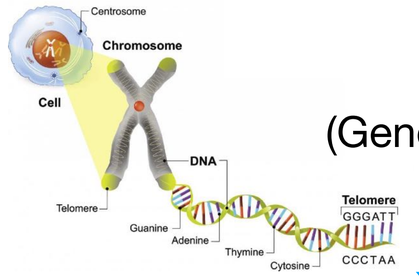
인공지능과 신약 개발

고준수
(주) 아론티어

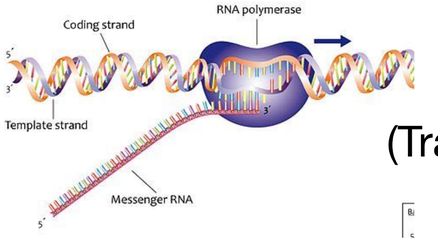
2019-06-21

ARONTIER

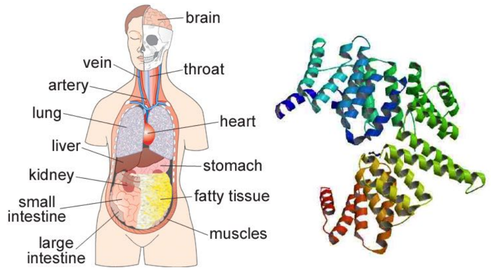
질병 & Drug Discovery



DNA
(Genome, 유전체)



RNA
(Transcriptome, 전사체)



Protein
(단백질)

Target

개수
&
품질
(모양)

"중심이론의 과정 중 어딘가에서 생긴 오류가 마지막 단계까지 쌓여서 문제가 발생한 것"

Drug Discovery

단백질의 잘못된 개수나 품질을 바로 잡을 수 있는 물질 또는 방법을 찾아내는 과정

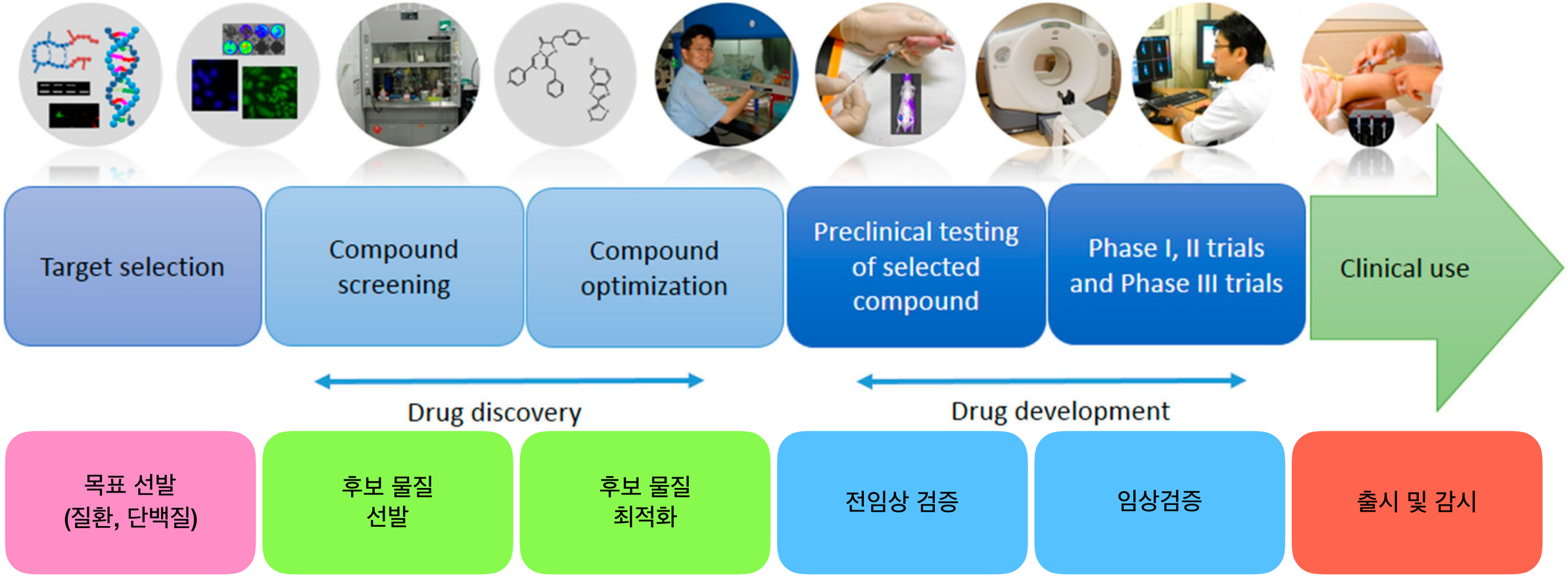
유전체 기반의
표적 탐색

신약 개발

https://www.ldoceonline.com/dictionary/body
https://www.biology-questions-and-answers.com/protein-structure.html
https://genius.com/Biology-genius-the-central-dogma-annotated
https://www.medicalnewstoday.com/articles/319818.php

