/Brenk structural-alert assessment of the top 20 RF-prioritized candidates, with detailed characterization of the highest-ranked candidate (Fig. 1).

2.2. Molecular Docking

The three-dimensional structure of the murine RANKL extracellular domain was obtained from the RCSB Protein Data Bank (PDB ID: 1S55; 1.9 Å resolution) [9]. The receptor structure was prepared using the established Python-based workflow used for this study, including removal of crystallographic water molecules and co-crystallized ligands, protonation at pH 7.4, and conversion to AutoDock-compatible PDBQT format. Ligands were prepared using RDKit, Open Babel and Meeko [10], including structure validation, three-dimensional structure generation, partial-charge assignment and PDBQT conversion [11]. The docking region was defined from the previously reported RANKL binding-site residues used in the original study [9]: LYS A:180, TYR A:234, GLN A:236, MET A:238, LYS A:256, ARG B:222, HIS B:224, GLU B:268, ASP B:299 and ASP B:303. The resulting grid-box centre was $X = 8.632$, $Y = 32.246$, $Z = 48.191$ Å, with dimensions $X = 30.673$, $Y = 31.586$, $Z = 26.502$ Å.

Molecular docking was performed using AutoDock Vina with exhaustiveness = 10, output modes = 1 and an energy range of 3.0 kcal/mol; the default Vina random seed was retained [10]. The top 1,000 docking poses were retained from the original screening workflow. For each retained record, the lowest available Vina affinity was used for ranking. Subsequent curation in Step 0 canonicalized the molecular structures and removed duplicate chemical structures, retaining

the record with the best (most negative) Vina score for each canonical SMILES [12].

Docking scores were additionally converted, for interpretive purposes only, to an indicative predicted inhibition constant (K_i) via the classical relationship linking free energy to an equilibrium binding constant (Eq. 1):

$$\Delta G = RT \ln K_i \Rightarrow K_i = \exp (\Delta G / RT) \dots\dots\dots (1)$$

where R is the universal gas constant (1.9872×10^{-3} kcal mol⁻¹ K⁻¹) and $T = 298.15$ K ($RT = 0.5925$ kcal/mol). The calculated (K_i) values were expressed in nanomolar (nM) for reporting and interpretive purposes. Values derived in this way are theoretical, scoring-function-derived quantities and are not a substitute for experimentally determined binding constants [13].

2.3. Molecular Descriptor and Fingerprint Calculation

For each curated structure, RDKit generated 217 molecular descriptors, a 166-bit MACCS structural-key fingerprint and a 1,024-bit ECFP4 fingerprint (bond radius 2), yielding 1,407 numerical features per compound [14,15]. The descriptor set included molecular weight, calculated LogP, TPSA, hydrogen-bond donor/acceptor counts, rotatable-bond count, ring descriptors, electrotopological descriptors and QED [16]. Descriptor calculation was performed on canonical SMILES generated during structure curation. The resulting $671 \times 1,407$ feature matrix was generated.

2.4. Drug-Likeness Assessment

Lipinski Rule-of-Five properties were retained as part of the RDKit descriptor set. For descriptive analysis, a compound was considered Ro5-compliant when the total number of violations was ≤ 1 , using the conventional cut-offs for molecular



weight, LogP, hydrogen-bond donors and hydrogen-bond acceptors [17].

$$N_{\text{viol}} = 1[\text{MW} > 500] + 1[\text{LogP} > 5] + 1[\text{HBD} > 5] + 1[\text{HBA} > 10] \dots\dots\dots (2)$$

2.5. Random-Forest Classification and Feature Selection

The docking score was used only to define the machine-learning target class: compounds with Vina scores at or below the 25th percentile of the curated distribution was labelled elite (class 1), while the remainder were labelled non-elite (class 0). For the 671-compound dataset, the threshold was -7.7815 kcal/mol, giving 168 elite and 503 non-elite structures. An 80:20 stratified train/test split was generated with random state = 42. Feature filtering was performed using only the training set: zero-variance removal followed by removal of descriptors with absolute pairwise correlation >0.90 . After filtering, 1,066 features remained in the full matrix used for final scaling. A Random Forest classifier with 500 trees, balanced class weighting and random state = 42 was used. The 20 highest-importance features from the training model were retained for the final classifier [17]. Standard scaling was applied consistently to the selected features. Five-fold stratified cross-validation was performed within the training set, and model performance was additionally evaluated on the held-out test set using accuracy, balanced accuracy, precision, recall, F1-score and ROC-AUC. Following validation, a final Random Forest model was refitted to all 671 curated structures for downstream AI prioritization. The Random Forest approach follows the established ensemble framework described by Breiman [18].

