Streamlining Drug Development with Artificial Intelligence

Using Algorithm-Based Structure Determination and Virtual Screening to Reduce Dependencies and Hasten Lead Discovery

February 11, 2022

CONTENTS

Abstract

Drug development is a resource/time-intensive process. Between target identification and animal trials, lead discovery and optimization has long required extensive effort via X-ray crystallography-driven protein-structure discovery and manual screening processes, such as high throughput screening (HTS).

HTS is often employed to test a wide variety of molecular compounds for effectivity with the desired target. While HTS works well, it is a cost- and labor-intensive process.

Recently, artificial intelligence (AI), specifically machine learning (ML) technology, has provided new capabilities with which to understand highly complex patterns. For instance, biomedical researchers use ML to better model relationships between tumors and drugs.

Meanwhile, virtual screening (VS) of possible drugs, where computers are used to match their molecular structure against target proteins, has risen as a cheaper/faster alternative to traditional HTS. VS leverages existing molecular and computational screening knowledge to predict small molecules that are most likely to be effective drug candidates (lead discovery and optimization) without the need for elaborate HTS efforts for many novel drug targets.

For those with biological and computational skills, there are unprecedented opportunities to help revolutionize the drug development process. Focusing cancer AI efforts, for example, in the HTS and/or pre-clinical space could dramatically expand available and relevant data. Large amounts of data could support identification of patterns of success across phases of investigation, enhance feature selection for cells and drugs, and provide greater insight into cell-line mechanisms of cancer biology and insight into drug structure.

Breakthrough AI-driven protein-structure determination by powerful, general purpose learning algorithms have reduced dependence on X-ray crystallography for protein structure determination and ever-evolving structure-based VS approaches offer to shift molecular screening in silica when appropriate.

These advances offer to streamline lead discovery and optimization and, thus, streamline drug development.

Artificial Intelligence in Contemporary Biomedical Research

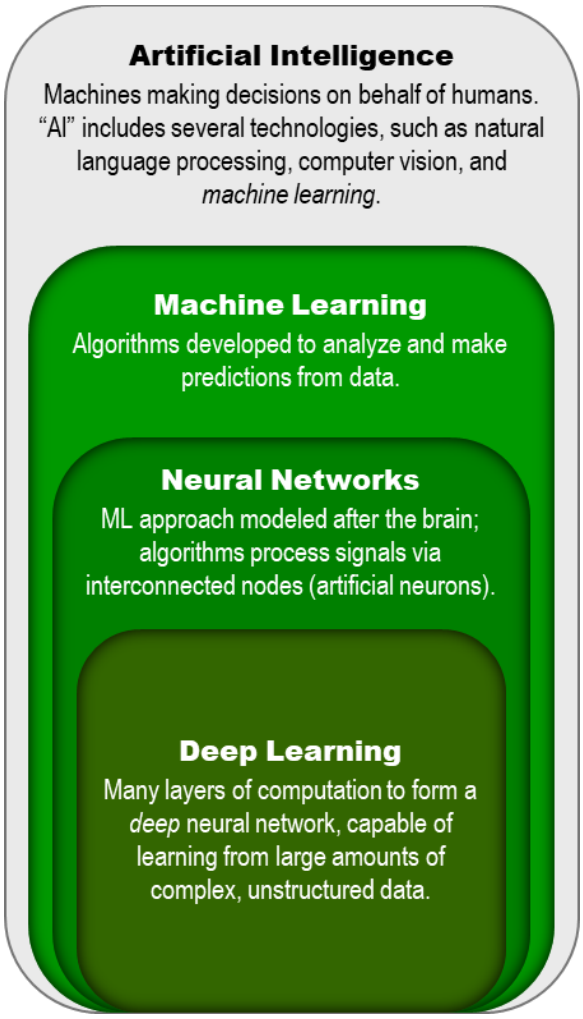
Global investments in oncology have substantially reduced mortality and improved patient outcomes[1], yet cancer remains the leading cause of death worldwide[2]. Even with successful treatment, patients' lives are often negatively affected with long term and severe side effects[3]. New technologies can assist development of more effective and better targeted cancer treatments, which remains a focus for researchers worldwide.