ARONTIER

신약 개발과정

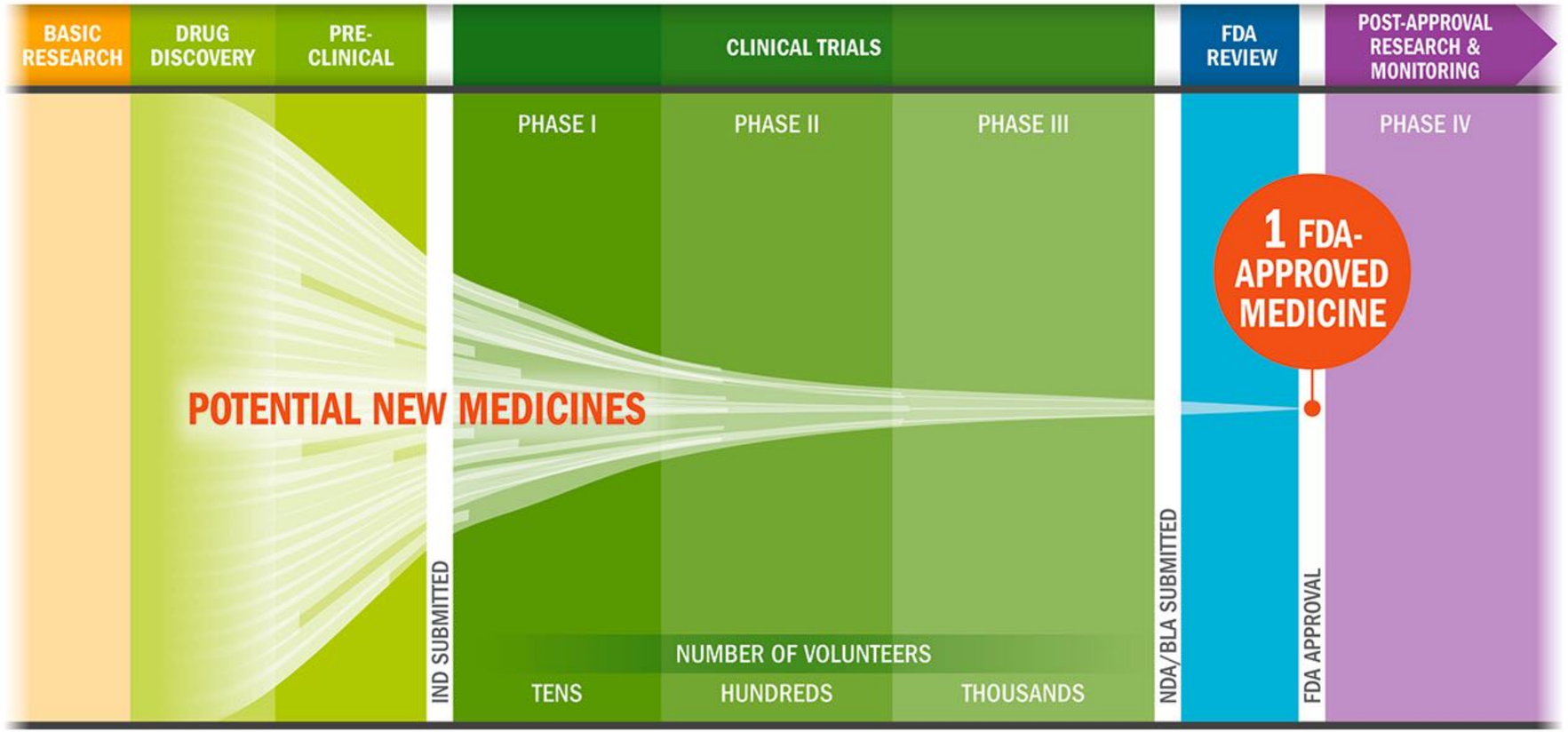


Kalimuthu S., et al. (2017) Int. J. Mol. Sci. 18(8), 1639

ARONTIER

신약 개발과정

From drug discovery through FDA approval, developing a new medicine takes at least 10 years on average and costs an average of $2.6 billion.* Less than 12% of the candidate medicines that make it into Phase I clinical trials will be approved by the FDA.



