Current Status of Computational Approaches for Small Molecule Drug Discovery

Weijun Xu

Experimental Drug Development Centre, 10 Biopolis Road, #05-01, Chromos, Singapore, 138670

Correspondence: xu_weijun@eddc.sg

2024 has been an exciting year for computational sciences for its recognition with the Nobel prize in physics awarded to “artificial neural network” and Nobel prize in chemistry presented to “protein structure prediction and design”. Given the rapid advancements in Computer-Aided Drug Design (CADD) and Artificial Intelligence in Drug Discovery (AIDD), a document summarizing their current standing and future directions would be timely and relevant to the readership of *Journal of Medicinal Chemistry*. This piece of commentary aims to highlight recent developments, key challenges, and potential synergies between these fields, contributing to ongoing discussions in the literature and scientific blogs.

Computer-Aided Drug Design (CADD) and Artificial Intelligence Drug Discovery (AIDD) have made significant advancements in recent years. These fields leverage physics-based computational methods and machine learning to enhance the efficiency and speed of drug design with an aim to revolutionize the way new therapeutics are discovered and optimized. Since the born the first molecular docking software DOCK[1], CADD has evolved considerably, with improvements in force field development and conformational sampling methods for both small molecules and macromolecules. New algorithms for molecular docking and scoring have improved the prediction of ligand-receptor interactions, increasing hit rates in virtual screening. The integration of pharmacophore modeling with quantitative structure-activity relationship (QSAR) methods has led to more robust predictive models. Platforms like Schrödinger[2], Molecular Operating Environment (MOE)[3], and OpenEye Scientific[4] have enhanced user interfaces and computational speed on cloud and interactive user modes, making them more accessible to researchers who are beyond the expertise of computational chemistry. Structure-based virtual screening has become a "state-of-the-art" tool to identify starting chemotypes for early discovery. The increasing space of "Make-on-demand" chemical libraries will enable the *in silico* screening of up to trillion compounds within weeks in the coming years, providing unprecedented opportunity to novel hit scaffold exploration. Moreover, accurate protein structure prediction by deep learning will increasingly contribute to successful applications of virtual screens to novel biological targets, for which ligand discovery through machine learning is unlikely to succeed due to the sparce chemical information available for the target. Over the past 5 years, virtual screening of compound libraries broadly covering large chemical space has been applied extensively in early drug discovery pipeline. Lyu et al used docked 138 million compounds in the dopamine D4 receptor and 99 million compounds in the AmpC β-lactamase, resulting in a diverse range of novel hits for the two targets[5]. Gorgulla et al docked 1.3 billion compounds in the KEAP1 redox sensor and confirmed a hit rate of 12% binders by surface plasmon resonance (SPR)[6]. Sadybekov et al proposed a "virtual synthon hierarchical enumeration screening" (V-SYNTHES) approach that performs docking calculations on a focused set of fragment compounds that are representative of all the scaffolds available for library synthesis[7]. Recursion successfully virtually screened nearly 15,000 human proteins containing more than 80,000 potential binding pockets using the Enamine REAL Space[8]. Meanwhile, the awareness of traditional modelling relying heavily on docking has called for the need of development of more accurate physics-based predictive approaches such as free energy perturbations (FEP) and thermo integration (TI). In recent years, relative binding free energy (RBFE) has been extensively validated for its power during lead optimization in a

wide range targets including protease, kinases and GPCRs[9]. Enhanced sampling molecular dynamics have been developed to detect cryptic pockets and predict ligand binding kinetics[10], an important parameter that correlates in vivo efficacy[11]. Thanks to the advancement of computational methods, availability of high-performance computing as well as GPU accelerated simulations, successful application of CADD workflow (Figure 1A) has led to a few clinical candidates. Nimbus partnered with Schrödinger and employed structure-based drug design strategy to drive the discovery of potentially best-in-class TYK2 inhibitor[12], Morphic Therapeutic leveraged "digital chemistry" - FEP to design a novel small molecule inhibitor of α4β7 integrin[13]. Relay therapeutics performed long time MD simulations and designed orthosteric binders that selectively and covalently engage the Cys residue from the "P-loop" of FGFR2[14].

