Journal of Cheminformatics

COMMENT

Open Access

AI/ML methodologies and the future-will they be successful in designing the next generation of new chemical entities?

Rachelle J. Bienstock[1*]

Abstract

Cheminformatics and chemical databases are essential to drug discovery. However, machine learning (ML) and artificial intelligence (AI) methodologies are changing the way in which chemical data is used. How will the use of chemical data change in drug discovery moving forward? How do the new ML methods in molecular property prediction, hit and lead and target identification and structure prediction differ and compare with previous computational methods? Will new ML methodologies improve chemical diversity in ligand design, and offer computational enhancements. There are still many advantages to physics based methods and they offer something lacking in ML/ AI based methods. Additionally, ML training methods often give the best results when experimental assay measurements are fed back into the model. Often modeling and experimental methods are not diametrically opposed but offer the greatest advantage when used complementary.

Keywords Alphafold, Machine learning (ML), Drug design, Virtual screening, Generative design, Graph neural networks, Docking, Artificial intelligence (AI)

Introduction

If you survey symposium topics presented at the Division of Chemical Information (CINF) over the past several years at an American Chemical Society (ACS) meeting you will notice a significant change. Topics selected often highlight the "hot" and current topics of interest in the application of cheminformatics methods to drug discovery. If we review the proposed symposia for the upcoming meeting, we will note topics such as "Machine Learning for Molecular Simulation and Design ", "Machine Learning and AI for Organic Chemistry", "Ethical issues of AI" and "Generative Modeling for Chemistry Biology and Material Discovery". Searching back only a few years ago in an ACS CINF program you would not find this focus on ML and AI methodologies. How are ML techniques having real and significant impact on cheminformatics, and dealing with chemical data in the drug discovery space? How much is "hype" versus significant improvement in identifying new chemical entities and exploring chemical space more efficiently with greater diversity?

AI/ML techniques have impacted the field of cheminformatics significantly, particularly with respect to applications and discussions within the fields of drug discovery. Molecular Databases and the representation, format, and treatment of small molecules for virtual high throughput screens, docking, modeling protein–ligand interactions, have changed and will continue to evolve with developments in the fields of graph neural networks, generative chemistry and alternative molecular representations [1] compatible with AI and ML methods. ML algorithms have impacted molecule property prediction, database searching, training datasets, and have brought about new methodologies such as active learning FEP; combined QSAR and FEP in cyclic active

*Correspondence:
Rachelle J. Bienstock
rachelleb1@gmail.com
[1] RJB Computational Modeling LLC, Chapel Hill, NC, USA

learning workflows; AI workflows in mining data sources; augmenting AI in structure based drug design by feeding back scoring in AI workflows and data imputation. As of Spring 2024, over 70 Investigational New Drug (IND) Applications have been filed with the Food and Drug Administration FDA which involve new chemical entities identified using AI/ML methods [2].

Discussion

In 2020 the BBC News ran a story "Artificial intelligence-created medicine to be used on humans for first time" [3] to report on the development of DSP-1181, a serotonin 5-HT1a receptor agonist, developed by Exscientia and Sumitomo Dainippon Pharma for the treatment of obsessive compulsive disorder (OCD). Chemical Abstract Services (CAS) [4] performed some background investigations regarding this compound. In reviewing the patent data filed (Patent US10800755) presenting DSP-1181 's molecular structure and the other novel AI disclosed molecules within the patent, it was revealed that they all shared a similar shape and molecular scaffold with haloperidol, an antipsychotic which has also been used to treat OCD, and that the majority of AI discovered molecules disclosed within this patent shared the same haloperidol scaffold. In this example, new ML techniques had not ventured into new chemical space or greater diversity, however, DSP-1181, using Exscientia's methods took only 12 months to develop, to phase 1 clinical trials.

