Supplementary Information

Gaps between medical biology and AI drug discovery

Yueming Yin[1], Afu Fu[1], and Lipo Wang[*1,2]

[1]Institute for Digital Molecular Analytics and Science, Nanyang Technological University, 59 Nanyang Drive, Singapore 636921
[2]School of Electrical and Electronic Engineering, Nanyang Technological University 639798, Singapore

Contents

[*]Corresponding author: elpwang@ntu.edu.sg

Supplementary Notes

1 Challenges in unified bioactivity modeling

The discrepancies in AI predictive models for bioactivity often stem from the diversity and complexity of assay measurement techniques. Different experimental methods, even when targeting the same bioactivity metric, introduce significant variability that current unified AI models fail to address. For example, federated multi-task learning in the pharmaceutical industry faces limitations, particularly in privacy-preserving settings, where underlying data heterogeneity and architectural choices reduce model performance.[1] They mitigated this by reducing task weights, reflecting the need to account for assay-specific variations.

Similarly, deep neural networks predicting protein-ligand binding affinities show no improvement when explicitly modeling noncovalent interactions compared to simpler molecular descriptors.[2] This counterintuitive result highlights a key issue: unified models obscure assay-specific nuances by aggregating disparate data. A single AI model attempting to generalize across such disparate data sources risks misinterpreting the underlying biological signals.

One recent work by[3] introduces an assay-aware bioactivity model that enriches proteochemometric (PCM) predictions by embedding free-text assay descriptions using BioBERT. They demonstrate that these embeddings cluster assays meaningfully and modestly improve R^2 in predicting ligand–protein interactions, from approximately 0.67 to 0.69. While this work marks an important step toward capturing assay context, it remains preliminary: the method relies solely on text embeddings and does not integrate structured metadata or mechanistic assay parameters. As a result, it only marginally enhances prediction and falls short of fully addressing the underlying biological variations across assays.

To address these challenges, future AI drug discovery methods must move beyond simplistic, unified models and incorporate assay-informed modeling, including assay method, experimental conditions, and measurement contexts. Integrating such metadata will allow models to account for the inherent variability in bioactivity measurements, leading to more robust and biologically meaningful predictions.

2 Potential of condition-value curves

Condition-value curves hold significant potential for improving bioactivity prediction due to their ability to capture dynamic relationships and provide a more comprehensive representation of experimental data. Pre- and post-condition readouts, such as the concentration-inhibition curve,[4] reflect the entire spectrum of response across varying conditions, such as concentrations or time points, rather than condensing complex interactions into single metrics like IC50 or EC50. This detailed data can reveal key aspects of molecular behavior, including dose-dependent effects,[5] non-linear responses,[6] and interaction thresholds,[7] which are often critical for understanding bioactivity but overlooked in standardized metrics.

For AI models, condition-value curves offer a richer feature space that enables learning nuanced patterns in molecular interactions. By incorporating the full trajectory of experimental responses, these curves can help models distinguish between molecules with similar standardized metrics but differing biological effects. For instance, molecules with comparable IC50 values might exhibit distinct inhibition profiles,[8] indicative of different mechanisms of action or off-target effects. Capturing this granularity allows AI models to make more accurate and physiologically relevant predictions.

Furthermore, condition-value curves align with the increasing complexity of modern drug discovery tasks, such as predicting combination therapy outcomes[9] or off-target toxicity[10], where interactions often exhibit emergent properties not evident in simplified readouts. By integrating this data, AI models can better account for experimental variability, such as differences in assay conditions or cellular contexts, leading to greater robustness and interpretability in predictions. This capability is particularly important in bridging the gap between *in vitro* and *in vivo* bioactivity assessments, where contextual factors significantly influence therapeutic efficacy.

3 Leveraging probabilistic insights

Probabilistic predictions handle missing data while enhancing model understanding. Unlike deterministic predictions,[11] which provide a single output for a given input, probabilistic methods[12] quantify the uncertainty in predictions and account for the range of possible outcomes. This is particularly advantageous in drug discovery, where experimental data is often incomplete, noisy, or subject to variability across measurement techniques and conditions. By explicitly modeling uncertainty, probabilistic predictions help AI models make more informed decisions and uncover patterns that might otherwise remain hidden.