2.6. Random-Forest Model Validation and AI Prioritization

Model validation showed a 5-fold cross-validation ROC-AUC of 0.648 ± 0.062 within the training partition. On the held-out 20% test partition, accuracy was 0.600, balanced accuracy 0.577, precision 0.321, recall 0.529, F1-score 0.400 and ROC-AUC 0.617. The corresponding confusion matrix contained 63 true non-elite and 18 true elite predictions, with 38 false-positive and 16 false-negative classifications. These results indicate useful but limited discrimination and were therefore used to support ranking/prioritization rather than to claim validated biological activity [19, 20].

2.7. Computational ADMET and Structural-Alert Screening

The top 20 RF-prioritized compounds were evaluated using RDKit-derived molecular weight, LogP, TPSA, hydrogen-bond donor/acceptor counts [21], and two rule-based permeability classifications such as Gastrointestinal (GI) absorption classification, and Blood-brain barrier (BBB) permeability classification [22]. The gastrointestinal-absorption classification used the Egan-style limits encoded in the script, while the BBB classification used the implemented molecular-weight, LogP, TPSA, HBA and HBD heuristic. PAINS and Brenk structural alerts were evaluated using the RDKit FilterCatalog, with the alert concepts contextualized by the published PAINS and Brenk literature [23, 24]. These outputs are cheminformatics rule-based predictions and were not generated by SwissADME, admetSAR3.0 [25, 26] or another external web platform.

2.8. Ligand Efficiency Metrics

For the RF-prioritized lead, ligand efficiency (LE) and lipophilic ligand efficiency (LLE) were calculated from the docking score, heavy-atom count and indicative K_i relationship solely for



interpretive comparison. Because the Vina score is a scoring-function output, the resulting K_i , LE and LLE values are theoretical and should not be interpreted as experimentally measured affinity metrics [27, 28].

$$LE = -\Delta G / N_{\text{heavy}} \dots\dots\dots (3)$$

$$LLE = pK_i - \text{LogP} \dots\dots\dots (4)$$

where $pK_i = -\log_{10}(K_i)$.

2.9. Statistical Analysis

Descriptive statistics (mean $\pm$ SD, median and range) were calculated for the curated 671-compound library using pandas. Model performance was summarized by 5-fold stratified cross-validation on the training partition and by evaluation on the held-out 20% test partition. Figures were generated in Python using matplotlib/seaborn; the two-dimensional lead structure was rendered using RDKit [29, 30].

3. RESULTS AND DISCUSSION

3.1. Virtual Screening and Docking Outcomes

The original 1,000 retained docking poses contained 329 duplicate chemical records after canonical-SMILES curation, leaving 671 unique structures for machine-learning analysis. Vina scores among the unique structures ranged from -8.742 to -7.229 kcal/mol (mean -7.609 ± 0.296 kcal/mol; median -7.540 kcal/mol) (Fig. 2). The strongest docking score was observed for ZINC70701164 (-8.742 kcal/mol), followed by ZINC70700221 (-8.669 kcal/mol) and ZINC15968013 (-8.599 kcal/mol). The RF-prioritized lead identified later in the workflow, ZINC70673359, had a Vina score of -8.130 kcal/mol and therefore was not the top-scoring docking compound.

3.2. Physicochemical and Drug-Likeness Profile

Across the 671 unique structures, mean molecular weight was 439.57 ± 53.23 Da, mean LogP was 0.43 ± 0.97 , mean TPSA was 96.26 ± 31.92 Å² and mean rotatable-bond count was 4.87 ± 2.88 . All 671 structures (100%) satisfied the study definition of Ro5 compliance (≤ 1 violation), while 610 compounds (90.9%) had zero violations. QED ranged from 0.119 to 0.852 (mean 0.508 ± 0.151); 234 compounds (34.9%) had QED ≥ 0.5 and 48 compounds (7.2%) had QED ≥ 0.7 (Table 1).

3.3. Random-Forest Feature Importance

The Random Forest identified a compact set of descriptor-level features associated with the docking-derived elite classification (Fig. 3; Table 2). Ipc was the most important feature (mean decrease in impurity = 0.02436), followed by VSA_EState3 (0.01745), MolLogP (0.01541), QED (0.01510) and BCUT2D_MRLOW (0.01504). The remaining high-importance features were dominated by electrotopological-state/van der Waals surface-area descriptors, including VSA_EState2, VSA_EState4, VSA_EState5, VSA_EState6, VSA_EState7 and EState_VSA2, together with SPS, MinEStateIndex, PEOE_VSA1/9, SMR_VSA1 and SlogP_VSA2. These importances describe the variables most frequently used by the fitted forest to reduce class impurity; they do not by themselves establish causal binding determinants.