Exscientia (together with Evotec) began Phase 1 clinical trials, of another AI discovered drug EXS21546, an adenosine A2a receptor antagonist, again as reported [4] the scaffold was similar in shape to the previously FDA approved A2a antagonists disclosed in Janssen patents in the late 2000s. Another AI Exscientia identified drug, DSP-0038, a dual 5-HT1a receptor agonist and 5-HT2a receptor antagonist shared scaffolds with previously FDA approved drugs used to treat psychiatric illnesses as well. However, designing selective dual activity molecules is a significant challenge for traditional drug discovery methods. This is a challenge as usually the goal in drug design is to optimize a drug for high affinity to a single identified target. Designing a ligand to hit multiple targets requires considering the relative binding affinity of each ligand considered to each of the multiple targets.

There have been projects with AI/ML discovered targets, small molecules and biologics discovered or optimized by AI, and drugs repurposed through AI techniques. Many claims have been made that AI/ML methods can strengthen and accelerate drug pipelines and impact target identification, hit finding and lead optimization [5].

What is the promise of new AI/ML methods for drug discovery in terms of taking us into new chemical space compared with the known currently used computational and medicinal chemistry methods?

How do AI/ML methods complement or compete with physics based methods, like absolute and relative free energy perturbation methods, MMGBSA and molecular dynamics studies? Can ML algorithms offer an assist to physics based methods?

One significant area where ML methods can play a role is in increasing chemical diversity in new chemical entities, not through searching or a virtual screen of the increasing larger databases (i.e. Enamine REAL Space, Wuxi Galaxi, OTAVA, ChemSpace, eMolecules and others containing as many as 10^{14} or greater molecules) but rather finding ways to increase chemical diversity through novel algorithms or methodologies.

Let's examine these different applications and see where ML algorithms have had a significant impact.

Areas where AI has successfully played a significant role in drug discovery:

(1) Predicting Properties and ADMET (absorption, distribution, metabolism)
(2) Hit identification (database searching methods)-small molecule ligand/chemical identification-neural networks, generative chemistry; AI enabled vHTS (virtual high throughput screening)
(3) Target identification and mechanism of action-Target/Protein Modeling and Structure Prediction; OpenFold, AlphaFold2,3, Bolt-1
(4) Docking- AlphaFold3
(5) Drug Design and optimization-including macromolecules and new molecular entities

Property prediction: automatic prediction of molecular properties using substructure vector embeddings within a feature selection workflow

AI/ML methods can be very useful in predicting molecular properties. Unsupervised, self-supervised learning, graph based and geometric models are used for molecular property prediction along with transformer-based language models. In this publication an example was given of the prediction of lipophilicity, logD, using a vector representation of molecular substructures so chemically similar substructures are aligned [6].

Problems of searching large databases

One way to make an ultra large database smaller for searching, while still achieving diversity is to search a small fragment database and then use combinatorics [7]. Chemical Space Docking is a method to accelerate the search through enormous "Chemical Spaces" starting with small fragments called "synthons", which are small

fragments of molecules that contain an extension vector. This vector features information on how the compound can grow through chemical reactions with other building blocks. Once these small fragments are docked at the target, they are expanded into larger, complete compounds. This happens through predefined chemical reactions that connect the initial synthon with other building blocks [8].

Thompson sampling is an active learning approach for virtual screening of large combinatorial libraries performing a probabilistic search in the reagent space, without full enumeration of the library. It can be applied to 2D and 3D similarity search, and docking. In a published study, Thompson sampling identified more than half of the top 100 molecules from a docking-based virtual screen of 335 million molecules by evaluating only 1% of the data set. The methods sole requirement is that the library used is described as a set of building blocks that can be assembled into the final molecules [9].

Generative design, graph neural networks

Atomwise published an extensive initiative using their AI based AtomNet platform to demonstrate competitiveness with traditional virtual HTS methods. AtomNet is a graph convolution network architecture with atoms represented as vertices and pair-wise, distance-dependent, edges representing atom proximities. They used their platform to identify novel bioactive scaffold hits for a diverse set of 235 out of 318 targets without any previously known binding ligands or x-ray structures. Their molecular hits were novel and not similar to the ones found by conventional HTS using standard libraries or databases. Several of their hits were first in class novel scaffold binders for their targets. They were able to identify hits for even some of the challenging targets such as allosteric binders and protein–protein interactions. The AtomNet method did not require a previous known active ligand or a target specific binding training set data [10].