For example, the value of probabilistic predictions can be seen by leveraging multiple molecular docking poses and a self-attention mechanism to infer binding sites and poses that are close to experimental data.[13] Instead of assuming a single "best" docking pose, the model explored a range of plausible configurations, improving its ability to identify interactions that align with underlying molecular mechanisms. This probabilistic approach effectively captured the complexity and variability of molecular interactions, uncovering insights into binding mechanisms that deterministic methods might overlook. Such strategies address concerns about the need for denser protein-ligand training matrices.[2] By incorporating probabilistic reasoning, AI models can approximate the missing complexity of experimental data and generate outputs that are not only accurate but also more interpretable.

Furthermore, probabilistic predictions enable AI models to account for experimental nuances, such as variability across assays or uncertainties in bioactivity measurements. For instance, Norinder et al.[14;15] incorporated conformal predictions[16] as additional features to improve bioactivity prediction accuracy. Conformal prediction frameworks provide confidence intervals or uncertainty bounds for each prediction, allowing models to better understand when and where their predictions may be unreliable. By integrating these uncertainty estimates, AI models can prioritize predictions with high confidence, optimizing the allocation of resources for experimental validation in resource-constrained scenarios.

The ability to model uncertainty also aligns well with the iterative nature of the drug discovery process. At early stages, where experimental data is sparse, probabilistic predictions can guide decisions on which data to acquire next, effectively reducing uncertainty[17] through active learning strategies. For example, when multiple molecular candidates exhibit similar predicted bioactivity, probabilistic insights can help identify those with the highest likelihood of success for further experimental validation. This strategy reduces investments in low-confidence predictions and accelerates the discovery of promising drug candidates, strengthening the foundation for addressing the complex challenges of modern drug discovery.

4 Challenges in AI interpretability

Beyond fragmented integration, a major challenge in AI-driven drug discovery is the interpretability of model predictions. While AI models have demonstrated high accuracy in predicting molecular properties,[18] biological activity,[13] and clinical outcomes,[19] understanding the underlying decision-making processes remains difficult. This lack of interpretability limits the ability of biologists and clinicians to validate AI-driven insights or design follow-up experiments. Recent efforts have introduced explainable AI (XAI) techniques to improve model transparency.[20]

Despite these advances, existing interpretability tools remain insufficient for addressing the complexity of biological systems and integrating expert domain knowledge. Many XAI techniques, such as saliency maps[21] and attention mechanisms[22], provide feature importance scores but lack biological grounding. This makes it challenging to determine whether a highlighted molecular substructure or pathway is mechanistically relevant or merely a statistical artifact. Additionally, most AI models operate at a single scale—such as molecular fingerprints in QSAR models[23] or protein-ligand binding scores[24]—without integrating information across multiple biological levels. Given that drug discovery spans molecular interactions, cellular networks, and whole-organism pharmacokinetics, bridging these scales remains an open challenge.

To enhance AI interpretability, several strategies can be adopted. Hybrid AI models that combine

mechanistic modeling (e.g., kinetic models[25] or pathway simulations[26]) with deep learning can improve biological relevance by incorporating known causal relationships. Knowledge graphs[27] can encode structured biological interactions, enabling AI models to reason beyond correlations. Causal discovery algorithms[28] and counterfactual explanations[29] can help distinguish causative factors from statistical associations, making AI-driven predictions more actionable for drug discovery.

Interpretability plays a vital role in building trust in AI systems and fosters collaboration across disciplines. When incorporated into AI frameworks, it enhances transparency, supports more effective hypothesis generation, and improves experimental validation strategies, contributing to greater success in AI-driven drug discovery. Expanding research in this area will further strengthen the integration of AI into biomedical science and keep pace with ongoing advances in the field.

