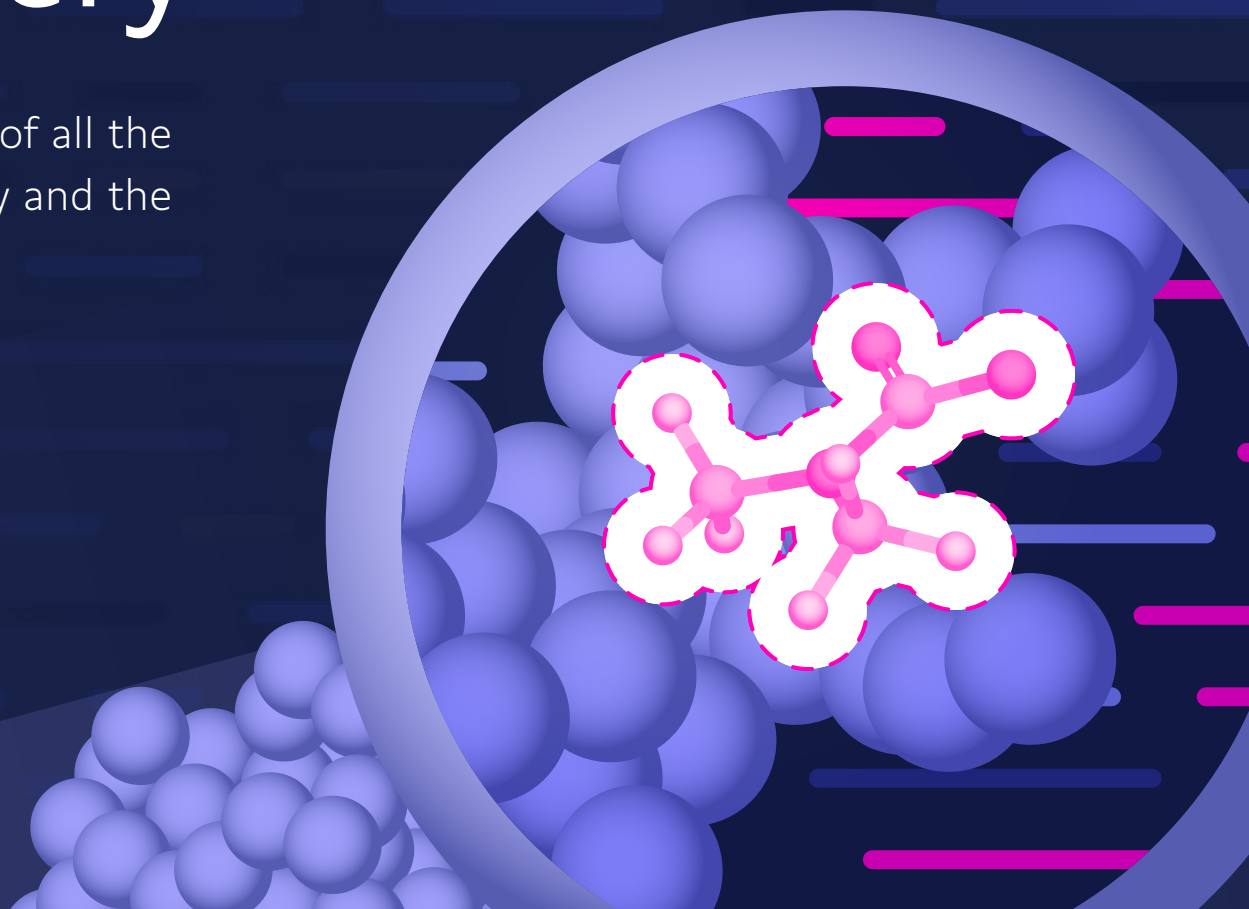
EBOOK

The Little Book Of Big Changes In AI-Powered Drug Discovery

A *somewhat* comprehensive rundown of all the ways AI is transforming drug discovery and the future of biomedical research.



DRUGBANK

Contents

Artificial Intelligence Has Changed The Game

AI has launched itself from the pages of science fiction and disrupted our industry.

230+ Startups using AI[1]

40+ Pharma companies using AI[2]

115+ Drugs developed as a result of AI[3]

$13.8B Invested in companies & partnerships leveraging AI[4]

At this point if you aren't using AI for drug discovery and repurposing, you're setting yourself up to be left behind. As the pace of innovation continues to surge, hundreds of startups are centering their entire organization around AI and seeing astounding results. Pharma companies aren't being shy either, with numerous big names quickly shifting to integrate AI into their practices.