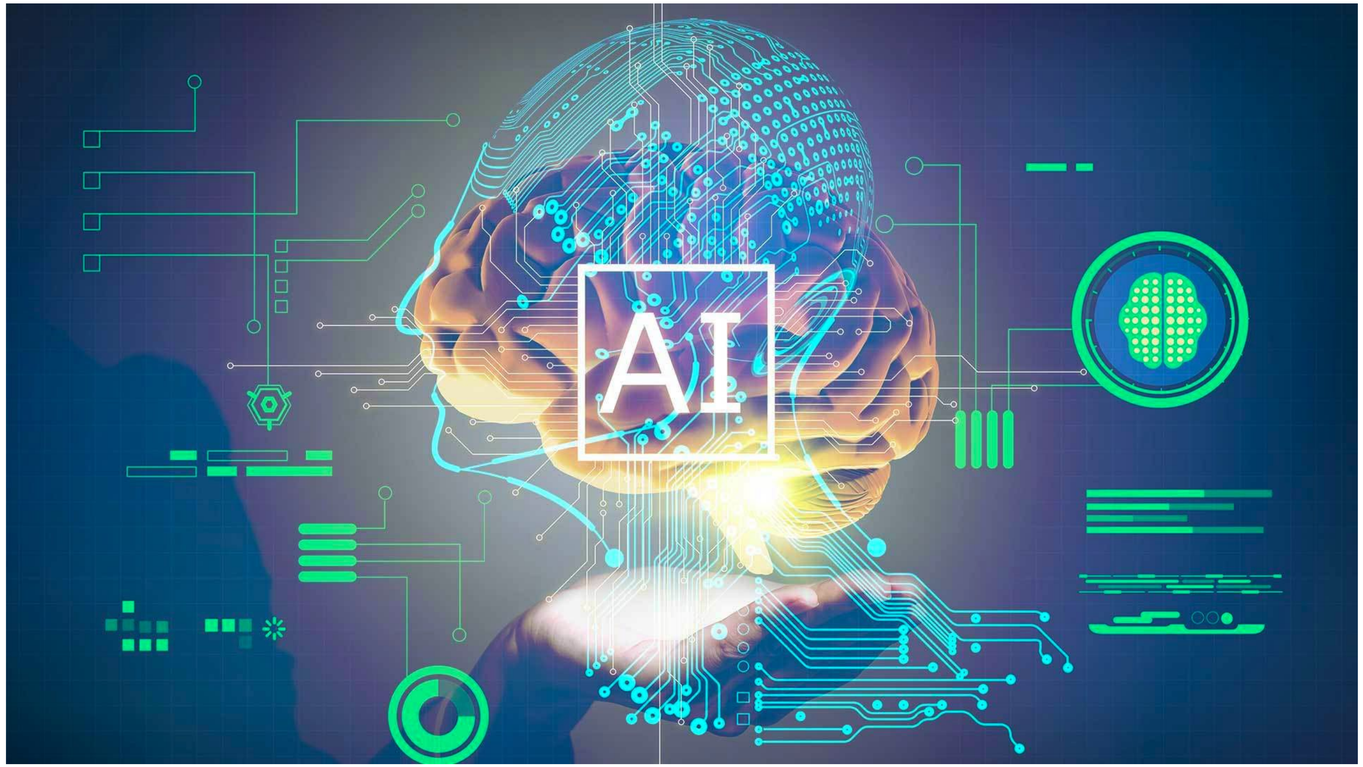
Key: IND: Investigational New Drug Application, NDA: New Drug Application, BLA: Biologics License Application

* The average R&D cost required to bring a new, FDA-approved medicine to patients is estimated to be $2.6 billion over the past decade (in 2013 dollars), including the cost of the many potential medicines that do not make it through to FDA approval.

Source: PhRMA adaptation based on Tufts Center for the Study of Drug Development (CSDD) Briefing: "Cost of Developing a New Drug," Nov. 2014. Tufts CSDD & School of Medicine., and US FDA Infographic, "Drug Approval Process," http://www.fda.gov/downloads/Drugs/ResourcesForYou/Consumers/UCM284393.pdf (accessed Jan. 20, 2015).

ARONTIER

인공지능

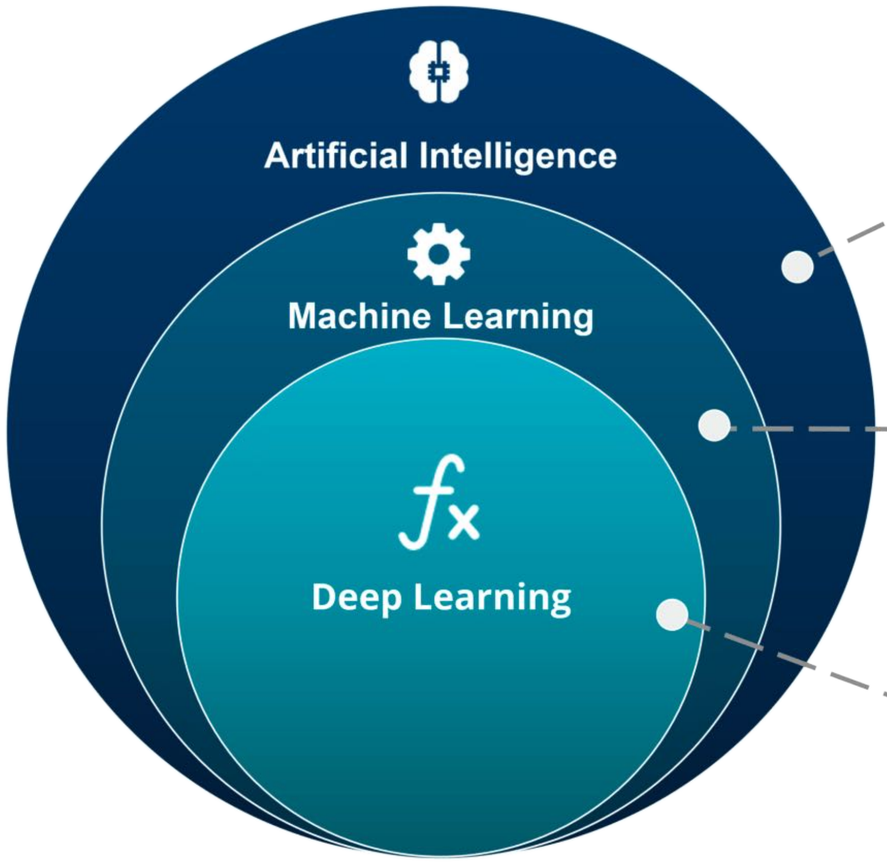


https://code.fb.com/ml-applications/facebook-to-open-source-ai-hardware-design/

http://www.ilovepc.co.kr/news/articleView.html?idxno=12761#09Si

ARONTIER

인공지능 & 기계학습 & 딥러닝



ARTIFICIAL INTELLIGENCE
A technique which enables machines to mimic human behaviour