5 Challenges after AI

Despite AI advances in hit identification and lead optimization, translating these successes into viable drug candidates remains challenging. Multi-objective molecular design methods[30;31] optimize both molecular activity and drug-likeness, and methods like leveraging gradients from molecular activity and property predictions to refine molecular designs.[32] While such approaches reduce the need for large-scale rescreening by balancing multiple drug properties, they fail to address the full complexity of downstream challenges, particularly in preclinical and clinical testing. Many drug candidates that successfully pass hit identification and lead optimization ultimately fail at these later stages.[33]

One major reason for these failures lies in the oversimplification of early-stage screening processes, which often neglect in situ bioactivity and physiological factors. AI models may optimize molecules for desirable drug-like properties under controlled conditions but fail to account for *in vivo* barriers such as bioavailability, pharmacokinetics, and tissue distribution. Molecules that perform well in early assays may fail to reach their targets effectively within the human body, limiting their therapeutic potential. For instance, Wu et al.[34] reviewed the bioavailability limitations of small-molecule drugs and highlighted strategies to address these barriers, such as prodrug approaches and nanoparticle delivery systems. These limitations highlight the need to include bioavailability and physiological constraints early in AI-driven drug design.

Additionally, conventional strategies often struggle to address "undruggable" targets—biomolecules that lack well-defined binding pockets or are inaccessible to small molecules. Recent advances have demonstrated promising strategies for tackling these targets, such as allosteric modulation, targeted protein degradation,[35] and exploiting transient binding sites.[36] While these approaches are still evolving, they highlight the potential for AI-driven methods to play a more significant role in addressing the complexity of undruggable targets. By incorporating insights from these emerging strategies, AI models can be further developed to identify novel binding mechanisms, predict transient interactions, and design molecules capable of modulating challenging targets. Expanding the scope of AI to focus on undruggable targets could unlock new therapeutic opportunities and overcome a critical bottleneck in drug discovery. For example, AI revolutionizes PROTAC technology for undruggable targets by streamlining target identification, ligand design, linker optimization, degradation prediction, specificity enhancement, and pharmacokinetic improvement.[35] It can re-analyze previous successful or unsuccessful drug candidates to predict new medicines that bind to targets and induce their degradation instead of merely inhibiting their activities. This dual capability—designing novel PROTACs and repurposing existing data—positions AI as a transformative force in developing therapies for previously intractable diseases.

6 Insufficient incorporation of physiological contexts

A significant gap in the current AI-driven drug discovery landscape is the insufficient incorporation of physiological contexts across the pipeline. Most AI models[37;38] are designed to optimize specific

molecular properties under controlled *in vitro* conditions but fail to simulate the dynamic and complex environments of living organisms. Critical factors such as immune system interactions,[39] tissue-specific delivery,[40] and spatial drug distribution[41] are often neglected, leading to failures in translating AI-optimized candidates into effective *in vivo* outcomes.

To address this limitation, the integration of AI simulations of physiological systems can serve as a promising direction for unifying the entire drug discovery pipeline. Physiologically based pharmacokinetic (PBPK)[42] models, for instance, can mathematically simulate drug absorption, distribution, metabolism, and excretion (ADME) within virtual human or animal systems. By coupling these models with AI-driven molecular design, it becomes possible to predict not only the efficacy of a candidate molecule but also its pharmacokinetics and potential toxicity in a specific biological context. Such simulations could inform early-stage decisions and prioritize molecules with higher probabilities of success in preclinical and clinical trials.

Furthermore, AI-based simulations of disease-specific physiological systems can improve the identification of therapeutic targets and the design of drugs tailored to specific patient populations. For example, virtual models of cancer microenvironments[43] or inflammatory pathways[44] could be used to predict how a drug interacts with its target in a biologically relevant context. This approach would enable a more holistic evaluation of drug candidates, reducing the reliance on oversimplified assays and improving the overall success rates of drug development.

Physiological context provides the critical link between molecular design and real-world efficacy, grounding AI applications in the complexity of biological systems. By simulating physiological processes, AI can unify the fragmented stages of drug discovery, enabling a more integrated and efficient pipeline from target identification to clinical implementation.

7 Features of the proposed biologically contextualized AI framework

- **Unify biology and AI:** Integrate structured biological knowledge into AI workflows, including tissue-specific responses, spatial transcriptomics, and pathway-level interactions. This allows models to interpret molecular features in the context of real biological systems, reducing misalignment between computational predictions and physiological outcomes.