Insilico Medicine have used their generative AI platform called Chemistry 42 to design lung fibrosis candidates as well as candidates for other therapeutic areas. Insilico medicine designed ISM012-042 for treating IBS using their AI Chemistry 42 generative drug design platform to identify a novel PHD inhibitor scaffold and it received approval for phase 1 clinical trials [11]. Insilico identified a target for idiopathic pulmonary fibrosis and designed novel compounds and completed preclinical testing within only 18 months. Their small molecule TNIK inhibitor ISM001-055 completed a phase 2a trial successfully. Insilico indicated that they typically synthesize on average only 70 AI designed molecules for each program [12].

A group at MIT and the Broad trained a deep neural network to predict molecules with antibacterial activity and applied their predictions to several chemical libraries to identify a novel compound, halicin, with antibacterial properties against *Mycobacterium tuberculosis, Clostridoides difficile* and *carbapenem resistant Enterobacteriaceae*. Examining the Zinc15 database, using the neural network model which they developed, they were able to identify 8 antibacterial compounds with novel scaffolds. Their training set was developed from a US FDA library screening for growth inhibition against *E Coli BW25113* and a natural product library, training them as hit or no hit. After developing their model, they applied it to identify antibiotic candidates from the Drug Repurposing Hub, and then larger databases-the Wuxi antituberculosis library and Zinc15 database. They then curated and assayed the hits with the highest scores and retrained their model. The group felt that "the success of deep neural network model guided antibiotic discovery rests heavily on the coupling of these approaches to appropriate experimental designs. The first consideration should be the assay design for training" [13].

Targets and protein structure prediction with ligands: AlphaFold3

Isomorphic Labs and Google DeepMind jointly developed AlphaFold3 (AF3) which predicts protein complexes including nucleic acids, ions, modified residues with ligands (small molecules) already bound within the complex. Alphafold3 directly predicts all these atom coordinates using a diffusion module [14].

So how do Alphafold3 predicted structures compare to a traditional cheminformatics approach of docking a database of ligands? PoseBusters is a benchmark dataset composed of 428 protein–ligand structures released to the PDB in 2021 or later. The main problem with AF3 seems to be maintaining stereochemistry. The AF3 model outputs do not seem to retain the proper chirality, even when reference structures with correct chirality are given as input. There frequently seem to be overlapping clashes seen in the AF3 models produced between the protein and ligand atoms. Clashes seem to frequently occur for nucleotides with the protein in protein-nucleic acid complexes. The modeled protein conformational states may not be correct for the specified ligands and other inputs. For example, E3 ubiquitin ligases natively adopt an open conformation in an apo state and have been observed only in a closed state when bound to ligands, but AF3 exclusively predicts the closed state for both holo and apo systems [15].

Other AI protein models, docking and virtual screening

The predictive protein structure field, advanced significantly with ML models such as AlphaFold, where prior only homology models with significant (> 40%) sequence homology to the target were anywhere close to predicting a correct protein structure, and other threading techniques were poor performers. In 2021, David Baker's, RosettaTTA was the first deep learning method to be successful at the CASP14 (Critical Assessment of Protein Structure) competition [16].

However, how good are AlphaFold, AlphaFold2 and other similar generated protein target models and are they good enough for high throughput ligand virtual screening and ligand docking studies for ligand design? Published studies indicate that using Alphafold2 for virtual screening does not lead to optimum results and that some post-processing modeling may be required in order to have an accurate binding site suitable for docking and computational screening studies [17].

It has been shown in studies with AF2 generated protein target models that small errors present within the predicted structures can cause inaccurate ligand recognition and pose prediction. Unrefined AF2 models have difficulty recognizing ligands and producing correct poses. In a published study, Bryan Roth and Brain Shoichet [18] took as examples two receptors, σ2 (EXPERA protein family) and 5-HT2A (GPCR) for a prospective test of the AF2 models and ligand docking prior to the publication of their crystal structures with ligands.