In the last decade, AIDD has gained a surge of momentum with the rise of machine learning and deep learning applications in drug discovery. Intriguingly, the term "AIDD" obtained high "quotes" in literature, public speech, and media. However, it is imperative to understand the underlying difference between CADD and AIDD for a better understanding of the best application domain that each may contribute towards a drug discovery project. AIDD utilizes large datasets from existing public or proprietary data repository, enabling the identification of new predictions through pattern recognition and pre-trained machine learning models. A revolutionary contribution is the proteome-wide prediction of the 3-dimensional structure by Google DeepMind's Alphafold platform, enabling structure-based virtual ligand discovery in a scale significantly superior to that from experimental structures[15]. On the other hand, the development and application of generative models, such as variational autoencoders (VAE)[16] and generative adversarial networks (GAN)[17], Chemical Language Model[18], reinforcement learning[19], transformer[20], and diffusion models[21] has allowed the design of molecular structures with desired biological and physicochemical properties based on the curation of a training set from which the machine learning algorithm can learn and generate new structures as hypotheses for experimental validation (Figure 1B). Benchmarking datasets such as MOSES and GuacaMol have been published for the purpose of comparing and validating generative models[21, 22]. In 2019, InsilicoMedicine published a seminal paper describing a breakthrough discovery of DDR1 inhibitor, leveraging on a deep generative tensorial reinforcement learning model (GENTRL) for de novo small-molecule design in only 21 days[23]. TNIK is proposed as a novel target for Idiopathic Pulmonary Fibrosis (IPF) by Insilico Medicine's AI target discovery engine platform "PandaOmics"[24]. Leveraging on the company's generative chemistry tool "Chemistry42", novel structure generation and medicinal

chemistry effort led to the discovery and development of INS018_055 currently in a phase II trial. After sitting on the fence for almost a decade, we also see the "wind down" of several candidates from AI companies: Exscientia-21546, a highly potent and selectively A2AR antagonist, prevented its further advancement at Phase I/II study; BenevolentAI terminated BEN-2293, a pan-Trk inhibitor for atopic dermatitis in a phase 2 trial. Since 2012, the incubation and financial seeding of more than a dozen of AI startups and biotech have gone through a decade of AI racing (Figure 1C). From the dealmaking aspect, acceleration of the discovery for both target and molecular entity has attracted significant partnerships between large biopharma and AI-focused biotechs[25]. Undoubtful, AIDD has become a mainstream in the pharmaceutical industry (Table 1), yet, before the wave of "AI-made" drugs reach to the market, several questions remain to be addressed: **First**, can we get high quality data for AI model to learn meaningful patterns and make predictions accordingly? **Second**, can AI really understand biology, which is complicated enough for its own science domain. This fundamental question relates to a debatable point on whether AI can discover a truly novel target for a disease indication? **Third**, can generative chemistry truly design an active molecule out of the training space. De novo generation of molecules based on protein binding pockets will likely address the caveat from the molecular generation based on ligands only as the training set[26].

Despite emerging advancements and successful stories, both CADD and AIDD face challenges. Affordable approaches to academy and start-up/biotechs to rank-order virtual screening hits from ultra-large screening campaign need to be developed to enable an economic way for prioritizing virtual hits that can be predicted with accuracy between that from docking scores and FEP calculations. Recently, Wu *et al* proposed treatment of solvation terms and GBMV method to eliminate "cheating molecules" from virtual screening[27]. We anticipate other methodologies that can address these gaps in the field. CADD or AIDD has showncase in delivering novel molecules (e.g. FIC or BIC) for targets that are generally amenable to computational approaches (e.g. high-quality structure with good druggability, well studied biology, training set molecules available). However, when it comes to the scenario that nearly nothing is known about the target – (a highly novel target), the applicability and success rate of applying machine learning for hit discovery, hit to lead and lead optimization remains enigmatic. This is exactly the case for a drug discovery program involving modalities beyond the traditional small molecule inhibitors. For instance, predicting the 3D structure of RNA is a significant challenge despite ongoing advancements in the field. Although AlphaFold3 has