- **Leverage non-traditional datasets:** Utilize underexploited data sources such as single-cell RNA sequencing, subcellular localization data, and protein-protein interaction networks. For example, the Tahoe-100M[45] dataset comprises 100 million cells and covers 60,000 conditions, 1,200 drug treatments, and 50 tumor models. A comprehensive dataset of protein-protein interactions and ligand binding pockets was proposed for advancing drug discovery.[46] It encompasses a diverse set of PPI complexes with more than 1,700 unique protein families. The RxRx3-core[47] dataset contains 6-channel Cell Painting images and associated embeddings from 222,601 wells. Trial-Bench[48] presents a comprehensive suite of 23 meticulously curated AI-ready datasets covering multi-modal input features and 8 crucial prediction challenges in clinical trial design, encompassing prediction of trial duration, patient dropout rate/event, serious adverse event, mortality event, trial approval outcome, trial failure reason, drug dose, and design of eligibility criteria. These datasets capture cellular heterogeneity and molecular interdependencies, enabling more precise and biologically meaningful predictions.

- **Introduce dynamic iterative AI loops:** Implement feedback loops where AI models are continuously refined based on experimental outcomes, including failures. This iterative approach mirrors real-world drug development, allowing models to learn from mistakes and progressively improve candidate selection.

- **Expand AI methodologies:** Move beyond standard deep learning to incorporate causal inference for mechanistic reasoning, hybrid graph neural networks for structural and relational understanding, and self-supervised learning to exploit unlabeled biological datasets. This broadens the AI toolkit for capturing complex, multi-scale biological phenomena.

- **Personalize drug discovery:** Integrate patient-specific information such as genomic variants, microbiome profiles, and immune system characteristics early in the pipeline. This supports stratified candidate selection and the design of drugs tailored to responder populations, improving both efficacy and safety.
- **Unlock non-classical targets:** Target challenging biomolecular interfaces, including transient allosteric sites, intrinsically disordered regions, and protein-protein interaction surfaces. AI can help identify druggable opportunities in regions traditionally considered intractable, expanding the landscape of potential therapeutics.
- **Ensure regulatory and ethical compliance:** Develop AI frameworks with regulatory foresight by emphasizing interpretability, robustness, and reproducibility. Ethical considerations, including data privacy, bias mitigation, and patient consent, are incorporated to ensure safe and responsible deployment in clinical contexts.