Everyone, at every level of your organization, needs to be thinking about applying AI as a means to reduce costs, increase speed, and ensure your time and efforts are in pursuit of successful drug candidates and treatments.

The Impact On Drug Discovery Is Undeniable

There are currently more than 100 drugs in the AI in drug discovery pipeline, and numerous drugs have been making their way to clinical trials in a matter of years — instead of decades.

2M+ Scientific journals published each year[5]

7% Of systematic reviews are inaccurate within 24hrs of publication[6]

23% Of reviews not updated in 2 years have incorrect conclusions[6]

As AI increasingly becomes standard practice, access to high-quality data is proving to be a true differentiator. You've likely heard, if not already said "garbage in, garbage out," when it comes to drug data. As the move to AI continues to advance it will only become more vital to have reliable, clean, and robust datasets. Without this, you'll find yourself spending more and more time getting datasets ready before you can even begin your research.

Yet, data is growing at such an exponential rate that it is becoming impossible to keep up with, maintain, structure, and normalise it at a pace and standard that enables anyone to use it to its fullest potential.

At DrugBank we work tirelessly to equip leading data scientists with the most in-depth, highest-quality, and up-to-date drug data on the market.

As a result we get to see firsthand how the industry and new technology is transforming the future of medicine and healthcare.

AI-Advancements

Over the past few decades, every facet of the drug discovery pipeline has in one way or another been disrupted by the move towards AI-guided approaches.

The exponential growth and improved coverage and availability of omics-datasets has pushed machine-learning-powered computational methods to become a critical tool in integrating, understanding, and utilizing these datasets to their fullest potential.

The AI in drug discovery timeline is matter of months — *without AI is 5+ years*[7]

Target identification efforts have been hugely improved through the integration of varying, often heterogeneous datasets such as pathways, RNA expression, animal models, mutations, and somatic and germline genetics. The machine learning approaches that are making this possible continue to be increasingly successful in finding targets for both new and existing diseases.

The application of **knowledge graphs** that integrate drugs, targets, diseases, pathways, and other entities, as well as their relationships, have led to a number of promising new approaches.

One example that is still in early stages and currently seeing an intense research focus is the use of knowledge graphs to capture a more representative model of structural biology.