The workflow produced a distinct RF-based prioritization of the curated library. After refitting the model to all 671 structures, compounds were ranked primarily by predicted probability of belonging to the docking-derived elite class, with Vina score retained as supporting information and a tie-breaker. ZINC70673359 ranked first, with a model-derived elite probability of 97.6% and a Vina score of -8.130 kcal/mol. Notably, the candidate was not the best-docked structure in the library, demonstrating that the RF ranking was not simply identical to the original Vina ranking. The



ZINC70673359 record belonged to the held-out test partition during model validation and was labelled elite (Vina = -8.130 kcal/mol); however, the reported 97.6% probability was generated after the model was refitted using all 671 structures and therefore is not a held-out test probability (Table 3). Because the class label itself was derived from Vina scores, the RF should be interpreted as a docking-score surrogate/prioritization model rather than as an experimentally validated bioactivity predictor.

3.4. RF Model Validation and AI Prioritization

The workflow produced a distinct RF-based prioritization of the curated library. After refitting the model to all 671 structures, compounds were ranked primarily by predicted probability of belonging to the docking-derived elite class, with Vina score retained as supporting information and a tie-breaker. ZINC70673359 ranked first, with a model-derived elite probability of 97.6% and a Vina score of -8.130 kcal/mol. Notably, the candidate was not the best-docked structure in the library, demonstrating that the RF ranking was not simply identical to the original Vina ranking. The ZINC70673359 record belonged to the held-out test partition during model validation and was labelled elite (Vina = -8.130 kcal/mol); however, the reported 97.6% probability was generated after the model was refitted using all 671 structures and therefore is not a held-out test probability (. Because the class label itself was derived from Vina scores, the RF should be interpreted as a docking-score surrogate/prioritization model rather than as an experimentally validated bioactivity predictor.

Using the docking score of -8.130 kcal/mol and 43 heavy atoms from the RDKit descriptor dataset, the indicative K_i calculated from $\Delta G = RT \ln K_i$ was approximately $1.10 \mu\text{M}$ ($1,098 \text{ nM}$; $\text{p}K_i \approx 5.96$). The corresponding theoretical ligand efficiency was $0.189 \text{ kcal mol}^{-1}$ per heavy atom

and the lipophilic ligand efficiency was approximately 7.10. These quantities are scoring-function-derived estimates and should not be presented as experimental binding constants or experimentally determined ligand efficiencies.

3.5. ADMET Profile, Structural Alerts and Ligand Efficiency

ZINC70673359 (Fig. 4; Table 4) was the highest-ranked RF candidate. Its calculated molecular weight was 549.413 Da , LogP -1.138 , TPSA $136.99 \text{ \AA}^2$, with 3 hydrogen-bond donors and 6 acceptors. The implemented rule-based analysis classified gastrointestinal absorption as Low and BBB permeability as No (-). No PAINS alerts were detected. Brenk filtering returned three alert patterns: imine_1, oxygen–nitrogen single bond and triple bond. The compound therefore combines high RF prioritization probability with several physicochemical and structural-alert features that warrant experimental and medicinal-chemistry scrutiny.

Using the docking score of -8.130 kcal/mol and 43 heavy atoms from the RDKit descriptor dataset, the indicative K_i calculated from $\Delta G = RT \ln K_i$ was approximately $1.10 \mu\text{M}$ ($1,098 \text{ nM}$; $\text{p}K_i \approx 5.96$). The corresponding theoretical ligand efficiency was $0.189 \text{ kcal mol}^{-1}$ per heavy atom and the lipophilic ligand efficiency was approximately 7.10 (Table 4; Fig. 4). These quantities are scoring-function-derived estimates and should not be presented as experimental binding constants or experimentally determined ligand efficiencies. The 2D structure of RF-prioritized candidate ZINC70673359 is presented in (Fig. 5). The ligand-protein interactions of ZINC70673359 with 1S55 are presented in (Fig. 6).