Supplementary Figures

Supplementary Figure 1: A biologically contextualized AI framework

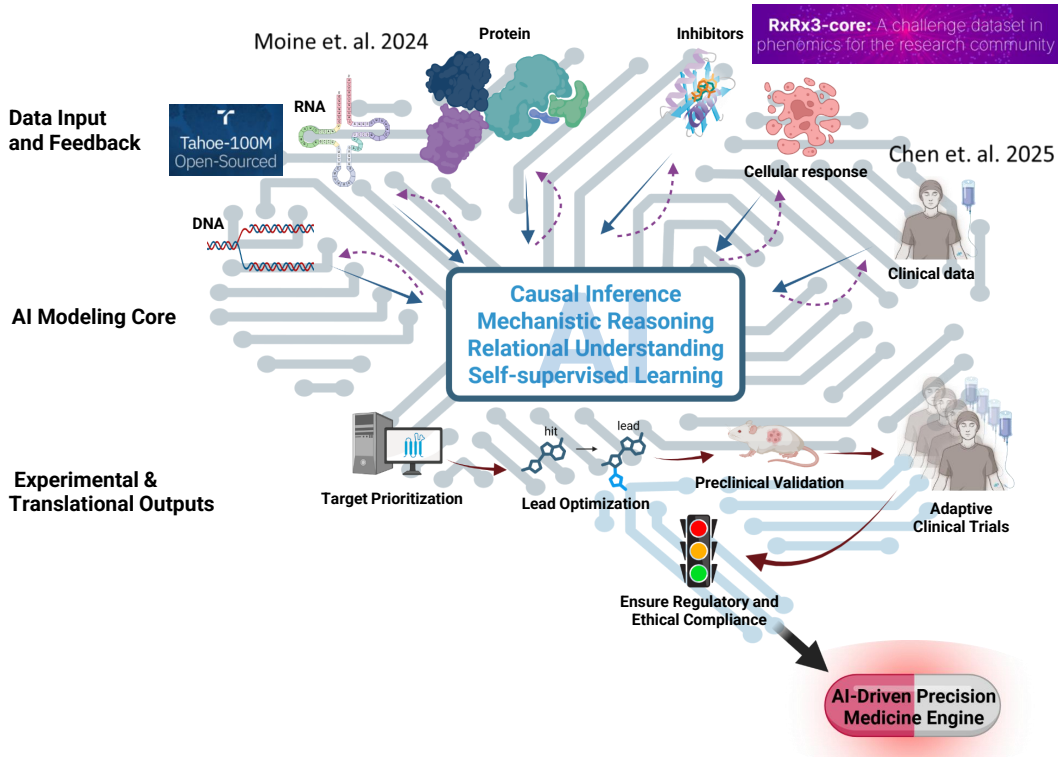

Figure 1: **A biologically contextualized AI framework** for drug discovery that integrates diverse biological data sources to optimize and accelerate therapeutic development. Blue arrows denote the incorporation of biological insights (derived from diverse and abundant datasets) into AI models, enhancing their capacity to identify targets and hits, optimize leads, and predict therapeutic outcomes. Red dotted arrows indicate AI feedback loops to refine input data. Red solid arrows mark the current fragmented applications that this framework aims to unify into a comprehensive and iterative AI modeling pipeline. Detailed framework features are described in Supplementary Note 7.

Supplementary References

References

[1] Humbeck Lina, Morawietz Tobias, Sturm Noe, et al. Don't overweight weights: Evaluation of weighting strategies for multi-task bioactivity classification models *Molecules.* 2021;26:6959.

[2] Volkov Mikhail, Turk Joseph-André, Drizard Nicolas, et al. On the frustration to predict binding affinities from protein–ligand structures with deep neural networks *Journal of medicinal chemistry.* 2022;65:7946–7958.

[3] Schoenmaker Linde, Sastrokarijo Enzo G, Heitman Laura H, Beltman Joost B, Jespers Willem, Westen Gerard JP. Toward Assay-Aware Bioactivity Model (er) s: Getting a Grip on Biological Context *Journal of Chemical Information and Modeling.* 2025.

[4] Dios Angeles, Mitchell Robert A, Aljabari Bayan, et al. Inhibition of MIF bioactivity by rational design of pharmacological inhibitors of MIF tautomerase activity *Journal of medicinal chemistry.* 2002;45:2410–2416.

[5] Zhang Hanrui, Wang Ziyan, Nan Yiyang, et al. Harmonizing across datasets to improve the transferability of drug combination prediction *Communications Biology.* 2023;6:397.

[6] Andersen Melvin E, Yang Raymond SH, French C Tenley, Chubb Laura S, Dennison James E. Molecular circuits, biological switches, and nonlinear dose-response relationships. *Environmental Health Perspectives.* 2002;110:971–978.

[7] Mukundan S, Deshpande Girish, Madhusudhan MS. High-affinity biomolecular interactions are modulated by low-affinity binders *NPJ Systems Biology and Applications.* 2024;10:85.

[8] Strelow John M. A perspective on the kinetics of covalent and irreversible inhibition *SLAS Discovery: Advancing Life Sciences R&D.* 2017;22:3–20.

[9] Rao Suman, Thibault Benoît, Peyrard Lisa, et al. Quantitative analysis of the potency of equimolar two-drug combinations and Combi-molecules involving kinase inhibitors in vitro: The concept of balanced targeting *International Journal of Molecular Sciences.* 2021;22:9569.

[10] Lin Ann, Giuliano Christopher J, Palladino Ann, et al. Off-target toxicity is a common mechanism of action of cancer drugs undergoing clinical trials *Science translational medicine.* 2019;11:eaaw8412.

[11] Tian Zhongda. A state-of-the-art review on wind power deterministic prediction *Wind Engineering.* 2021;45:1374–1392.