Recently, artificial intelligence (AI), specifically *machine learning* (ML) technology, has provided new capabilities with which to understand highly complex patterns. Neural-network-based (a type of statistical model) ML has found increasing interest within the biomedical research sector.[4] Neural networks can be particularly powerful when researchers have a wealth of data but lack a complete enough understanding of the underlying mechanisms to drive traditional reductionist hypothesis-driven science.[5]

In one such area of interest, researchers use ML to better model relationships between tumors and drugs. ML can capture complex patterns in data and thus potentially

- predict optimal treatment protocol for individual patients (i.e., personalized medicine),
- predict novel effective drug structures, and
- elucidate novel cancer pathways for future therapeutics research.

Figure 1: AI, ML, and Neural Nets



Several technologies are converging to fundamentally change how new disease treatments are discovered. ML is beginning to change the way we approach all aspects of drug development. As it has in fields from travel to finance, this shift to ML/data-driven solutions promises to bring new

[1] National Cancer Institute, *Annual Plan & Budget Proposal for Fiscal Year 2021*, NIH Publication No. 19-7957 (September 2019)
[2] World Health Organization, Cancer (February 3, 2022)
[3] Friese, Christopher R. et al., "Treatment-associated toxicities reported by patients with early-stage invasive breast cancer", *Cancer*, Vol. 123, Issue 11: 1925-1934 (June 1, 2017)
[4] Shah, Pratik et al., "Artificial intelligence and machine learning in clinical development: a translational perspective", *NPJ Digital Medicine*, Vol. 2 69 (July 26, 2019)
[5] Ho, Joshua W. K. & Giannoulatou Eleni, "Big data: the elements of good questions, open data, and powerful software", *Biophysical Reviews*, 11: 1–3 (January 25, 2019)

efficiencies and accelerators to biomedical research.

Virtual screening (VS) of possible drugs, where computers are used to match their molecular structure against target proteins, has risen as a cheaper/faster alternative to traditional, high throughput [drug] screening (HTS). A current limitation of VS is that it requires detailed structural knowledge of target proteins, accessible through painstaking X-ray crystallography experiments. The recent breakthrough of AI-driven protein-structure determination by DeepMind learning algorithms has reduced dependence on X-ray crystallography. Biomedical researchers anticipate that better prediction of protein folding would have wide applications in basic cancer research as well as drug development.[6]

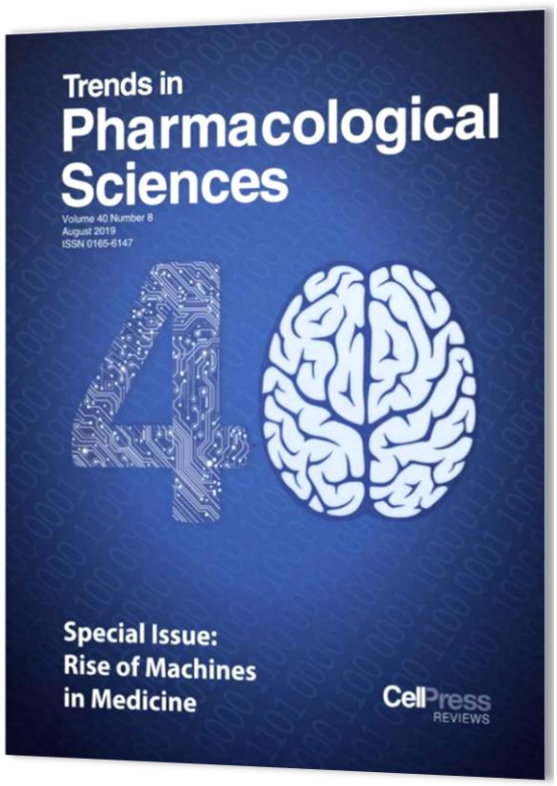
Combined with recent advances in neural networks for VS, this new structural knowledge is poised to vastly increase the number utility of drug candidates.[7]

High Throughput Screening

HTS is a discovery process. It uses several technologies, including automation and high-speed computing, to test thousands to millions of samples per day/week for biological activity at the model organism, cellular, pathway, or molecular levels (i.e., DNA, RNA, proteins, other types of molecules and chemical compounds).