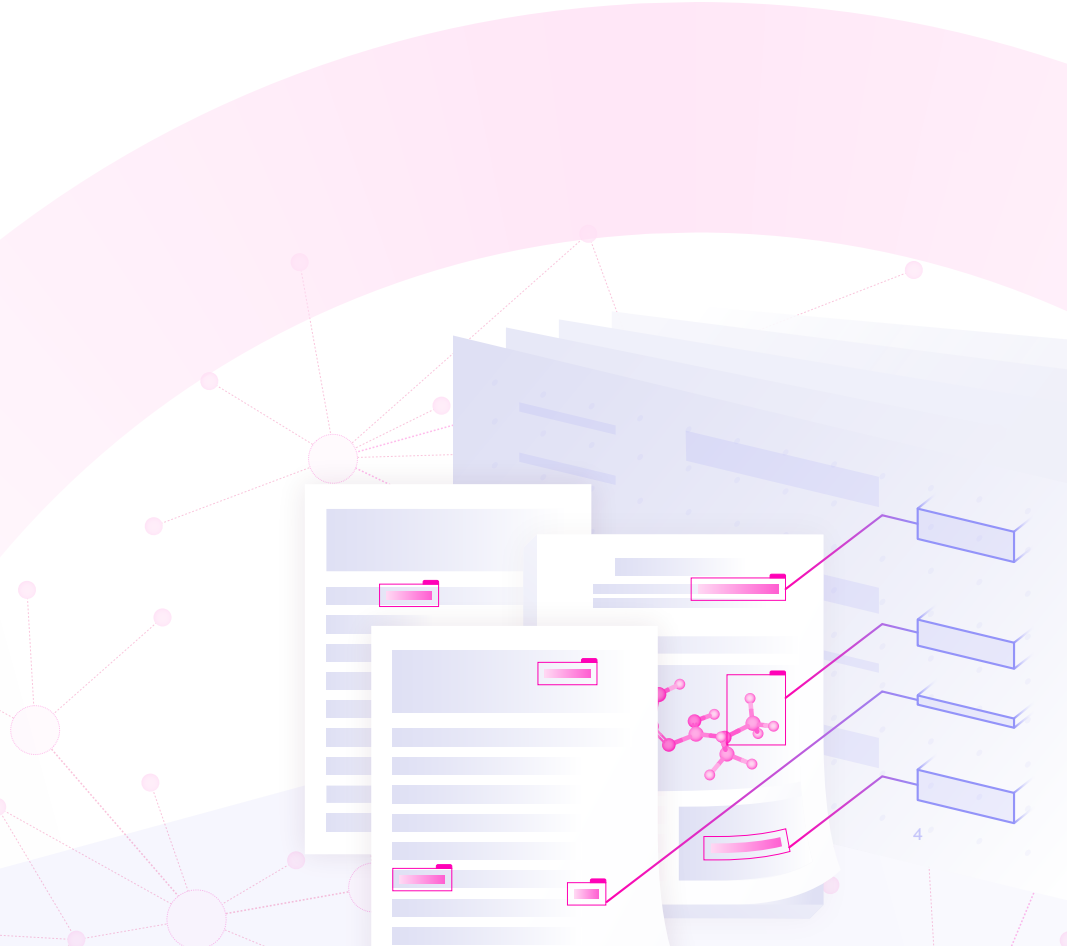
Big Deal Breakthroughs

Although now well-known in the drug-discovery world, the release of protein structure predictions from AlphaFold will have wide-ranging implications across the AI in drug discovery field. Beyond AlphaFold, there have also been significant advances in **cryo-electron microscopy**. Combining these advances introduces a new model for structural biology[8], one which may well become a fertile ground for drug discovery research over the coming decade.

The gaps that exist in AlphaFold predictions (namely conformational dynamics, and disordered proteins and regions) align well with the areas where **cryo-EM / cryo-ET** provide useful information. The combination of AlphaFold predictions with the information on tomography being realized with cryo-EM technology[9] will lead to near-atomic-resolution models of complexes, in their physiological context, inside the cell.

These advances in ML and experimental methods represent the starting point for new approaches to AI-focused drug discovery. The explosion in available **high-quality 3D protein structures** and the continued advancements in **structured, machine-learning-ready drug data** will open up new approaches, and demand entirely new optimizations to handle the intersection between the growing protein structure and chemical compound spaces.

Additionally, the COVID-19 pandemic opened the door to many new innovative, collaborative research projects focused on applying machine learning to better understand the SARS-CoV-2 virus, as well as identify novel antiviral drug candidates.



Research Worth Talking About

AI is making a huge impact across the commercial and academic research space. Below we explore six cases where recent key advances in AI-powered drug discovery have resulted in innovative tools and exciting research findings.

COMMERCIAL

Insilico Medicine

Insilico Medicine *is a Hong Kong, China based company specializing in the development of new AI technologies.*

WHAT THEY'RE UP TO

Insilico has made it their mission to accelerate drug discovery and drug development by continuously inventing and deploying new artificial intelligence technologies. Nearly a decade old, they now have several oncology candidates in their pipeline, and are pursuing the development of both drugs and biomarkers in areas ranging from fibrosis, infectious diseases, immunology, and the process of aging.

HOW THEY'RE USING AI

One of the most noteworthy advancements Insilico has developed involved applying a generative pipeline to complete hit discovery, optimization, synthesis, and validation on candidates against discoidin domain receptor 1 (DDR1), a kinase target implicated in fibrosis and other diseases.

This approach uses a two-step algorithm. The first step involves learning a mapping of the chemical space; the second step explores this mapping using their proprietary deep reinforcement learning platform GENTRL (General Tensorial Reinforcement Learning) to learn DDR1 and common kinase inhibitors. GENTRL utilized three distinct Kohonen-based self-organizing maps (SOMs) as reward functions for the reinforcement learning step: the trending SOM (scores compound novelty based on patent disclosure dates), the general kinase SOM (distinguishes kinase inhibitors), and the specific kinase SOM (isolates DDR1 inhibitors).

This approach identified four active compounds[10], two active in cellular assays, and one lead candidate that demonstrated favorable pharmacokinetics in mice.

Insilico has also advanced our understanding of aging. Applying several supervised machine learning approaches, including neural networks, Insilico built a panel of tissue-specific biomarkers of aging.

WHY WE'RE SO IMPRESSED

Insilico's DDR1 research was able to save significant development time, completing the process in 46 days which was 15-fold faster than traditional approaches.

Insilico was also able to identify the genes most important for age prediction[11], achieving Pearson correlation of 0.91 for the actual age values of the muscle tissue samples.