shown break-through predictive performance for proteins, accurate structural prediction for RNAs remains challenging due to fundamental differences between proteins and RNAs and the sparce structural data as training set[28]. In fact, given the nature of RNAs, small molecules targeting RNA *per se* may be intractable due to the highly polar and dynamic feature of the RNAs. Scoring functions that are suitable for small molecule RNA docking are yet to be improved[29]. Molecular glues including heterobifunctional molecules (PROTAC) and monovalent degrader (intrinsic degrader) have gained increasing attention for small molecule drug discovery over the past 5 years. Cherkasov group developed an integrative computational pipeline of 3-D modeling and deep learning for the automation of *de novo* design of PROTACs[30]. Monte Rosa built the AI algorithms named fAIceit™, an ultra-fast engine to scan through thousands of proteins to determine those with structural features called degrons that allow an E3 ligase to bind to and degrade them[31]. It remains to be seen whether generative AI may deliver promising small molecules targeting RNA and molecular glues in the near future.

Looking forward, the future of CADD and AIDD appears promising, with a potential focus on complementing physics-based methods with AIDD techniques to leverage the strengths of both domains while mitigating caveats stemming from these computational approaches. Interestingly, the Nobel Prize in Chemistry 2013 was awarded to Martin Karplus, Michael Levitt and Arieh Warshel for the development of multiscale models for complex chemical systems. In 2024, Geoffrey Hinton, the 'godfather of AI', and John Hopfield co-honoured for their work on artificial neural networks and won the Nobel prize in physics. Meantime, the Nobel prize in chemistry was awarded to David Baker, Demis Hassabis and John Jumper for their contribution towards protein design and prediction of complex protein structure. What a big year for AI in Science! In the coming years, we will witness the wave of "match-making" CADD and AIDD, offering unprecedented opportunities to accelerate drug discovery and development. As technology continues to evolve, the integration of these computational tools into traditional research paradigms holds the promise of transforming the pharmaceutical landscape. Continued investment in research, collaboration, and addressing existing challenges will be the key to unlocking their full potential of computational approaches for small molecule drug discovery.

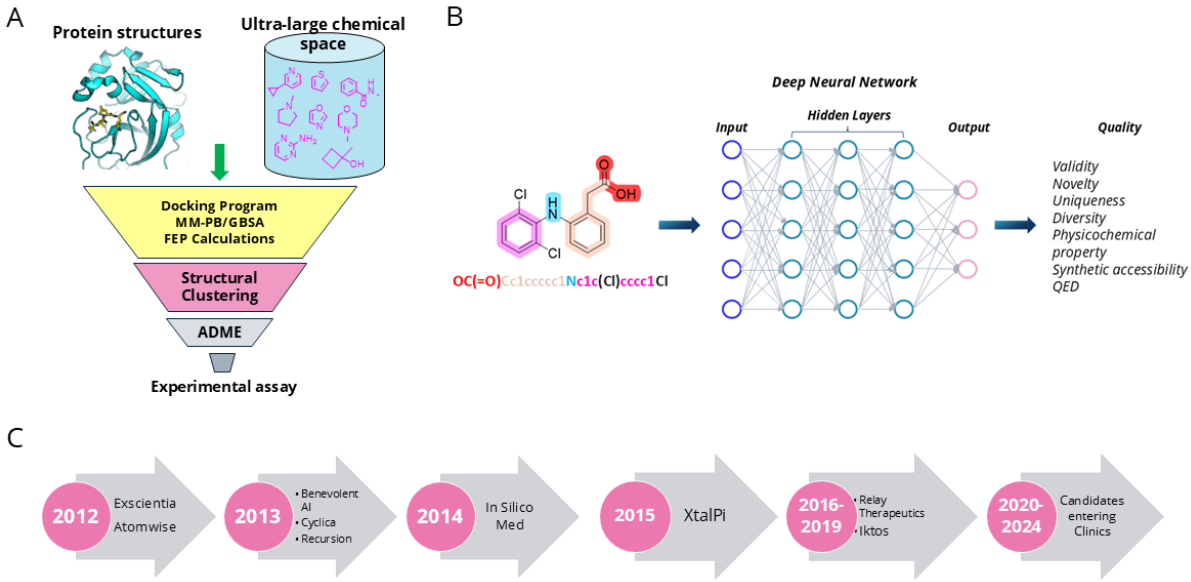

Figure 1. (A) General workflow of CADD. (B) Molecular generation by deep learning architecture. (C) Examples of the evolving AI companies in the past decade.