MACHINE LEARNING
Subset of AI technique which use statistical methods to enable machines to improve with experience

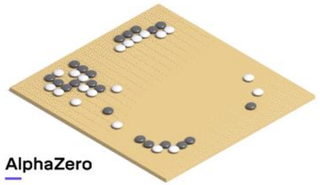
DEEP LEARNING
Subset of ML which make the computation of multi-layer neural network feasible

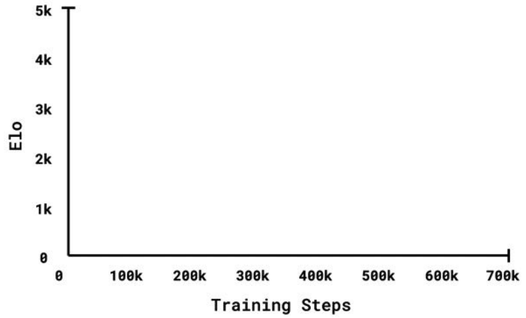
https://www.edureka.co/blog/ai-vs-machine-learning-vs-deep-learning/

인공지능

2016

2017





In chess, AlphaZero first outperformed Stockfish after just 4 hours; in shogi, AlphaZero first outperformed Elmo after 2 hours; and in Go, AlphaZero first outperformed the version of AlphaGo that beat the legendary player Lee Sedol in 2016 after 30 hours. Note: each training step represents 4,096 board positions.

2018

2년만에 프로 바둑기사의 '교과서'된 알파고

인공지능(AI)이 프로 바둑·장기 기사들의 연습 파트너로 자리매김하고 있다. 2016년 이세돌 9단과 알파고가 '세기의 대국'을 하기 전만해도 AI의 실력이 인간을 뛰어넘지 못할 것이란 평가가 지배적이었으나 지금은 세계 정상급 선수들의 교과서가 되고 있다.

http://nowseouledu.com/2016/12/03.php

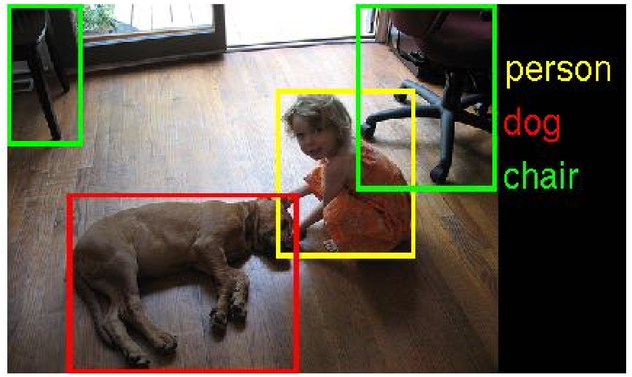
https://deepmind.com/blog/alphazero-shedding-new-light-grand-games-chess-shogi-and-go/