CelerisTx

Celeris Therapeutics *is a deep learning company focused on developing therapeutics that work to degrade disease-producing proteins. Based in the US, they also have a presence in Austria.*

WHAT THEY'RE UP TO

Celeris Therapeutics focuses on undruggable pathogenic proteins that cause serious conditions such as Alzheimer's and Parkinson's disease. Their current pipeline includes programs in neurology and oncology. Celeris is also using graph neural networks to predict the properties of molecules.

HOW THEY'RE USING AI

In their Xanthos Match Maker platform, Celeris encodes molecular structures in a graph along with features such as the number of hydrogens, valence, and aromaticity and then applies deep neural networks where information about molecules and proteins are processed into an increasingly high-level form.

To make a molecular graph more performant, Celeris uses additional ML techniques to improve molecular fragment linking (linking two fragments binding in nearby subpockets together has become an important technique in fragment-based drug discovery to optimize the binding potency of fragment hits).

In late 2021, Celeris published work[12] describing the use of Variational Autoencoders (VAEs) to augment existing data with a bond-angle-torsion coordinate system, trained on the ZINC dataset, that demonstrated an improvement of 9.3 percent (79 to 88.3) over the previous model (DeLinker).

Applying these approaches and models, Celeris was able to identify novel West Nile Virus NS2B/NS3 protease inhibitors[13]. The viral NS2B/NS3 protease is critical to the viral replication process. Using these deep learning approaches, Celeris was able to identify novel, unexplored drug candidates that demonstrate an inhibition score statistically neighboring experimentally confirmed inhibitors, presenting new candidates for treating West Nile Virus.

WHY WE'RE SO IMPRESSED

Drugs are currently limited in their ability to treat diseases caused by pathogenic proteins.

By establishing reliable, AI-backed methods to leverage the body's natural cell-based mechanisms to degrade these proteins, they were able to identify novel potential treatment options.

Cyclica *is a drug discovery company headquartered in Toronto, Canada, with teams also located in the US and UK. They work to harness AI and machine learning and utilize a custom-built interactome library to model potential protein interactions.*

WHAT THEY'RE UP TO

Cyclica takes an interdisciplinary, collaborative approach to identify molecules that address protein malfunction. The company leverages polypharmacology, a method of concurrently evaluating interactions, to discover new drugs.

HOW THEY'RE USING AI

Cyclica's machine learning platform, MatchMaker, combines features derived from protein targets and small molecules to distinguish binding from non-binding protein-ligand pairs. MatchMaker, trained on ~1.5M human bioactivities (including DrugBank), innovates on existing drug-target interaction models by augmenting the protein representations with structural data.

This approach[14] involves mapping drug-target interaction pairs onto protein binding sites of 3D protein structures. Uncertainties in these predicted and experimental mappings are handled in the model using a deep neural network method called Filtered Transfer Learning (FTL). FTL defines multiple tiers of data confidence as separate tasks in a transfer learning setting. Fine-tuning of the DNN is achieved in a hierarchical process by iteratively removing data points with lower label confidence and retraining.

In 2021, Cyclica applied many of these techniques to quickly identify repurposed drug candidates for COVID-19. Cyclica developed a database called PolypharmDB[15], a deep learning-based resource based on the DrugBank knowledge base. This resource was then utilized by researchers at Ryerson University to predict novel drug candidates[16].

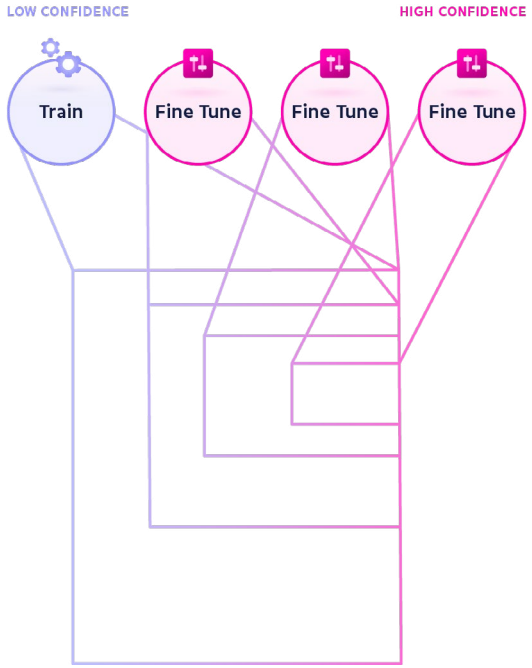


Figure*: Data Partitioning by Confidence for Filtered Transfer Learning*

WHY WE'RE SO IMPRESSED

Cyclica's platform MatchMaker, with its unique approach and generalization capacities, has demonstrated an ability to improve its speed, efficiency, and success rates when predicting potential molecules of interest.