Table 1. Selected examples of AI-discovered/designed small molecules in clinical stage. The information was collected from website of BioPharmaTrend and respective company's website.

Small Molecule	Company	Target	Stage	Indication
REC-2282	Recursion	HDAC	Phase 2/3	Neurofibromatosis Type 2
REC-994	Recursion	Superoxide	Phase 2	Cerebral Cavernous Malformation
REC-4881	Recursion	MEK	Phase 2	Familial Adenomatous Polyposis
INS018_055	Insilico Medicine	TNIK	Phase 2	IPF
ISM3091	Insilico Medicine	USP1	Phase 1	BRCA-mutant cancer
RLY-4008	Relay Therapeutics	FGFR2	Phase1/2	FGFR2-altered cholangiocarcinoma
GTAEXS617	Exscentia	CDK7	Phase1/2	solid tumors
EXS4318	Exscentia	PKC-theta	Phase 1	Inflammatory and immunologic diseases
BEN-8744	BenevolentAI	PDE10	Phase 1	Ulcerative Colitis

Acknowledgements: We thank Dr. Hao Fan from Bioinformatics Institute, A*STAR of Singapore for insightful discussion.

References

(1) Shoichet, B. K.; Kuntz, I. D. Protein docking and complementarity. *J Mol Biol* **1991**, *221* (1), 327-346. DOI: 10.1016/0022-2836(91)80222-g From NLM Medline.
(2) Schrödinger, LLC, New York, NY, https://www.schrodinger.com/.
(3) Molecular Operating Environment,https://www.chemcomp.com/en/Products.htm.
(4) OpenEye Scientific, https://www.eyesopen.com/.
(5) Lyu, J.; Wang, S.; Balius, T. E.; Singh, I.; Levit, A.; Moroz, Y. S.; O'Meara, M. J.; Che, T.; Algaa, E.; Tolmachova, K.; et al. Ultra-large library docking for discovering new chemotypes. *Nature* **2019**, *566* (7743), 224-229. DOI: 10.1038/s41586-019-0917-9 From NLM Medline.
(6) Gorgulla, C.; Boeszoermenyi, A.; Wang, Z. F.; Fischer, P. D.; Coote, P. W.; Padmanabha Das, K. M.; Malets, Y. S.; Radchenko, D. S.; Moroz, Y. S.; Scott, D. A.; et al. An open-source drug discovery platform enables ultra-large virtual screens. *Nature* **2020**, *580* (7805), 663-668. DOI: 10.1038/s41586-020-2117-z From NLM Medline.
(7) Sadybekov, A. A.; Sadybekov, A. V.; Liu, Y.; Iliopoulos-Tsoutsouvas, C.; Huang, X. P.; Pickett, J.; Houser, B.; Patel, N.; Tran, N. K.; Tong, F.; et al. Synthon-based ligand discovery in virtual libraries of over 11 billion compounds. *Nature* **2022**, *601* (7893), 452-459. DOI: 10.1038/s41586-021-04220-9 From NLM Medline.
(8) https://ir.recursion.com/news-releases/news-release-details/recursion-bridges-protein-and-chemical-space-massive-protein.
(9) Wang, L.; Wu, Y.; Deng, Y.; Kim, B.; Pierce, L.; Krilov, G.; Lupyan, D.; Robinson, S.; Dahlgren, M. K.; Greenwood, J.; et al. Accurate and reliable prediction of relative ligand binding potency in prospective drug discovery by way of a modern free-energy calculation protocol and force field. *J Am Chem Soc* **2015**, *137* (7), 2695-2703. DOI: 10.1021/ja512751q From NLM Medline.
(10) Conflitti, P.; Raniolo, S.; Limongelli, V. Perspectives on Ligand/Protein Binding Kinetics Simulations: Force Fields, Machine Learning, Sampling, and User-Friendliness. *J Chem Theory Comput* **2023**, *19* (18), 6047-6061. DOI: 10.1021/acs.jctc.3c00641 From NLM Medline. Wang, J.; Arantes, P. R.; Bhattarai, A.; Hsu, R. V.; Pawnikar, S.; Huang, Y. M.; Palermo, G.; Miao, Y. Gaussian accelerated molecular dynamics (GaMD): principles and applications. *Wiley Interdiscip Rev Comput Mol Sci* **2021**, *11* (5). DOI: 10.1002/wcms.1521 From NLM PubMed-not-MEDLINE.
(11) Lee, K. S. S.; Yang, J.; Niu, J.; Ng, C. J.; Wagner, K. M.; Dong, H.; Kodani, S. D.; Wan, D.; Morisseau, C.; Hammock, B. D. Drug-Target Residence Time Affects in Vivo Target Occupancy through Multiple Pathways. *ACS Cent Sci* **2019**, *5* (9), 1614-1624. DOI: 10.1021/acscentsci.9b00770 From NLM PubMed-not-MEDLINE.
(12) Leit, S.; Greenwood, J.; Carriero, S.; Mondal, S.; Abel, R.; Ashwell, M.; Blanchette, H.; Boyles, N. A.; Cartwright, M.; Collis, A.; et al. Discovery of a Potent and Selective Tyrosine Kinase 2 Inhibitor: TAK-279. *J Med Chem* **2023**, *66* (15), 10473-10496. DOI: 10.1021/acs.jmedchem.3c00600 From NLM Medline.
(13) https://www.schrodinger.com/life-science/learn/case-studies/morphic-therapeutic-leverages-digital-chemistry-strategy/. (accessed.
(14) Subbiah, V.; Sahai, V.; Maglic, D.; Bruderek, K.; Toure, B. B.; Zhao, S.; Valverde, R.; O'Hearn, P. J.; Moustakas, D. T.; Schonherr, H.; et al. RLY-4008, the First Highly Selective FGFR2 Inhibitor with Activity across FGFR2 Alterations and Resistance Mutations. *Cancer Discov* **2023**, *13* (9), 2012-2031. DOI: 10.1158/2159-8290.CD-23-0475 From NLM Medline.