3.6. Limitations



Several limitations must be acknowledged. First, AutoDock Vina provides an approximate scoring function and does not fully account for receptor flexibility, explicit solvent effects or entropic contributions; docking was performed against a single RANKL conformation (PDB 1S55). Second, the machine-learning target is not an experimental activity label: the elite/non-elite class was defined directly from the Vina-score distribution. Consequently, the RF model learns a chemical-feature representation of docking-derived class membership and cannot, from this dataset alone, establish biological RANKL inhibition. Third, model performance was modest (5-fold training ROC-AUC 0.648 ± 0.062 ; held-out test ROC-AUC 0.617), so the 97.6% probability assigned by the final refitted model should be regarded as a prioritization score rather than calibrated probability of biological activity. Fourth, the final model was refitted to all 671 structures for prioritization, so downstream ranking should not be confused with independent test-set prediction. Fifth, the ADMET outputs were generated using RDKit rule-based heuristics and structural-alert filters rather than experimental pharmacokinetics or a dedicated external ADMET platform. Finally, the Brenk alerts and low predicted gastrointestinal absorption of ZINC70673359 indicate liabilities that should be addressed before considering the compound a lead for synthesis or biological testing.

CONCLUSION

The workflow combined structure-based docking, chemical-structure curation, RDKit descriptor/fingerprint profiling, docking-derived Random-Forest classification, model validation, RF-based prioritization and rule-based ADMET/structural-alert screening. From 1,000 retained docking poses, canonical-structure curation produced 671 unique compounds. The RF model defined the top 25% of the docking-score

distribution as the elite class and achieved a 5-fold training ROC-AUC of 0.648 ± 0.062 and a held-out test ROC-AUC of 0.617. After refitting to the complete curated dataset, ZINC70673359 was ranked first with a model-derived elite probability of 97.6%, despite having a Vina score of -8.130 kcal/mol rather than the best docking score. Its rule-based profile showed low predicted gastrointestinal absorption, no predicted BBB permeability, no PAINS alerts and three Brenk structural-alert patterns. Accordingly, ZINC70673359 is best described as an RF-prioritized computational candidate for experimental validation, not as an experimentally confirmed RANKL inhibitor. The present results provide an auditable computational prioritization framework while highlighting the need for biochemical binding assays, cellular osteoclast assays and, if supported, subsequent structure-based optimization.

Declarations

• Funding

The author(s) reported no funding associated with the work featured in this article.

• Conflict of interest

The author declares that there are no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

• Author contributions

C.C.B.: Conceptualization, methodology, investigation, data curation, formal analysis, writing-original draft, and visualization.

R.D.: Investigation, data curation, and writing-review and editing.

S.P.: Investigation, data curation, and writing-review and editing.



S.K.M.: Supervision, conceptualization, and writing-review and editing.

All authors have read and approved the final manuscript.

• Ethics approval

This study did not involve human participants, animal experiments, or clinical data; therefore, ethical approval was not required.

• Consent for publication

Not applicable, as the manuscript does not contain any individual person's data in any form.

• Data Availability

The datasets generated and/or analyzed during the current study are available from the corresponding author upon reasonable request.

• Acknowledgements

The authors are thankful to the University Department of Pharmaceutical Sciences, Utkal University for providing computational facilities and infrastructural support for the present study.