DeepCE

In a collaborative project between Ohio State University, City University of New York, and Cornell University, researchers Thai-Hoang Pham, Yue Qiu, Jucheng Zeng, Lei Xie, and Ping Zhang developed DeepCE, a mechanism-driven neural network-based method.

WHAT IT DOES

DeepCE expands on phenotype-based compound screening by modelling chemical substructure-gene and gene-gene associations, predicting the differential gene expression profile perturbed by *de novo* chemicals. Essentially, DeepCE uses deep learning to predict how drugs will influence the amounts of RNA, and therefore the amounts of various proteins, produced by a cell, which in turn provides insights into how the drug may modulate the disease.

HOW THEY'RE USING AI

DeepCE uses a neural network-based model for gene expression profile prediction consisting of several components. A graph convolutional network is used to learn a vector representation for each chemical compound from its graph structure. A feed-forward neural network is used to learn vector representations for cell line and chemical dose size.

These vector representations are then put into the interaction component (two multihead attention modules, concatenated into a normalization layer followed by feed-forward layer and another normalization layer) to learn high-level feature associations, including chemical substructure-gene and gene-gene feature associations. Finally, the prediction component (two-layer feed-forward neural network with a rectified linear unit activation function) takes the interaction component's outputs as inputs to simultaneously predict the gene expression values for all L1000[17] genes.

To validate the model's effectiveness, the authors utilized DrugBank as a source for clinically relevant drug-target and disease relationships. The results indicated that integrating gene expression profiles generated with DeepCE can solve problems related to unreliable data in the standard (L1000) dataset, leading to better performances on downstream prediction tasks. This specific application of DeepCE represents the first work of phenotype-based drug repurposing for COVID-19.

Going one step further, Deep CE was applied to the full DrugBank dataset, combined with gene expression data from patients with SARS-CoV-2, to predict highly relevant drug candidates.

WHY WE'RE SO IMPRESSED

DeepCE offers improved performance compared to existing methods and has the advantage of providing data augmentation, which makes it possible to tackle areas with minimal or unreliable data.

DOCKSTRING

A research team from the Engineering, Chemistry, and Statistical Laboratory departments at the University of Cambridge, UK created DOCKSTRING, a software and data bundle for meaningful and robust comparison of machine learning models.

WHAT IT DOES

One challenge in drug discovery is being able to utilize the full spectrum of knowledge available. Often, approaches that would be beneficial require the researchers to have a deep level of understanding of the underlying biology. One example of this is molecular docking. It requires extensive domain knowledge to set up experiments and train machine learning correctly. DOCKSTRING was created to help address this challenge.

HOW THEY'RE USING AI

As machine learning methods for drug discovery continue to be developed, benchmarks are required to compare performance against experimental data, giving an indication of what performance can be expected in the real world.

While other benchmarking methods exist, DOCKSTRING offers standardized and accessible benchmarking capabilities based on molecular docking. The three-component DOCKSTRING bundle includes code, datasets, and benchmarking tasks which allow ML practitioners without biological expertise to obtain meaningful docking scores.

The code is an open-source Python package, and the dataset is the first to include docking poses. It is also the most extensive dataset that offers a complete matrix of docking scores for all ligand-target-pairs. This feature enables experiments in transfer learning and multi-objective optimization.

WHY WE'RE SO IMPRESSED

DOCKSTRING, like other recent tools, is aiming to lower the barrier to entry for drug discovery startups.

It goes beyond structure-based modelling and brings more complex techniques for predicting binding affinity into more ligand design pipelines.

EGNNs

The UvA-Bosch Delta Lab at the University of Amsterdam focuses on the fundamentals of deep learning. At the 38th International Conference on Machine Learning in 2021 their team of researchers introduced Equivariant Graph Neural Networks (EGNNs)[18].

WHAT IT DOES

Satorras, Hoogeboom, and Welling introduced the EGNN architecture for graphs that is translation, rotation, reflection, and permutation equivariant.

Trained and tested against the QM9[19 20] dataset (a standard in ML for chemical property prediction tasks), Equivariant Graph Neural Networks (EGNNs) produces highly competitive results in all property prediction tasks while remaining simple, not requiring the use of higher-order molecular representations, molecular angles, or spherical harmonics.