(15) Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Zidek, A.; Potapenko, A.; et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **2021**, *596* (7873), 583-589. DOI: 10.1038/s41586-021-03819-2 From NLM Medline.
(16) Lim, J.; Ryu, S.; Kim, J. W.; Kim, W. Y. Molecular generative model based on conditional variational autoencoder for de novo molecular design. *J Cheminform* **2018**, *10* (1), 31. DOI: 10.1186/s13321-018-0286-7 From NLM PubMed-not-MEDLINE.
(17) Goodfellow, I. P.-A., J.; Mirza, M.; Xu, B.; Warde-; Farley, D. O., S.; Courville, A.; Bengio, Y. . Generative AdversarialNetworks. *Commun. ACM* **2020**, *63* (111), 139–144.

(18) Moret, M.; Pachon Angona, I.; Cotos, L.; Yan, S.; Atz, K.; Brunner, C.; Baumgartner, M.; Grisoni, F.; Schneider, G. Leveraging molecular structure and bioactivity with chemical language models for de novo drug design. *Nat Commun* **2023**, *14* (1), 114. DOI: 10.1038/s41467-022-35692-6 From NLM Medline.
(19) Dodds, M.; Guo, J.; Lohr, T.; Tibo, A.; Engkvist, O.; Janet, J. P. Sample efficient reinforcement learning with active learning for molecular design. *Chem Sci* **2024**, *15* (11), 4146-4160. DOI: 10.1039/d3sc04653b From NLM PubMed-not-MEDLINE.
(20) Bagal, V.; Aggarwal, R.; Vinod, P. K.; Priyakumar, U. D. MolGPT: Molecular Generation Using a Transformer-Decoder Model. *J Chem Inf Model* **2022**, *62* (9), 2064-2076. DOI: 10.1021/acs.jcim.1c00600 From NLM Medline.
(21) Brown, N.; Fiscato, M.; Segler, M. H. S.; Vaucher, A. C. GuacaMol: Benchmarking Models for de Novo Molecular Design. *J Chem Inf Model* **2019**, *59* (3), 1096-1108. DOI: 10.1021/acs.jcim.8b00839 From NLM Medline.
(22) Polykovskiy, D.; Zhebrak, A.; Sanchez-Lengeling, B.; Golovanov, S.; Tatanov, O.; Belyaev, S.; Kurbanov, R.; Artamonov, A.; Aladinskiy, V.; Veselov, M.; et al. Molecular Sets (MOSES): A Benchmarking Platform for Molecular Generation Models. *Front Pharmacol* **2020**, *11*, 565644. DOI: 10.3389/fphar.2020.565644 From NLM PubMed-not-MEDLINE.
(23) Zhavoronkov, A.; Ivanenkov, Y. A.; Aliper, A.; Veselov, M. S.; Aladinskiy, V. A.; Aladinskaya, A. V.; Terentiev, V. A.; Polykovskiy, D. A.; Kuznetsov, M. D.; Asadulaev, A.; et al. Deep learning enables rapid identification of potent DDR1 kinase inhibitors. *Nat Biotechnol* **2019**, *37* (9), 1038-1040. DOI: 10.1038/s41587-019-0224-x From NLM Medline.