HTS can hasten drug discovery, allowing efficient (high rate, short time) screening of large compound libraries. The output of HTS offers a starting point for drug design and elaboration (i.e., pharmacological probe) used to generate lead compounds, or "hits," with appropriate physicochemical properties for therapeutic indications.

Figure 2: AI Trending in Clinical Trial Design



Current Drug Development Is Slow and Expensive

Drug development is a resource intensive process, requiring on average a 1.3-billion-dollar investment and 10 years to 15 years of investigative effort to achieve FDA approval for a single drug.[8] The process begins by identifying a molecular target, often a protein. Subsequently, a search begins for small molecules (drugs) that bind to and thus modify the activity of said protein in a beneficial way, while minimally affecting other biological processes via binding to unintended proteins. This is typically done via omics and genetic knockout studies. Once a target is identified, molecular screening begins to search for potential drug

[6] Sharpless, Norman E. & Kerlavage, Anthony R., "The potential of AI in cancer care and research", *Biochimica et Biophysica Acta (BBA) - Reviews on Cancer*, Vol. 1876, Issue 1 (August 2021)
[7] Hale, Conor, "Alphabet launches AI drug discovery venture built on DeepMind's protein-folding expertise", *Fierce Biotech* (November 4, 2021)
[8] Harrer, Steven et al., "Artificial Intelligence for Clinical Trial Design", *Trends in Pharmacological Sciences*, Vol. 40 No. 8 (July 17, 2019)

candidates which can effectively modulate said target of interest. HTS is often employed to test a wide variety of molecular compounds for effectivity with the desired target. While HTS works well, it is a cost- and labor-intensive process.

Virtual Screening: An Effective and *Efficient* Alternative

Virtual screening (VS) leverages existing molecular and computational screening knowledge to predict small molecules that are most likely to be effective drug candidates (lead discovery and optimization) without the need for elaborate HTS efforts for many novel drug targets.

Virtual Screening Approaches

VS models can be extended to propose novel drug structures (de novo drug design, a CADD technique to identify drug-like novel chemical structures from a huge chemical search space) and predict likely safety hazards prior to proposition of high resource/lengthy clinical trials.

VS consists of both structure and ligand-based approaches where structure-based approaches require high resolution knowledge of protein structure and/or ligand (drug candidate) orientation within the protein binding site.

Ligand-based approaches rely on databases containing expansive small molecule structures with extensive functional annotation. Historically, determination of three-dimensional (3D) protein structure and/or ligand orientation and binding site has been resource intensive, requiring substantial effort from chemists to manually perform X-ray crystallography on each protein of interest.

Structure-based VS consists of

1. computational representation of proteins and ligands,
2. machine learning classification approaches,
3. data repositories and
4. model evaluation.

At the simplest level, proteins and ligands can be represented as sequence strings. More commonly, 3D grids with each segment of the grid containing a variety of information about the protein/ligand's atomic components can capture more molecular information. They can also be represented as graphs with atoms as nodes with edges representing molecular relationships between atoms. Graphs capture additional spatial information; however, both graphs and grids must be converted to numerical vectors to be used as model inputs.

Generally, the more detailed the representation, the higher the potential for advanced artificial intelligence methods, such as deep learning, to accurately predict ligand binding to produce better drug candidates. In some cases, different types of representation can be used as inputs to the same model to leverage their different strengths, as appropriate.