[12] Vershynin Roman. Four lectures on probabilistic methods for data science in *The mathematics of data*:231–271 2018.

[13] Yin Yueming, Lam Hilbert Yuen In, Mu Yuguang, Li Hoi Yeung, Kong Adams Wai-Kin. Advancing Bioactivity Prediction through Molecular Docking and Self-Attention *IEEE Journal of Biomedical and Health Informatics.* 2024.

[14] Norinder Ulf, Spjuth Ola, Svensson Fredrik. Using predicted bioactivity profiles to improve predictive modeling *Journal of chemical information and modeling.* 2020;60:2830–2837.

[15] Norinder Ulf, Spjuth Ola, Svensson Fredrik. Synergy conformal prediction applied to large-scale bioactivity datasets and in federated learning *Journal of Cheminformatics.* 2021;13:77.

[16] Fontana Matteo, Zeni Gianluca, Vantini Simone. Conformal prediction: a unified review of theory and new challenges *Bernoulli.* 2023;29:1–23.

[17] Yang Yi, Ma Zhigang, Nie Feiping, Chang Xiaojun, Hauptmann Alexander G. Multi-class active learning by uncertainty sampling with diversity maximization *International Journal of Computer Vision.* 2015;113:113–127.

[18] Zheng Yizhen, Koh Huan Yee, Ju Jiaxin, et al. Large language models for scientific discovery in molecular property prediction *Nature Machine Intelligence.* 2025:1–11.

[19] Khalifa Mohamed, Albadawy Mona. Artificial intelligence for clinical prediction: exploring key domains and essential functions *Computer Methods and Programs in Biomedicine Update.* 2024:100148.

[20] Urvi , Sharma Parul, Goyal Kanu, Sharma Shweta. Real-World Applications of Explainable AI in Healthcare *Explainable Artificial Intelligence in the Healthcare Industry.* 2025:451–466.

[21] Wang Conghao, Kumar Gaurav Asok, Rajapakse Jagath C. Drug discovery and mechanism prediction with explainable graph neural networks *Scientific Reports.* 2025;15:179.

[22] Yin Yueming, Hu Haifeng, Yang Zhen, Xu Huajian, Wu Jiansheng. Realvs: toward enhancing the precision of top hits in ligand-based virtual screening of drug leads from large compound databases *Journal of chemical information and modeling.* 2021;61:4924–4939.

[23] Myint Kyaw-Zeyar, Wang Lirong, Tong Qin, Xie Xiang-Qun. Molecular fingerprint-based artificial neural networks QSAR for ligand biological activity predictions *Molecular pharmaceutics.* 2012;9:2912–2923.

[24] Sim Jaemin, Kim Dongwoo, Kim Bomin, Choi Jieun, Lee Juyong. Recent advances in AI-driven protein-ligand interaction predictions *Current Opinion in Structural Biology.* 2025;92:103020.

[25] Claas Allison M, Lee Meelim, Huang Pai-Hsi, et al. Viral kinetics model of SARS-CoV-2 infection informs drug discovery, clinical dose, and regimen selection *Clinical Pharmacology & Therapeutics.* 2024;116:757–769.

[26] Hwang Woochang, Hwang Yongdeuk, Lee Sunjae, Lee Doheon. Rule-based multi-scale simulation for drug effect pathway analysis in *BMC medical informatics and decision making*;13:1–9Springer 2013.

[27] Meijer David, Beniddir Mehdi A, Coley Connor W, et al. Empowering natural product science with AI: leveraging multimodal data and knowledge graphs *Natural Product Reports.* 2025.

[28] Malinsky Daniel, Danks David. Causal discovery algorithms: A practical guide *Philosophy Compass.* 2018;13:e12470.

[29] Wellawatte Geemi P, Seshadri Aditi, White Andrew D. Model agnostic generation of counterfactual explanations for molecules *Chemical science.* 2022;13:3697–3705.

[30] Perron Quentin, Mirguet Olivier, Tajmouati Hamza, et al. Deep generative models for ligand-based de novo design applied to multi-parametric optimization *Journal of Computational Chemistry.* 2022;43:692–703.