(24) Ren, F.; Aliper, A.; Chen, J.; Zhao, H.; Rao, S.; Kuppe, C.; Ozerov, I. V.; Zhang, M.; Witte, K.; Kruse, C.; et al. A small-molecule TNIK inhibitor targets fibrosis in preclinical and clinical models. *Nat Biotechnol* **2024**. DOI: 10.1038/s41587-024-02143-0 From NLM Publisher.
(25) Dealmakers, B. Generative AI platforms drive drug discovery dealmaking, https://www.nature.com/articles/d43747-024-00084-w. **2024**.
(26) Huang, L.; Xu, T.; Yu, Y.; Zhao, P.; Chen, X.; Han, J.; Xie, Z.; Li, H.; Zhong, W.; Wong, K. C.; Zhang, H. A dual diffusion model enables 3D molecule generation and lead optimization based on target pockets. *Nat Commun* **2024**, *15* (1), 2657. DOI: 10.1038/s41467-024-46569-1 From NLM Medline. Atz, K.; Cotos, L.; Isert, C.; Hakansson, M.; Focht, D.; Hilleke, M.; Nippa, D. F.; Iff, M.; Ledergerber, J.; Schiebroek, C. C. G.; et al. Prospective de novo drug design with deep interactome learning. *Nat Commun* **2024**, *15* (1), 3408. DOI: 10.1038/s41467-024-47613-w From NLM Medline.
(27) Wu, Y.; Liu, F.; Glenn, I.; Fonseca-Valencia, K.; Paris, L.; Xiong, Y.; Jerome, S. V.; Brooks, C. L., 3rd; Shoichet, B. K. Identifying Artifacts from Large Library Docking. *J Med*

Chem **2024**, *67* (18), 16796-16806. DOI: 10.1021/acs.jmedchem.4c01632 From NLM Medline.
(28) Bernard C, P. G., Ghannay S, Tahi F, Has AlphaFold 3 reached its success for RNAs? https://www.biorxiv.org/content/10.1101/2024.06.13.598780v1.
(29) Gunasinghe, K. K. J.; Ginjom, I. R. H.; San, H. S.; Rahman, T.; Wezen, X. C. Can Current Molecular Docking Methods Accurately Predict RNA Inhibitors? *J Chem Inf Model* **2024**, *64* (15), 5954-5963. DOI: 10.1021/acs.jcim.4c00235 From NLM Medline.
(30) Mslati, H.; Gentile, F.; Pandey, M.; Ban, F.; Cherkasov, A. PROTACable Is an Integrative Computational Pipeline of 3-D Modeling and Deep Learning To Automate the De Novo Design of PROTACs. *J Chem Inf Model* **2024**, *64* (8), 3034-3046. DOI: 10.1021/acs.jcim.3c01878 From NLM Medline.
(31) https://www.linkedin.com/pulse/how-monte-rosa-therapeutics-integrates-ai-big-pljcf/.