REFERENCES

- 1 Smolen, JS, Aletaha D, McInnes IB: Rheumatoid arthritis, *Lancet* (2016), 388: 2023–2038. [https://doi.org/10.1016/S0140-6736\(16\)30173-8](https://doi.org/10.1016/S0140-6736(16)30173-8).
- 2 McInnes IB, Schett G: Pathogenetic insights from the treatment of rheumatoid arthritis, *Lancet* (2017), 389: 2328–2337. [https://doi.org/10.1016/S0140-6736\(17\)31472-1](https://doi.org/10.1016/S0140-6736(17)31472-1).
- 3 Köhler BM, Günther J, Kaudewitz D, Lorenz HM: Current therapeutic options in the treatment of rheumatoid arthritis, *J. Clin. Med.* (2019), 8: 938. <https://doi.org/10.3390/jcm8070938>.
- 4 Burmester GR, Pope JE: Novel treatment strategies in rheumatoid arthritis, *Lancet* (2017), 389: 2338–2348. [https://doi.org/10.1016/S0140-6736\(17\)31491-5](https://doi.org/10.1016/S0140-6736(17)31491-5).
- 5 Jiang M, Peng L, Yang K, Wang T, Yan X, Jiang T, et al.: Receptor activator of nuclear factor- κ B ligand (RANKL)-receptor activator of nuclear factor- κ B (RANK) protein–protein interaction inhibitors based on a benzene sulfonamide scaffold, *J. Med. Chem.* (2019), 62: 576–594. <https://doi.org/10.1021/acs.jmedchem.8b02027>.
- 6 Nakai Y, Okamoto K, Terashima A, Ehata S, Nishida J, Imamura T, et al.: Efficacy of an orally active small-molecule inhibitor of RANKL in bone metastasis, *Bone Res.* (2019), 7: 26. <https://doi.org/10.1038/s41413-018-0036-5>.
- 7 Dey M, Amour A: Applications of artificial intelligence in drug discovery. *Emerg Top Life Sci.* 2025. doi:10.1042/ETLS20250001.
- 8 Gangwal A, Lavecchia A: IMPACT framework: establishing global standards for artificial intelligence implementation, methodology, and translation in drug discovery. *WIREs Comput Mol Sci.* (2026), 16(2). doi:10.1002/wcms.70072.
- 9 Huang D, Zhao C, Li R, Chen B, Zhang Y, Sun Z, et al.: Identification of a binding site on soluble RANKL that can be targeted to inhibit soluble RANK–RANKL interactions and treat osteoporosis, *Nat. Commun.* (2022), 13: 5817. <https://doi.org/10.1038/s41467-022-33006-4>.
- 10 Eberhardt J, Santos-Martins D, Tillack AF, & Forli S: AutoDock Vina 1.2.0: New docking methods, expanded force field, and Python bindings. *Journal of Chemical Information and Modeling* (2021), 61(8): 3891–3898. <https://doi.org/10.1021/acs.jcim.1c00203>.



- 11 Santos-Martins D, He Y, Eberhardt J, Sharma P, Bruciaferri N, Holcomb M, Llanos MA, Hansel-Harris AT, Barkdull AP, Tillack AF, Bianco G, Paulsen ML, Mato J, Taneja I, & Forli S: Meeko: Molecule Parametrization and Software Interoperability for Docking and Beyond. *Journal of Chemical Information and Modeling*, (2025), 65(24): 13045–13050. <https://doi.org/10.1021/acs.jcim.5c02271>.
- 12 Forli S, Huey R, Pique ME, Sanner MF, Goodsell DS, & Olson AJ: Computational protein–ligand docking and virtual drug screening with the AutoDock suite. *Nature Protocols*, (2016), 11: 905–919. <https://doi.org/10.1038/nprot.2016.051>.
- 13 Trott O, Olson AJ: AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J Comput Chem*. (2010), 31(2):455-461. <https://doi.org/10.1002/jcc.21334>.
- 14 Rogers D, Hahn M: Extended-connectivity fingerprints. *J Chem Inf Model*. (2010), 50(5): 742-754. <https://doi.org/10.1021/ci100050t>.
- 15 Landrum G: RDKit: Open-source cheminformatics and machine learning software. <https://www.rdkit.org> (accessed 2026).
- 16 Bickerton GR, Paolini GV, Besnard J, Muresan S, Hopkins AL: Quantifying the chemical beauty of drugs. *Nat Chem*. (2012), 4(2): 90-98. <https://doi.org/10.1038/nchem.1243>.
- 17 Yuzhe W, Li Y, Chen J, Lai L: Modeling protein-ligand interactions for drug discovery in the era of deep learning, *Chemical Society Reviews* (2025), 54(23): 11141–11183. <https://doi.org/10.1039/D5CS00415B>.
- 18 Breiman L: Random Forests, *Machine Learning*, (2001), 45: 5–32. <https://doi.org/10.1023/A:1010933404324>.
- 19 Tran-Nguyen VK, Camproux AC and Taboureau O: ClassyPose: A Machine-Learning Classification Model for Ligand Pose Selection Applied to Virtual Screening in Drug Discovery, *Advanced Intelligent Systems*, (2024), 6: 2400238. <https://doi.org/10.1002/aisy.202400238>.
- 20 George O, Ibomoye DM, Oluwaseun FE, Ikiomoye DE, Adeola O, Blessing O, Pere M, Kehinde A: Supervised machine learning in drug discovery and development: Algorithms, applications, challenges, and prospects, *Machine Learning with Applications*, (2024), 17:100576. <https://doi.org/10.1016/j.mlwa.2024.100576>.
- 21 Kralj S, Jukič M, Bren U: Molecular Filters in Medicinal Chemistry, *Encyclopedia*, (2023), 3(2):501–511. <https://doi.org/10.3390/encyclopedia3020035>.
- 22 Fu L, Shi S, Yi J, et al. ADMETlab 3.0: an updated comprehensive online ADMET prediction platform enhanced with broader coverage, improved performance, API functionality and decision support. *Nucleic Acids Research*, (2024), 52(W1): W422–W431. <https://doi.org/10.1093/nar/gkae236>.
- 23 Baell JB, Holloway GA: New substructure filters for removal of pan assay interference compounds (PAINS) from screening libraries and for their exclusion in bioassays. *J. Med. Chem*. (2010), 53: 2719–2740. <https://doi.org/10.1021/jm901137j>.
- 24 Ramírez-Márquez CD, Medina-Franco JL: KNIME Workflows for Chemoinformatic Characterization of Chemical Databases, *Molecular Informatics*, (2025), 44: e202400337. <https://doi.org/10.1002/minf.202400337>.
- 25 Daina A, Michielin O, Zoete V: SwissADME: a free web tool to evaluate pharmacokinetics, drug-likeness and medicinal chemistry