[31] Tang Xiangru, Dai Howard, Knight Elizabeth, et al. A survey of generative AI for de novo drug design: new frontiers in molecule and protein generation *Briefings in Bioinformatics.* 2024;25:bbae338.

[32] Yin Yueming, Hu Haifeng, Yang Jitao, et al. OLB-AC: toward optimizing ligand bioactivities through deep graph learning and activity cliffs *Bioinformatics.* 2024;40:btae365.

[33] Hwang Thomas J, Carpenter Daniel, Lauffenburger Julie C, Wang Bo, Franklin Jessica M, Kesselheim Aaron S. Failure of investigational drugs in late-stage clinical development and publication of trial results *JAMA internal medicine.* 2016;176:1826–1833.

[34] Wu Ke, Kwon Soon Hwan, Zhou Xuhan, et al. Overcoming Challenges in Small-Molecule Drug Bioavailability: A Review of Key Factors and Approaches *International Journal of Molecular Sciences.* 2024;25:13121.

[35] Danishuddin , Jamal Mohammad Sarwar, Song Kyoung-Seob, Lee Keun-Woo, Kim Jong-Joo, Park Yeong-Min. Revolutionizing drug targeting strategies: integrating artificial intelligence and structure-based methods in PROTAC development *Pharmaceuticals.* 2023;16:1649.

[36] Xie Xin, Yu Tingting, Li Xiang, et al. Recent advances in targeting the "undruggable" proteins: from drug discovery to clinical trials *Signal transduction and targeted therapy.* 2023;8:335.

[37] Wei Yu, Li Shanshan, Li Zhonglin, Wan Ziwei, Lin Jianping. Interpretable-ADMET: a web service for ADMET prediction and optimization based on deep neural representation *Bioinformatics.* 2022;38:2863–2871.

[38] Yi Jia-Cai, Yang Zi-Yi, Zhao Wen-Tao, et al. ChemMORT: an automatic ADMET optimization platform using deep learning and multi-objective particle swarm optimization *Briefings in Bioinformatics.* 2024;25:bbae008.

[39] Pichler Werner J, Beeler Andreas, Keller Monika, et al. Pharmacological interaction of drugs with immune receptors: the pi concept *Allergology International.* 2006;55:17–25.

[40] Zhao Zongmin, Ukidve Anvay, Kim Jayoung, Mitragotri Samir. Targeting strategies for tissue-specific drug delivery *Cell.* 2020;181:151–167.

[41] Vendel Esmée, Rottschäfer Vivi, Lange Elizabeth CM. The need for mathematical modelling of spatial drug distribution within the brain *Fluids and Barriers of the CNS.* 2019;16:1–33.

[42] Espié Pascal, Tytgat Dominique, Sargentini-Maier Maria-Laura, Poggesi Italo, Watelet Jean-Baptiste. Physiologically based pharmacokinetics (PBPK) *Drug metabolism reviews.* 2009;41:391–407.

[43] Wu Mingming, Swartz Melody A. Modeling tumor microenvironments in vitro *Journal of biomechanical engineering.* 2014;136:021011.

[44] Pinton Philippe. Computational models in inflammatory bowel disease *Clinical and Translational Science.* 2022;15:824–830.

[45] Zhang Jesse, Ubas Airol A, Borja Richard, et al. Tahoe-100M: A Giga-Scale Single-Cell Perturbation Atlas for Context-Dependent Gene Function and Cellular Modeling *bioRxiv.* 2025:2025–02.

[46] Moine-Franel Alexandra, Mareuil Fabien, Nilges Michael, Ciambur Constantin Bogdan, Sperandio Olivier. A comprehensive dataset of protein-protein interactions and ligand binding pockets for advancing drug discovery *Scientific Data.* 2024;11:402.

[47] Kraus Oren, Comitani Federico, Kenyon-Dean Kian, et al. RxRx3-core: Benchmarking drug-target interactions in high-content microscopy in *Learning Meaningful Representations of Life (LMRL) Workshop at ICLR 2025* 2025.

[48] Chen Jintai, Hu Yaojun, Cai Mingchen, et al. TrialBench: Multi-modal AI-ready datasets for clinical trial prediction *Scientific Data.* 2025;12:1564.