In retrospective docking screens against the σ2 and 5-HT2A receptors, the AF2 predicted structures had difficulties in selecting the same ligands that were found docking against the receptors' experimental structures. Large library docking studies with the AF2 receptor models, yielded similar hit rates for both receptors as did docking against the experimentally-derived structures. Docking with the AF2 receptor models was successful despite the differences in the binding pocket residue conformations for both of the receptor target models as compared with the experimental solved structures. The results were interpreted to suggest that the AF2 models may sample conformations that are relevant for ligand discovery, indicating that docking studies with the AF2 models were no less effective than those against experimental structures. The hit rates were high for both the σ2 and the 5HT2A receptors across hundreds of molecules experimentally tested against each of the models for both targets, and were not significantly different between the modeled and experimental structures. For the σ2 receptor, 54% of the AF2 model docking hits were active at 1 μM, and for the crystal structure the docking hit rate was 51%. For the 5-HT2A receptor, 26% of the molecules from the AF2- derived model bound at 10 μM, while for the cryoEM experimental structure 23%.While in this particular example AlphaFold2 performed well, it is questionable whether AlphaFold models can be used for virtual screens and replace experimentally solved structures for all protein targets.

PoseBusters checks the quality of docked ligand structures using the RDKit Distance Geometry Module rules-evaluating stereochemistry and inter and intramolecular measurements- bond lengths, planarity of aromatics and atom clashes. In the evaluation and comparison of five deep learning "AI" docking methods- DeepDock, DiffDock, EquiBind, TankBind and Uni-Mol, compared with traditional physics based docking methods -AutoDock Vina and CCDC Gold, the physics based docking methods limited the degrees of movement in the ligand to only the permissible rotatable bonds in the ligand and included penalties for protein and ligand clashes. The conclusion reached by this published study was that "no deep learning-based method yet outperforms classical docking tools". And "molecular mechanics force fields contain docking-relevant physics missing from deep-learning methods" [19].

Another published study indicated that the physics based methods Surflex-Dock, Glide, Vina, and Gnina all performed better than DiffDock (an RF diffusion AI model) on ligand re-docking studies in the known binding- site [20].

In comparing docking of ligands with Alphafold models to docking with homology models for trace amine–associated receptor 1(TAAR1), a set of 30 and 32 highly ranked compounds from the AlphaFold and homology model screens, were experimentally evaluated. Of these, 25 were TAAR1 agonists with potencies ranging from 12 to 0.03 μM. The docking screen with the Alphafold model yielded a more than twofold higher hit rate (60%) than the homology model and discovered the most potent agonists. In this particular example, an AlphaFold modeled structure was demonstrated to outperform a homology model in a virtual screening application [21].

Protein–protein interactions; larger ligands (macromolecules) and new modalities (i.e. molecular glues)

One of the exciting new areas is the development of drugs targeting protein–protein interactions, molecular glues, and new modalities, e.g. PROTACS. An example of a new "fingerprinting" approach to address drug design in this space, is the use of geometric deep learning for molecular surface interaction fingerprinting (MaSIF). Developed through training neural networks on the interactions between proteins and ligands to characterize these interactions and create defined protein–ligand neosurfaces. These neosurfaces, surfaces from protein

ligand complexes, can then be used to predict and design new protein–protein interactions, for example, designing molecular glues or new PPI (protein–protein inhibitors). In some published studies, MaSIF has already been applied designing new drug-inducible protein binders recognizing the B-cell lymphoma 2 (Bcl2) protein in complex with the inhibitor venetoclax; progesterone-binding antibody DB3 in complex with its ligand; and peptide deformylase1 (PDF1) protein from *Pseudomonas aeruginosa* in complex with an antibiotic, actinonin. The method works by finding surface patch descriptors (fingerprints), so that patches with complementary geometry and chemistry have similar fingerprints, whereas non-interacting patches have low fingerprint similarity [22].