https://www.hankyung.com/article/201806272073i

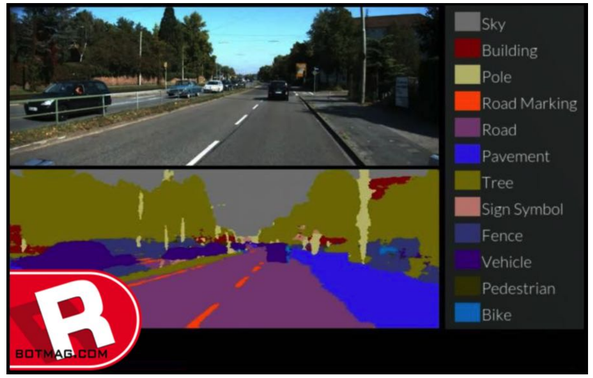
ARONTIER

인공지능의 활용

이미지 분석



자율 주행



음성인식 스피커

문서 작성

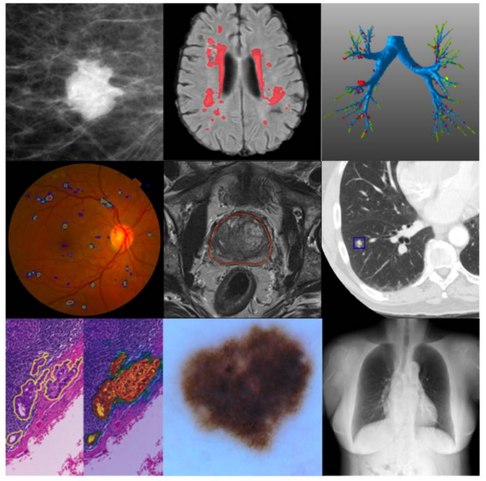
신문기사

번역기

Google Translate

이미지 변환

의료 영상 분석



ARONTIER

신약개발 & 인공지능

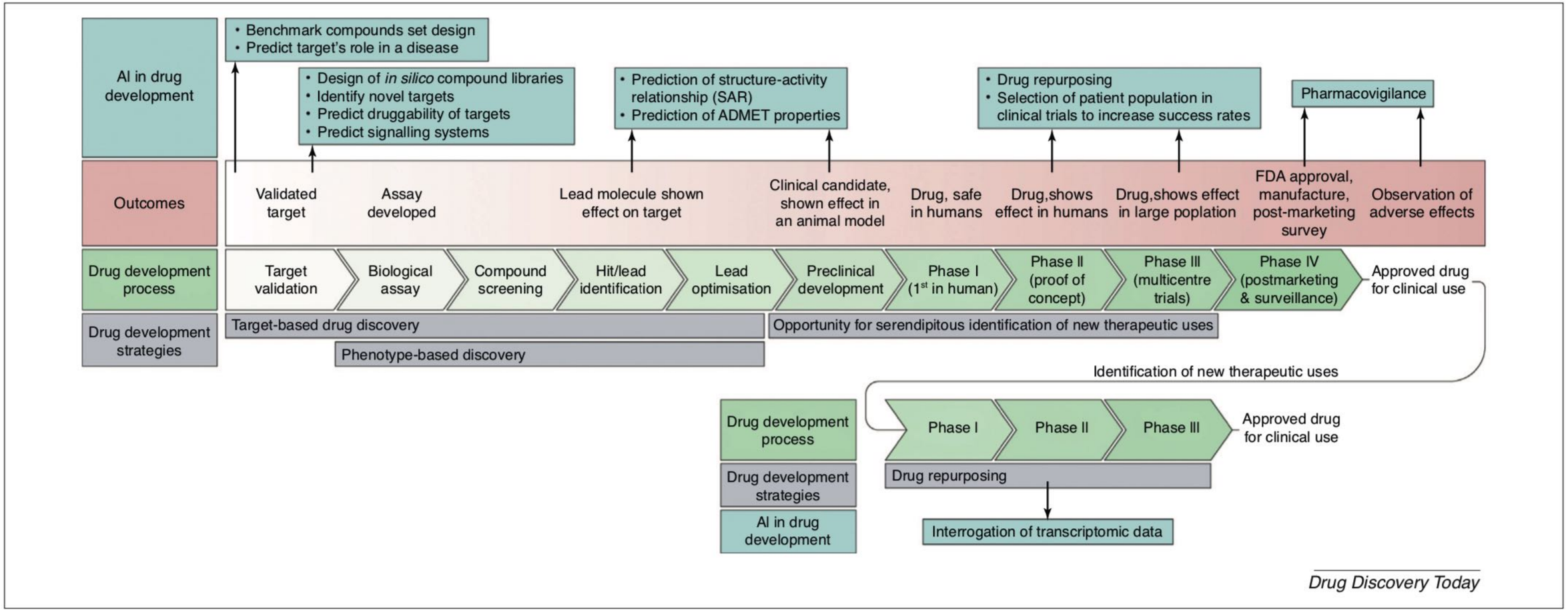


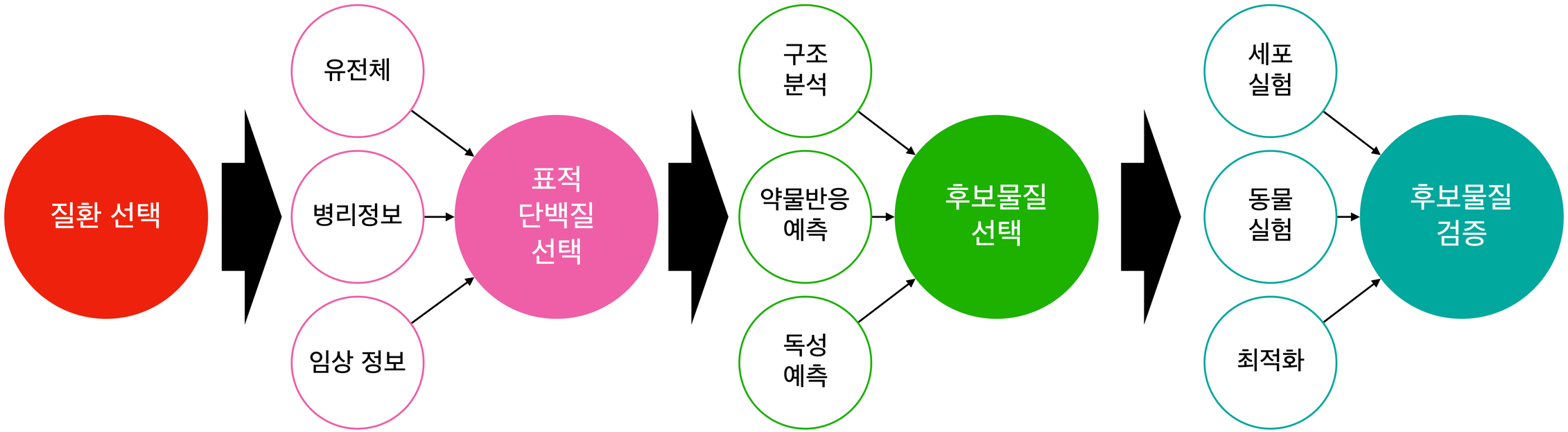
FIGURE 2

Utilisation of artificial intelligence (AI) in the drug development process. The outcomes and the strategies of the various components of the drug development process are described. The applications of AI at each stage of drug development are also shown.