HOW THEY'RE USING AI

Computational *de novo* design of new drugs and optimization of known compounds requires rigorous and unbiased exploration of chemical compound space. Recent advances in Graph Neural Networks (GNNs) have made significant improvements in this space in terms of accuracy and computational efficiency.

The EGNN-based model can predict all features from the QM9 dataset including equilibrium geometries, frontier orbital eigenvalues, dipole moments, harmonic frequencies, polarizabilities, and thermochemical energetics corresponding to atomization energies, enthalpies, and entropies at ambient temperature.

WHY WE'RE SO IMPRESSED

Graph Neural Networks (GNNs) can accelerate the drug discovery process by providing an ability to analyze molecules and their properties at a previously unattainable level, and EGNNs in particular represent a step forward in terms of simplicity and efficiency.

Quality Data, Defined

If there's one thing we know for sure, AI is at its best, and your research will be too, when you have the highest quality drug data that spans a vast range and depth of detail.

In order to ensure that our data is of the highest quality we uphold strict criteria. Before we are satisfied with the quality of our datasets we ask ourselves a series of questions.

Does it have quality coverage?

Coverage means knowing that our data sufficiently captures all relevant medical information.

Is it consistent?

All data must be input in a consistent manner. At DrugBank we have strict curation specifications that all of our data must meet before it is incorporated in our datasets. By standardizing this multi-step peer review process we ensure consistency and accuracy.

Does it tie back to common data entities?

As we add more data to our datasets we create and strengthen many new connections between data points. Each additional connection improves the interoperability of our datasets.

Is it hierarchical and flexible?

Quality data enables you to zoom out or drill down to the appropriate level to adapt the data to the problem being solved. Our data needs to encompass the full variety and complexity of information available, so we incorporate as many levels of detail as possible.

How structured is it?

Structured data is easier to search, find, use, and reason with. We work to create highly structured, detailed data so that our users have total control in how they manipulate and explore it.

Is it evidence-based and can you follow the data-lineage?

Quality data is evidence-based. By ensuring that all our data is based on evidence and its lineage can be traced we are able to review and adjust our data as the underlying evidence changes overtime.

Does it have the appropriate meta-data?

We make sure that we can trace all data curation actions so that we can internally audit and question our data. At any time, we can review who added or updated data and when they did it.

Whether you're midway through years of research or just starting out, it is always worthwhile to assess the quality and source of your data to ensure that you're operating as efficiently as possible, and getting results you can trust.

Hi, We're DrugBank!

DrugBank was founded at the University of Alberta and is the world's first intelligent and comprehensive drug knowledge platform.

With the help of artificial intelligence, our team of medical and scientific experts gather, author, verify, and organize all of the latest, most relevant biomedical information into one machine-learning ready platform. This platform is accessible through data downloads or software integrations and is constantly updated to include the latest findings.

We're working to augment human intelligence so that the world's medical information can be used to its fullest potential and ensure that everyone has access to the best possible medical outcomes.

Our datasets are ideal for all kinds of machine learning, drug discovery applications. As a result we get to work alongside many leading researchers and institutions.

Here's a handful of our most requested datasets:

- **Adverse Effects**
- **Metabolism**
- **Indications**
- **Targets**
- **Protein Relationships**
- **Chemical Structure**
- **Drug Categories**
- **Pharmacology**

Contact us to learn more about DrugBank, our comprehensive drug database, & potential applications.

info@drugbank.com

drugbank.com/datasets

Appendix

Resources

PAPERS

SURVEY PAPERS

Walters and Barzilay, 2021. *Critical assessment of AI in drug discovery.*

Coley, 2020. *Defining and Exploring Chemical Spaces*.

Chuang et al, 2020. *Learning Molecular Representations for Medicinal Chemistry.*

Walters and Barzilay, 2020. *Applications of Deep Learning in Molecule Generation and Molecular Property Prediction.*

Cai et al, 2020. *Transfer Learning for Drug Discovery.*

REPRESENTATION AND TRANSFER LEARNING

Ahmad et al, 2021. *ChemBERTa-2: Towards Chemical Foundation Models.* **[Code]**

Satorras et al, 2021. *E(n) Equivariant Graph Neural Networks.* **[Code]**

Townshend et al, 2021. *ATOM3D: Tasks On Molecules in Three Dimensions.*

Chuang and Keiser, 2020. *Attention-Based Learning on Molecular Ensembles.*

Li and Fourches, 2020. *Inductive transfer learning for molecular activity prediction: Next-Gen QSAR Models with MolPMoFiT.* **[Code]**