AI Deep learning methods have also been used to design macromolecular drugs. The deep learning-based RFdiffusion method was used to design antivenoms to target the short-chain and long-chain α-neurotoxins and cytotoxins from the 3FTx snake venom toxin family [23].

Conclusions

Cheminformatics and chemical data will be used differently in drug discovery and may require different representations moving forward. Deep Learning, graphical neural networks, generative chemistry and other ML methods will call for different representations of ligands in addition to SMILES, and SELFIES. ML methods will be most effective when used in conjunction with experimental data and physics based methods in cyclic retraining workflow methods. As experimental representative protein datasets increase, ML methods for protein structure prediction will improve. More ML methods will be developed like MaSIF and neosurfaces and applied to new motifs- PPIs (protein–protein inhibitors), PROTACS, molecular glues and ADCs (antibody drug conjugates). ML methods and combinatorics will be used more as ultra large screening databases continue to increase in size. This will be an exciting time to see how increased computational power, quantum computing and other computational methods and advances will impact cheminformatics and drug discovery.

Abbreviations

AI	Artificial intelligence
ML	Machine learning
FEP	Free energy perturbation
QSAR	Quantitative structure activity relationship
vHTS	Virtual high throughput screening
AF	AlphaFold
PROTACS	Proteolysis targeting chimera

Author contributions

R.J.B. wrote the manuscript.

Funding

There are no external funders of this work.

Availability of data and materials

No datasets were generated or analysed during the current study.

Declarations

Competing interests

The authors declare no competing interests.

Received: 28 February 2025 Accepted: 22 March 2025
Published online: 06 April 2025

References

1. Leon M, Perezhohin Y, Peres, et al (2024) Comparing SMILES and SELFIES tokenization for enhanced chemical language modeling. Sci Rep 14:25016. https://doi.org/10.1038/s41598-024-76440-8
2. Fierce Biotech , Where is AI winning in drug discovery? 4 use cases to know, https://www.fiercebiotech.com/sponsored/where-ai-winning-drug-discovery-4-use-cases-know-0, https://www.fda.gov/vaccines-blood-biologics/artificial-intelligence-and-machine-learning-aiml-biological-and-other-products-regulated-cber. Accessed 16 Dec 2024.
3. Wakefield J (2020) Artificial Intelligence-created medicine to be used on humans for first time, BBC News, https://www.bbc.com/news/technology-51315462. Accessed 30 Jan 2020.
4. Willis T (2022) AI drug discovery: assessing the first AI-designed drug candidates to go into human clinical trials, https://www.cas.org/resources/cas-insights/ai-drug-discovery-assessing-the-first-ai-designed-drug-candidates-to-go-into-human-clinical-trials
5. Lowe D (2024) In the Pipeline Science Magazine Commentary Blogs : AI Drugs So Far. Science: https://www.science.org/content/blog-post/ai-drugs-so-far. Accessed 13 May 2024.
6. Jung SG, Jung G, Cole JM (2025) Automatic prediction of molecular properties using substructure vector embeddings within a feature selection workflow. J Chem Inf Model 65(1):133–152. https://doi.org/10.1021/acs.jcim.4c01862
7. Beroza P, Crawford JJ, Ganichkin O et al (2022) Chemical space docking enables large-scale structure-based virtual screening to discover ROCK1 kinase inhibitors. Nat Commun 13:6447. https://doi.org/10.1038/s41467-022-33981-8
8. Müller J, Klein R, Tarkhanova O, Gryniukova A, Borysko P, Merkl S, et al. (2022) Magnet for the Needle in Haystack: "Crystal Structure First" Fragment Hits Unlock Active Chemical Matter Using Targeted Exploration of Vast Chemical Spaces. J. Med. Chem. 65: 15663–15678.https://doi.org/10.1021/acs.jmedchem.2c00813.
9. Klarich K, Goldman B, Kramer T, Riley P, Walters WP (2024) Thompson sampling an efficient method for searching ultralarge synthesis on demand databases. J Chem Inf Model 64:1158–1171. https://doi.org/10.1021/acs.jcim.3c01790
10. The Atomwise AIMS Program (2024) AI is a viable alternative to high-throughput screening: a 318-target study. Sci Rep 14:7526. https://doi.org/10.1038/s41598-024-54655-z
11. Ivanenkov YA, Polykovskiy D, Bezrukov D, Zagribelnyy B, Aladinskiy V, Kamya P, Aliper A, Ren F, Zhavoronkov A (2023) Chemistry42: an AI-driven platform for molecular design and optimization. J Chem Inf Model 63:695–701. https://doi.org/10.1021/acs.jcim.2c01191
12. Fu Y, Ding X, Zhang M et al (2024) Intestinal mucosal barrier repair and immune regulation with an AI-developed gut-restricted PHD inhibitor. Nat Biotechnol. https://doi.org/10.1038/s41587-024-02503-w
13. Ren F, Aliper A, Chen J et al (2025) A small-molecule TNIK inhibitor targets fibrosis in preclinical and clinical models. Nat Biotechnol 43:63–75. https://doi.org/10.1038/s41587-024-02143-0)
14. Stokes JM et al (2020) A deep learning approach to antibiotic discovery. Cell 180:688–702. https://doi.org/10.1016/j.cell.2020.01.021
15. Abramson J, Adler J, Dunger J et al (2024) Accurate structure prediction of biomolecular interactions with AlphaFold 3. Nature 630:493–500. https://doi.org/10.1038/s41586-024-07487-w