Mak, K.K., Pichika, M.R. Artificial intelligence in drug development: present status and future prospects, Drug Discov Today (2018)

ARONTIER

신약 후보 물질 발굴



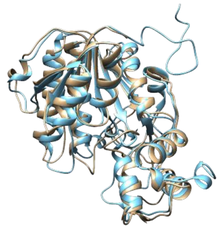
ARONTIER

신약 후보 물질 발굴

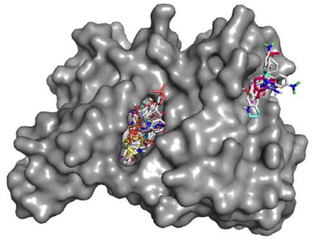
질환 선택

표적 단백질 선택

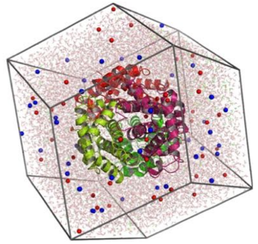
Structure-Based Drug Design (SBDD)



단백질 구조

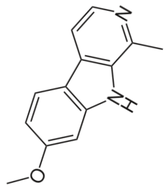


Docking

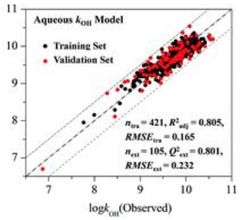


Molecular Dynamics Simulation

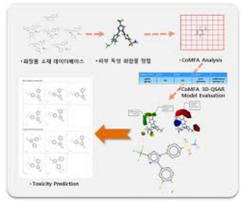
Computer Aided Drug Design (CADD)



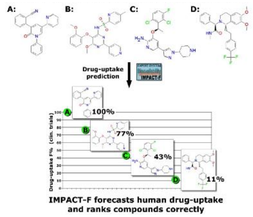
후보 물질 선택



활성도 예측



독성 예측



후보물질 최적화

Ligand-Based Drug Design (LBDD)

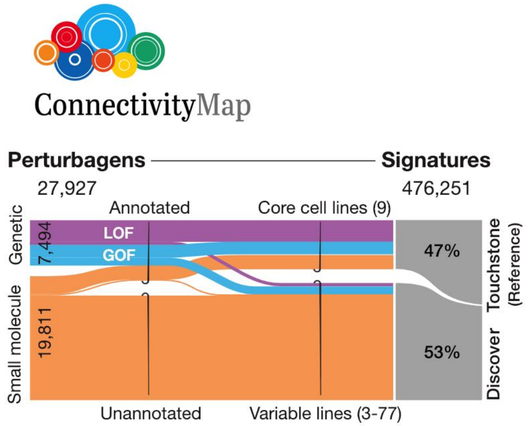
후보물질 검증

https://www.creative-proteomics.com/services/molecular-dynamics-simulation-service.htm

ARONTIER

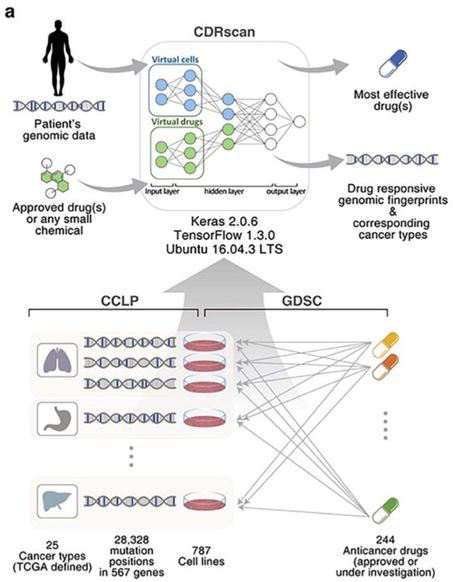
LBDD : 인공 지능을 이용한 기술 변화

유전체 기반의 약물 반응성 DB



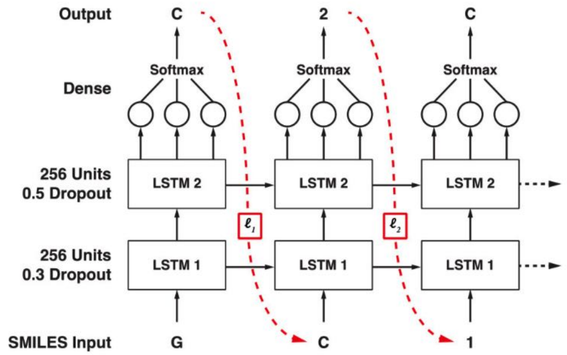
Subramanian, A. et al. (2017).
Cell, 171(6), 1437–1452.e17.