GENERATIVE ALGORITHMS

Bengio et al, 2021. *Flow Network based Generative Models for Non-Iterative Diverse Candidate Generation.* **[Code]**

Berenger and Tsuda, 2021. *Molecular generation by Fast Assembly of (Deep)SMILES fragments.* **[Code]**

Gao et al, 2021. *Amortized Tree Generation for Bottom-up Synthesis Planning and Synthesizable Molecular Design.* **[Code]**

Takeuchi et al, 2021. *R-group replacement database for medicinal chemistry.*

Imrie et al, 2020. *Deep Generative Models for 3D Linker Design.* **[Code]**

Jin et al, 2020. *Hierarchical Generation of Molecular Graphs using Structural Motifs.* **[Code]**

Polishchuk, 2020. *CReM: chemically reasonable mutations framework for structure generation.* **[Code]**

HIT FINDING AND POTENCY PREDICITON

Bender et al, 2021. *A practical guide to large-scale docking.*

García-Ortegón et al, 2021. *DOCKSTRING: easy molecular docking yields better benchmarks for ligand design.* **[Code]** **[Data]**

Graff et al, 2021. *Accelerating high-throughput virtual screening through molecular pool-based active learning.* **[Code]**

Gentile et al, 2020. *Deep Docking: A Deep Learning Platform for Augmentation of Structure Based Drug Discovery.* **[Code]**

Cáceres et al, 2020. *Adding Stochastic Negative Examples into Machine Learning Improves Molecular Bioactivity Prediction.*

Lin et al, 2019. *Ultra-large library docking for discovering new chemotypes.*

ADME AND TOXICITY PREDICTION

Siramshetty et al, 2021. *Validating ADME QSAR Models Using Marketed Drugs.*

Göller et al, 2020. *Bayer's in silico ADMET platform: a journey of machine learning over the past two decades.*

Ryu et al, 2020. *DeepHIT: a deep learning framework for prediction of hERG-induced cardiotoxicity.* **[Code]**

SYNTHETIC ACCESSABILITY AND RETROSYNTHETIC PLANNING

Fortunato et al, 2020. *Data augmentation and pretraining for template-based retrosynthetic prediction in computer-aided synthesis planning.*

Koch et al, 2020.*Reinforcement Learning for Bioretrosynthesis.*

Somnath et al, 2020. *Learning Graph Models for Retrosynthesis Prediction.*

VISUALIZATION AND INTERPRETABILITY

Humer et al, 2021. *ChemInformatics Model Explorer (CIME): Exploratory analysis of chemical model explanations.* **[Code]**

Matveieva and Polishchuk, 2021. *Benchmarks for interpretation of QSAR models.* **[Code]**

DATA SETS

- ADME@NCATS
- AMED Cardiotoxicity Database
- BindingDB
- ChEMBL
- DrugBank Online
- DrugMatrix
- Enamine Real database
- hERG Central
- MoleculeNet
- MONA: DB of Mass spec + other readouts
- NPASS database of natural products
- PubChem
- The Open Reaction Database
- Therapeutic Data Commons
- Zinc

For more up-to-date resources, please visit the Awesome Small Molecule Machine Learning GitHub repository.