- friendliness of small molecules, *Scientific Reports*, (2017), 7:42717. <https://doi.org/10.1038/srep42717>.
- 26 Gu Y, Yu Z, Wang Y, et al.: admetSAR3.0: a comprehensive platform for exploration, prediction and optimization of chemical ADMET properties, *Nucleic Acids Research*, (2024), 52(W1): W432–W438. <https://doi.org/10.1093/nar/gkae298>.
- 27 Gallego RA, Edwards MP, Montgomery TP: An update on lipophilic efficiency as an important metric in drug design, *Expert Opinion on Drug Discovery*, (2024), 19: 917–931. <https://doi.org/10.1080/17460441.2024.2368744>.
- 28 Quancard J, et al.: The European Federation for Medicinal Chemistry and Chemical Biology (EFMC) Best Practice Initiative: Hit to Lead. *ChemMedChem*. (2025). <https://doi.org/10.1002/cmdc.202400931>.
- 29 Lo Y-C, et al.: A comprehensive comparison of molecular feature representations for use in predictive modeling. *Journal of Cheminformatics*. (2021), 13: 6. <https://doi.org/10.1186/s13321-020-00473-2>.
- 30 Boudjelal M, et al.: Random forest classification for predicting lifespan-extending chemical compounds. *Scientific Reports*. (2021), 11: 12673. <https://doi.org/10.1038/s41598-021-93070-6>.

HOW TO CITE: Chinmaya Chidananda Behera, Rohit Dash, Satya Prakash Panda, Sagar Kumar Mishra, Structure-Based Virtual Screening and AI-Assisted Prioritization of Small-Molecule RANKL-Binding Candidates for Rheumatoid Arthritis-Associated Bone Erosion, *Int. J. of Pharm. Sci.*, 2026, Vol 4, Issue 10, 1-13, <https://doi.org/10.5281/zenodo.23074259>

Table 1. Descriptive statistics for the 671-compound curated screening library.

Parameter	Mean $\pm$ SD	Median	Range
Vina score (kcal/mol)	-7.61 ± 0.30	-7.54	-8.742 to -7.229
Molecular weight (Da)	439.57 ± 53.23	432.37	299.20 to 636.54
LogP	0.43 ± 0.97	0.46	-2.20 to 6.27
TPSA ($\text{\AA}^2$)	96.26 ± 31.92	99.52	17.07 to 182.28
Rotatable bonds	4.87 ± 2.88	5	0 to 12
H-bond donors	1.78 ± 1.17	2	0 to 5
H-bond acceptors	4.97 ± 1.87	5	1 to 10
QED	0.508 ± 0.151	0.515	0.119 to 0.852
Heavy atom count	34.36 ± 4.02	34	23 to 50

Table 2. Top 20 molecular features ranked by Random-Forest mean decrease in impurity.

Rank	Feature	Importance
1	Ipc	0.02436
2	VSA_EState3	0.01745
3	MolLogP	0.01541
4	QED	0.01510
5	BCUT2D_MRLOW	0.01503
6	VSA_EState7	0.01435
7	MaxAbsEStateIndex	0.01429
8	VSA_EState4	0.01420
9	MinAbsEStateIndex	0.01402
10	VSA_EState2	0.01365



11	SPS	0.01273
12	MinEStateIndex	0.01268
13	PEOE_VSA9	0.01197
14	VSA_EState5	0.01180
15	EState_VSA2	0.01169
16	VSA_EState6	0.01153
17	PEOE_VSA1	0.01095
18	SMR_VSA1	0.01058
19	SlogP_VSA2	0.01019
20	VSA_EState1	0.00989

Table 3. Random-Forest validation performance for the docking-derived elite-class classifier.