16. Buttenschoen M, Morris GM, Deane CM (2024) PoseBusters: AI-based docking methods fail to generate physically valid poses or generalize to novel sequences. Chem Sci 15:3130–3139
17. Comment SS (2022) Method of the Year 2021: protein structure prediction. Nat Methods 19:1. https://doi.org/10.1038/s41592-021-01380-4
18. Díaz-Rovira AM, Martín H, Beuming T, Díaz L, Victor L, Guallar V, Ray SS (2023) Are deep learning structural models sufficiently accurate for virtual screening? Application of docking algorithms to AlphaFold2 predicted structures. J Chem Inf Model 63:1668–1674. https://doi.org/10.1021/acs.jcim.2c01270
19. Buttenschoen M, Morris GM, Deane CM (2024) PoseBusters: AI-based docking methods fail to generate physically valid poses or generalize to novel sequences. Chem Sci 15:3130. https://doi.org/10.1021/acs.jcim.2c01191
20. Jain AN, Cleves AE, Walters WP (2024) Deep-Learning Based Docking Methods: Fair Comparisons to Conventional Docking Workflow, arXiv preprint arXiv:2412.02889. https://doi.org/10.48550/arXiv.2412.02889.
21. Díaz-Holguín A, Saarinen M, Vo DD, Sturchio A, Branzell N, Cabeza de Vaca I, Hu H, Mitjavila-Domènech N, Lindqvist A, Svenningsson P (2024) AlphaFold accelerated discovery of psychotropic agonists targeting the trace amine–associated receptor 1. Sci Adv. https://doi.org/10.1126/sciadv.adn1524
22. Lyu J, Kapolka N, Gumpper R, Alon A (2025) Targeting protein-ligand neosurfaces using a generalizable deep learning approach. Nature. https://doi.org/10.1038/s41586-024-08435-4
23. Marchand A, Buckley S, Schneuing A, Pacesa M, Gainza P, Elizarova E, Neeser RM, Lee PW, Reymond L, Elia M, Scheller L, Georgeon S, Schmidt J, Schwaller P, Maerkl SJ, Bronstein M, Correia BE, Torres VS, Benard VM, Mackessy SP et al (2025) De novo designed proteins neutralize lethal snake venom toxins. Nature. https://doi.org/10.1038/s41586-024-08393-x

Publisher's Note