Machine Learning: CNNs, RNNs, and GNNs

Machine learning approaches attempt to determine which ligands are likely to have high binding affinity with the protein of interest via classification (yes, no) or regression (likelihood of association). Classic machine learning approaches like support vector machines and random forest classifiers have and continue to be effectively used for these purposes. However, deep learning models based on neural nets (statistical models) continue to lead in the segment due to their ability to classify very complex relationships in data.

Model specifics can be nuanced. However,

neural nets typically fall within either convoluted or recurrent paradigms and are designed for either 3D grid or graph data. Convolutional neural networks (CNNs) use hidden layers convolutions (i.e., filters) on 1D, 2D, or 3D data. Frequently used for image analysis, CNNs perform well, but can be slow to train and require input and output of predetermined dimensions.

In contrast, recurrent neural networks (RNNs) have layers which can feed information back on themselves and are well suited to sequential/linear data like auto correct for text and video analysis. As the field moves from sequence-based chemical descriptions to formats which capture more 3D spatial information, CNNs are becoming more appropriate in most cases, but RNNs remain highly viable for preliminary computational exploration.

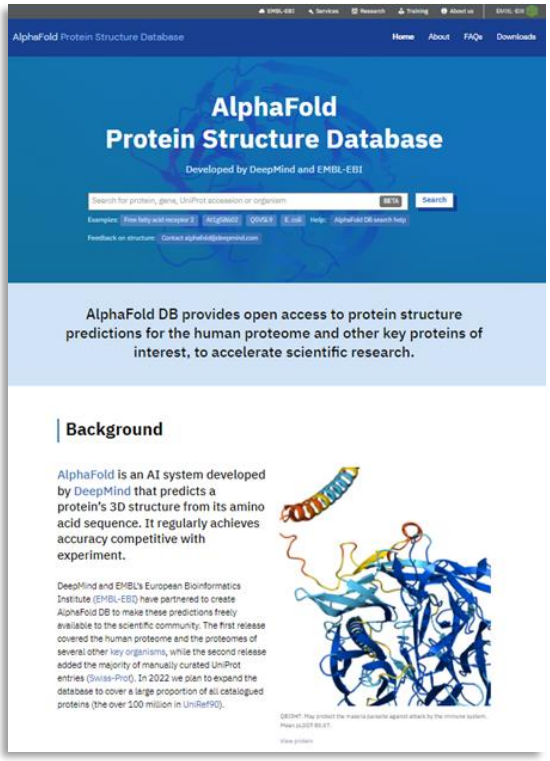
For cases where the input chemical structure is represented as a graph, graph neural networks (GNNs) can use that representation directly.

Neural nets have been used by a wide variety of fields to uncover novel patterns and reach unprecedented predictive accuracy.[9] One thing all neural network approaches share is a requirement for relatively large (by the standards of traditional hypothesis-driven research) amounts of data.

Deep Learning Models and Big Data

Deep learning models require large amounts of input data for training and to screen for predicted efficacy. PDBbind, BindingDB, BindingMOAD each contain tens of thousands of structure-based, protein-ligand binding affinity scores, appropriate for neural network model training. Moreover, the AlphaFold Protein Structure Database includes predicted structure for 20,000 (and

Figure 3: AlphaFold Database



growing) human proteins, computed by DeepMind systems. PubChem, and ChEMBL are major databases containing tens to hundreds of millions of small molecules and their activities (downloadable in SMILES format or as InChI chemical identifier strings or a variety of 2D and 3D image types). By training neural nets on protein/ligand pairs with known binding affinities, potential small molecules with known activity scores can be virtually screened for their ability to bind to diseases' related proteins of interest. To compare across methods, several benchmark datasets exist, including the directory of useful decoys enhanced (DUDE) and maximum unbiased validation (MUV).

A vital step in VS is evaluating model performance. For models whose output is *classification* (i.e., hit or miss), *sensitivity* (proportion of known hits that the model

[9] Widrow, Bernard et al. "Neural Networks: Applications in Industry, Business and Science", *Communications of the ACM*, Vol. 37, Issue 3: 93-105 (March 1994)

calls hits), *specificity* (proportion of known misses that the model calls misses), and their *derivatives* (e.g., overall accuracy and area under the ROC curve) can be computed if known data are split into training and validation sets. In those cases where model outputs are continuous values (e.g., predicted inhibition scores), mean square error rate (the average squared difference between the estimated values and the actual value) or correlation (degree to which predicted values are related to actual values) are often used to evaluate and compare model performance.