약물 반응성 예측



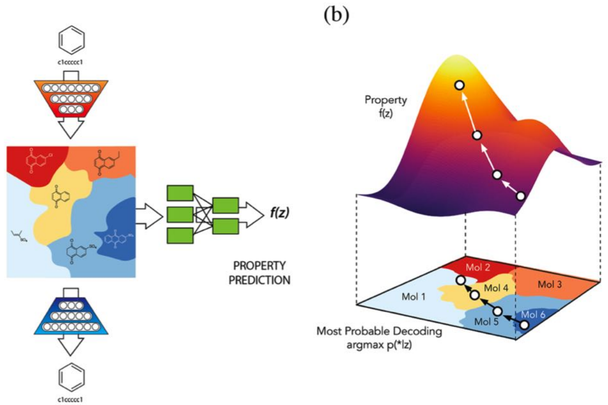
Chang, Y. et al. (2018).
Scientific Reports,
8(1), 8857.

활성약물 자동 디자인



Gupta, A., et al. (2018).
Molecular Informatics,
37(1-2), 1700111.

인공지능을 이용한 약물 최적화



Gómez-Bombarelli, et al.
(2016, October 8)

ARONTIER

SBDD : 인공 지능을 이용한 기술 변화

Support The Guardian
Contribute →
Subscribe →
Sign in
The Guardian
News Opinion Sport Culture Lifestyle
World UK Science Cities Global development Football Tech Business More

Science

Google's DeepMind predicts 3D shapes of proteins

AI program's understanding of proteins could usher in new era of medical progress

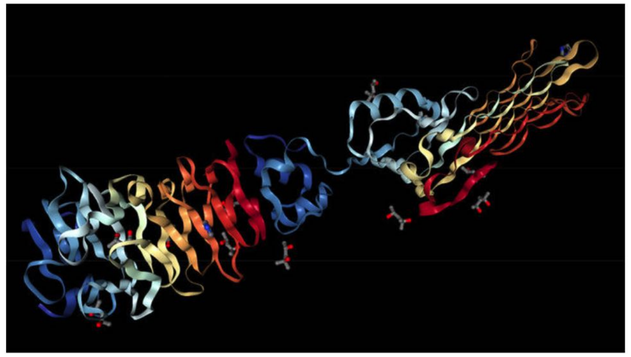
▲ Google's DeepMind artificial intelligence program, AlphaGo, plays South Korean professional Go player, Lee Sedol. Photograph: Ahn Young-joon/AP

Ian Sample *Science editor*

@iansample

Sun 2 Dec 2018 20.55 GMT

This article is over **1 month old**



Complex of bacteria-infecting viral proteins modeled in CASP 13. The complex contains four separate subunits that were modeled individually. PROTEIN DATA BANK

Google's DeepMind aces protein folding

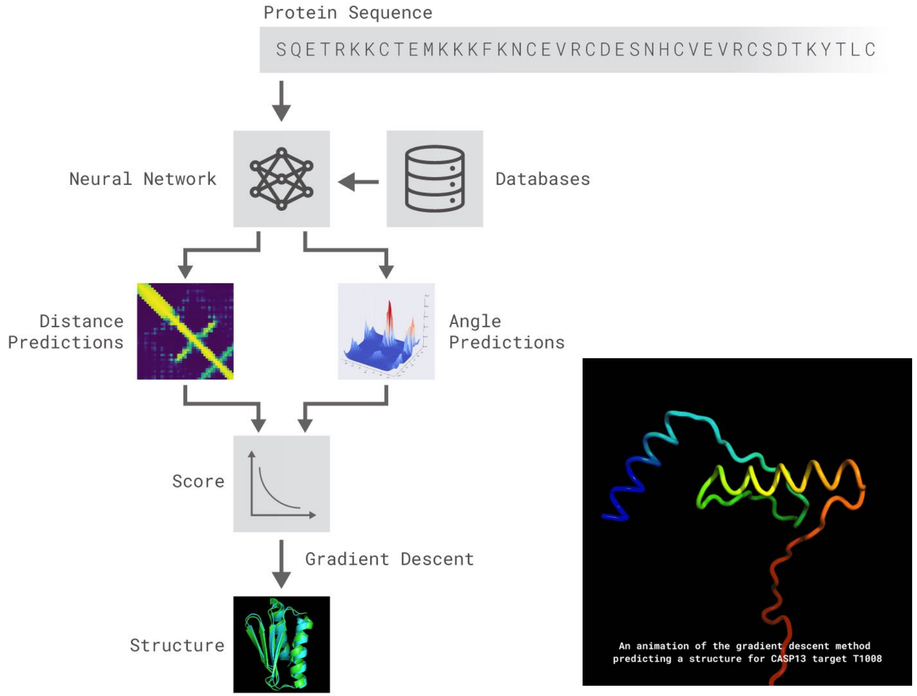
By **Robert F. Service** | Dec. 6, 2018 , 12:05 PM

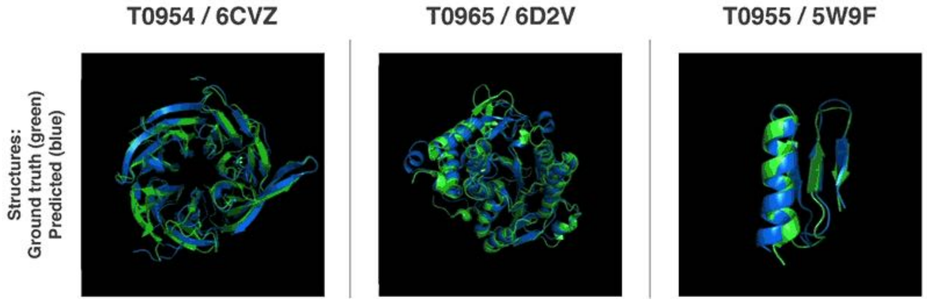
Turns out mastering chess and Go was just for starters. On 2 December, the Google-owned artificial intelligence firm DeepMind took top honors in the 13th Critical Assessment of Structure Prediction (CASP), a biannual competition aimed at predicting the 3D structure of proteins.