References

1. Smith, S. 230 *Startups Using Artificial Intelligence in Drug Discovery*. 2021 Jun 10.
2. Smith, S. 43 *Pharma Companies Using Artificial Intelligence in Drug Discovery*. 2021 May 25; Accessed January 27, 2022.
3. Smith, S. 116 *Drugs in the Artificial Intelligence in Drug Discovery Pipeline*. 2021 May 14; Accessed January 27, 2022.
4. Zhang D, Mishra S, Brynjolfsson E, Etchemendy J, Ganguli D, Grosz B, Lyons T, Manyika J, Niebles JC, Sellitto M, Shoham Y, Clark J, Perrault R. *The AI Index 2021 Annual Report*. AI Index Steering Committee: Human-Centered AI Institute, Stanford University. 2021 March.
5. Davis B. *How many scientific papers are published each year?* 2021 Jun 1; Accessed January 27, 2022.
6. Shojania KG, Sampson M, Ansari MT, Ji J, Doucette S, Moher D. *How quickly do systematic reviews go out of date? A survival analysis*. Ann Intern Med. 2007 Aug 21;147(4):224-33. doi: 10.7326/0003-4819-147-4-200708210-00179. Epub 2007 Jul 16. PMID: 17638714.
7. Tsao N. *AI in Drug Discovery Cuts Timeline From 5+ Years to Months*. 2021 Jun 10; Accessed January 27, 2022.
8. Subramaniam S, Kleywegt GJ. *A paradigm shift in structural biology*. Nat Methods 19, 20–23. 2022 Jan 11.
9. de Oliveira TM, van Beek L, Shilliday F, Debreczeni JÉ, Phillips C. *Cryo-EM: The Resolution Revolution and Drug Discovery*. SLAS Discov. 2021 Jan;26(1):17-31. doi: 10.1177/2472555220960401. Epub 2020 Oct 5. PMID: 33016175.
10. Zhavoronkov A, Ivanenkov YA, Aliper A, et al. *Deep learning enables rapid identification of potent DDR1 kinase inhibitors*. Nat Biotechnol 37, 1038–1040. 2019 Sep 2.
11. Mamoshina P, Volosnikova M, Ozerov IV, Putin E, Skibina E, Cortese F, Zhavoronkov A. *Machine Learning on Human Muscle Transcriptomic Data for Biomarker Discovery and Tissue-Specific Drug Target Identification.* Frontiers in Genetics. 2018 Jul 12..
12. Flack M, Webber N, Trummer C. *Decoupled coordinates for machine learning-based molecular fragment linking*. 2021 Nov 1.
13. Celeris, *Idendification of Novel West Nile Virus NS2B/NS3 Protease Inhibitors*. 2021.
14. Cyclica, *Comparison of MatchMaker to DeepConv-DTI reveals superior performance and compute efficiency for predicting drug-target interactions*. 2021.
15. Redka DS, MacKinnon SS, Landon M, et al. *PolypharmDB, a Deep Learning-Based Resource, Quickly Identifies Repurposed Drug Candidates for COVID-19*. ChemRxiv; (2020). DOI: 10.26434/chemrxiv.12071271.v1.
16. Sugiyama MG, Cui H, Redka DS, Karimzadeh M, Rujas E, Maan H, Hayat S, Cheung K, Misra R, McPhee JB, Viirre RD, Haller A, Botelho RJ, Karshafian R, Sabatinos SA, Fairn GD, Madani Tonekaboni SA, Windemuth A, Julien JP, Shahani V, MacKinnon SS, Wang B, Antonescu CN. *Multiscale interactome analysis coupled with off-target drug predictions reveals drug repurposing candidates for human coronavirus disease*. Sci Rep. 2021 Dec 2;11(1):23315. doi: 10.1038/s41598-021-02432-7. PMID: 34857794; PMCID: PMC8640055.
17. Subramanian A, Narayan R, Corsello SM, Peck DD, Natoli TE, Lu X, Gould J, Davis JF, Tubelli AA, Asiedu JK, Lahr DL, Hirschman JE, Liu Z, Donahue M, Julian B, Khan M, Wadden D, Smith IC, Lam D, Liberzon A, Toder C, Bagul M, Orzechowski M, Enache OM, Piccioni F, Johnson SA, Lyons NJ, Berger AH, Shamji AF, Brooks AN, Vrcic A, Flynn C, Rosains J, Takeda DY, Hu R, Davison D, Lamb J, Ardlie K, Hogstrom L, Greenside P, Gray NS, Clemons PA, Silver S, Wu X, Zhao WN, Read-Button W, Wu X, Haggarty SJ, Ronco LV, Boehm JS, Schreiber SL, Doench JG, Bittker JA, Root DE, Wong B, Golub TR. *A Next Generation Connectivity Map: L1000 Platform and the First 1,000,000 Profiles*. Cell. 2017 Nov 30;171(6):1437-1452.e17. doi: 10.1016/j.cell.2017.10.049. PMID: 29195078; PMCID: PMC5990023.
18. Satorras VG., Hoogeboom E, Welling M. *E(n) Equivariant Graph Neural Networks*. 2021.
19. L. Ruddigkeit, R. van Deursen, L. C. Blum, J.-L. Reymond, *Enumeration of 166 billion organic small molecules in the chemical universe database* GDB-17, J. Chem. Inf. Model. 52, 2864–2875, 2012.
20. R. Ramakrishnan, P. O. Dral, M. Rupp, O. A. von Lilienfeld, *Quantum chemistry structures and properties of 134 kilo molecules*, Scientific Data 1, 140022, 2014. [bibtex]

info@drugbank.com | drugbank.com