Metric	Value	Evaluation set
Elite threshold	-7.7815 kcal/mol	Full curated dataset
Elite / non-elite	168 / 503	Full curated dataset
CV ROC-AUC	0.648 ± 0.062	5-fold CV; training partition
Accuracy	0.600	Held-out test set
Balanced accuracy	0.577	Held-out test set
Precision	0.321	Held-out test set
Recall	0.529	Held-out test set
F1-score	0.400	Held-out test set
ROC-AUC	0.617	Held-out test set
Confusion matrix	63 TN; 38 FP; 16 FN; 18 TP	Held-out test set

Table 4. Integrated docking, RF-prioritization and rule-based ADMET/structural-alert profile of ZINC70673359.

Parameter	Value
Ligand	ZINC70673359
Vina score	-8.130 kcal/mol
RF AI rank	1
RF elite probability	97.6% (final model refit on 671 structures)
RF predicted class	Elite
Molecular weight	549.413 Da
LogP	-1.138
TPSA	136.99 Å ²
H-bond donors / acceptors	3 / 6
Rotatable bonds	10
Heavy atom count	43
QED	0.234
Predicted GI absorption	Low (implemented RDKit heuristic)
Predicted BBB permeability	No (-) (implemented heuristic)
PAINS alerts	0 (Pass)
Brenk alerts	3 patterns: imine_1; oxygen-nitrogen single bond; triple bond
Indicative Ki	≈1,098 nM (pKi ≈ 5.96; theoretical)
Ligand efficiency	0.189 kcal mol ⁻¹ per heavy atom (theoretical)
Lipophilic ligand efficiency	≈7.10 (theoretical)

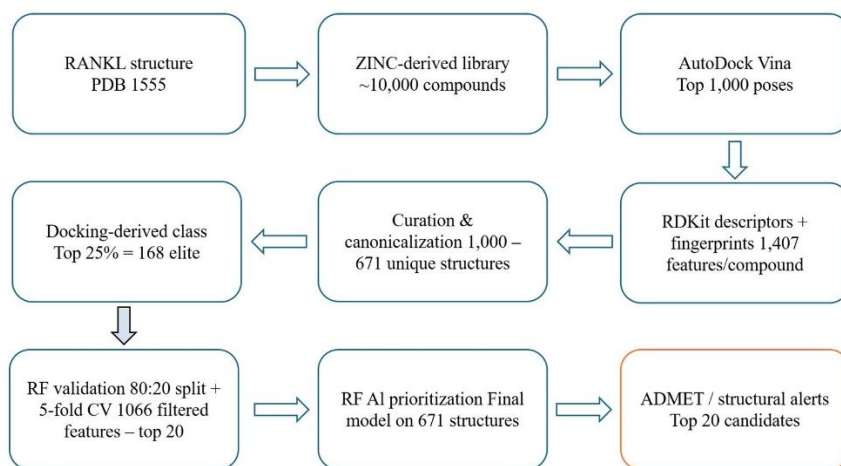


Fig. 1. Computational workflow used for AI-assisted prioritization of RANKL-binding candidates.

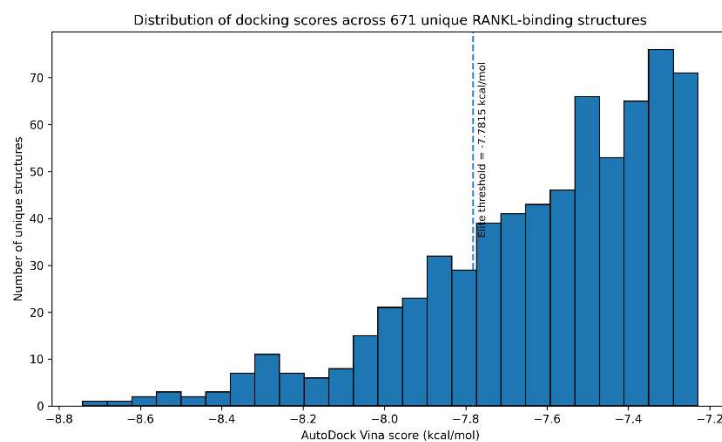


Fig. 2. Distribution of AutoDock Vina scores across the 671 curated unique structures. The dashed line marks the elite-class threshold (−7.7815 kcal/mol).