Breakthroughs in Computing Protein Structure

Recent advancements in computational protein structure determination directly from amino acid sequences have been achieved by DeepMind, an AI focused sister company of Google. This breakthrough discovery is poised to massively expand the number of high-resolution protein structures available for research. Moreover, computational chemists extended the utility of DeepMind's protein folding software to show it capable of reliably determining orientation of ligands in protein binding sites. This creates a very favorable set of conditions for the growth and utility of structure based VS.

Several recent attempts have demonstrated the power of deep neural networks to leverage protein and small molecule structural information to vastly improve VS. For example, Atomwise applied deep convolutional neural networks (i.e., the AtomNet architecture) to chemical data from ChEMBL and DUDE.[10] AtomNet

Figure 4: Drug-Target Interaction Prediction Proposed via DEEPScreen

combines traditional ligand information with a 3D grid representation of protein-ligand binding interactions.

DEEPScreen also uses deep convolutional neural networks but takes 2D structural drawings of molecules as input.[11] This representation of the data allowed the model to learn novel binding rules without being limited by existing assumptions common in pre-defined "fingerprint" approaches. DEEPScreen is trained on known protein/ligand pairs from CHEMBL.

ParaVS is a GNN-based procedure for virtual screening.[12] It combines a docking-

[10] Wallach, Izhar et al., *AtomNet: A Deep Convolutional Neural Network for Bioactivity Prediction in Structure-based Drug Discovery*, arXiv:1510.02855 (Cornell University; October 10, 2015)
[11] Rifaioglu, Ahmet Sureyya et al., "DEEPScreen: high performance drug–target interaction prediction with convolutional neural networks using 2-D structural compound representations", *Chemical Science*, 2020,11, 2531-2557 (January 8, 2020)
[12] Wu, Junfeng et al., *ParaVS: A Simple, Fast, Efficient and Flexible Graph Neural Network Framework for Structure-Based Virtual Screening*, arXiv:2102.06086 (Cornell University; February 8, 2021)

based screen (for high prediction accuracy) with a deep learning non-docking-based approach to vastly reduce the overall computational requirements.

All three approaches performed well on benchmark databases and many more highly effective approaches can be expected to be developed/integrated in drug development pipelines in coming years.

Conclusion(s)

Applying Biological Knowledge and Computational Skills to Support Identification, Enhance Selection, and Gain Insight

For those with biological and computational skills, there are unprecedented opportunities to help revolutionize the drug development process. Focusing cancer AI efforts in the high-throughput screening (HTS) and/or pre-clinical space could dramatically expand available data (10,000 HTS candidates per FDA approved drug[13]) and provide data relevant to identify novel drug targets. These greater amounts of data could support identification of patterns of success across phases of investigation, enhance feature selection for cells and drugs, and provide greater insight into cell-line mechanisms of cancer biology and insight into drug structure. All of that could be leveraged in various ways to substantially benefit clinical cancer AI.

Streamlining Drug Development

Drug development is a resource/time-intensive process. Between target identification and animal trials, lead discovery and optimization has long required extensive effort via X-ray crystallography-driven protein-structure discovery and high throughput and other manual screening processes. Breakthrough AI-driven protein structure determination by DeepMind learning algorithms have reduced dependence on X-ray crystallography for protein structure determination and ever-evolving structure-based VS approaches offer to shift molecular screening in silica when appropriate.

Together, these advances offer to streamline lead discovery and optimization, representing valuable steps forward in the effort to streamline drug development.

[13] Harrer, Steven et al., "Artificial Intelligence for Clinical Trial Design", *Trends in Pharmacological Sciences*, Vol. 40 No. 8 (July 17, 2019)