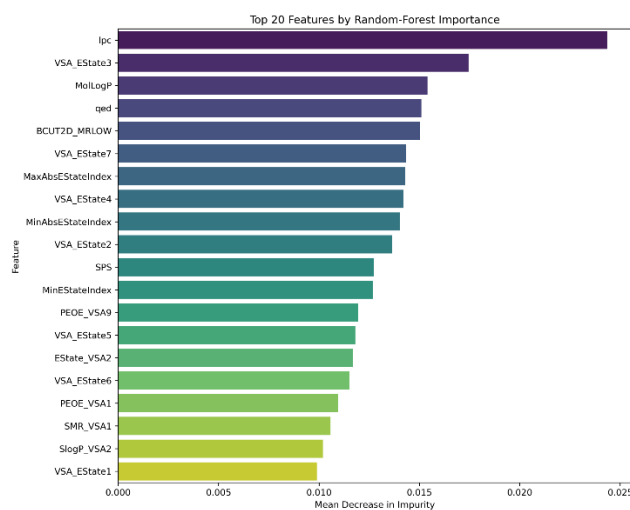


Fig. 3. Top 20 molecular features ranked by Random-Forest mean decrease in impurity.

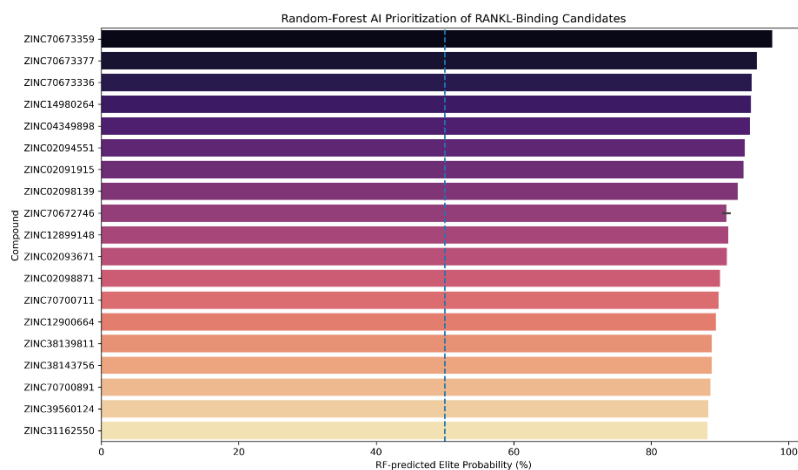


Fig. 4. RF-predicted elite probability for the top 20 prioritized RANKL-binding candidates.

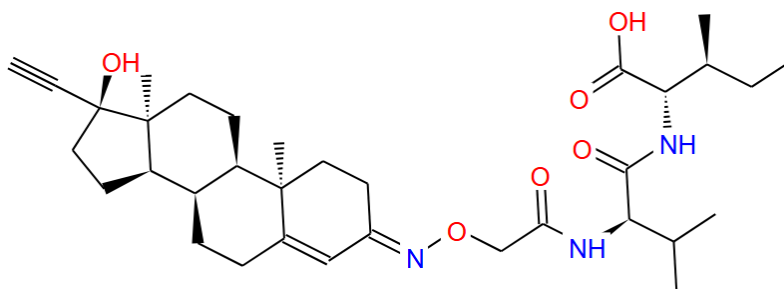


Fig. 5. Two-dimensional representation of RF-prioritized candidate ZINC70673359 based on the canonical SMILES used for analysis.

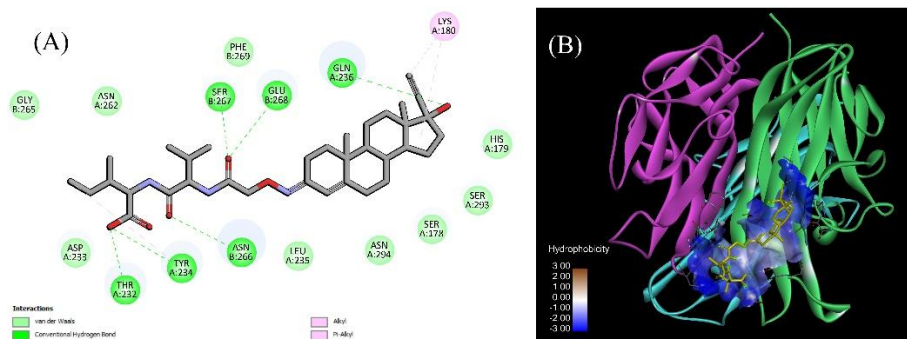


Fig. 6. Receptor – Ligand interactions (A)-2D, (B)-3D of ZINC70673359 with RANKL trimeric protein structure 1S55.