The contest worked like this: Competing teams were given the linear sequence of amino acids for 90 proteins for which the 3D shape is known but not yet published. Teams then computed how those sequences would fold. Though London-based DeepMind had not previously joined this competition, the predictions of its AlphaFold software were, on average, more accurate than those of its 97 competitors.

How close was the race? By one metric, not very. For protein sequences for which no other information was known—43 of the 90—AlphaFold made the most accurate prediction 25 times. That far outpaced the second place finisher, which won three of the 43 tests.

SIGN UP FOR OUR DAILY NEWSLETTER
Get more great content like this delivered right to you!

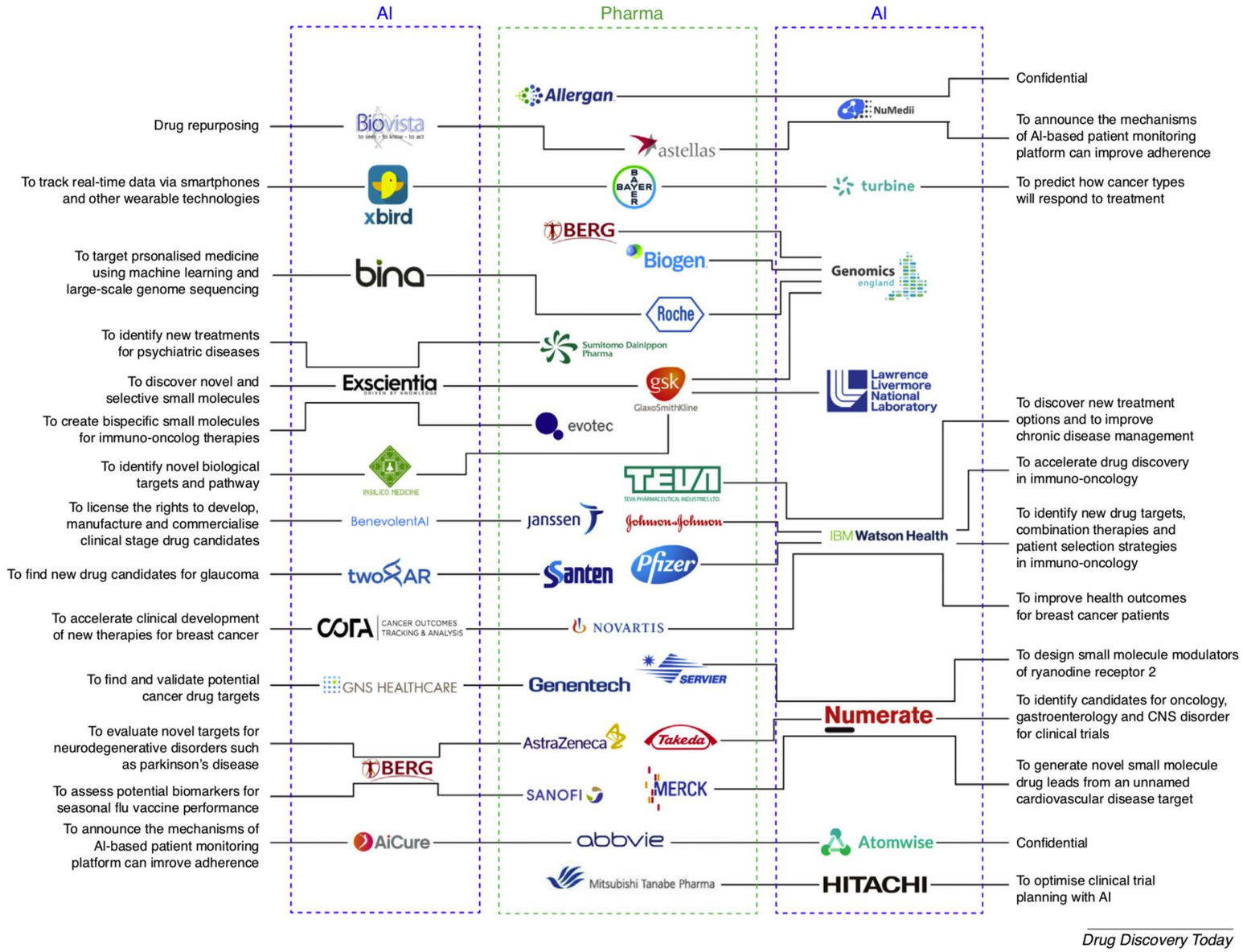




https://www.sciencemag.org/news/2018/12/google-s-deepmind-aces-protein-folding

ARONTIER

인공지능 기반의 신약 개발



Drug Discovery Today

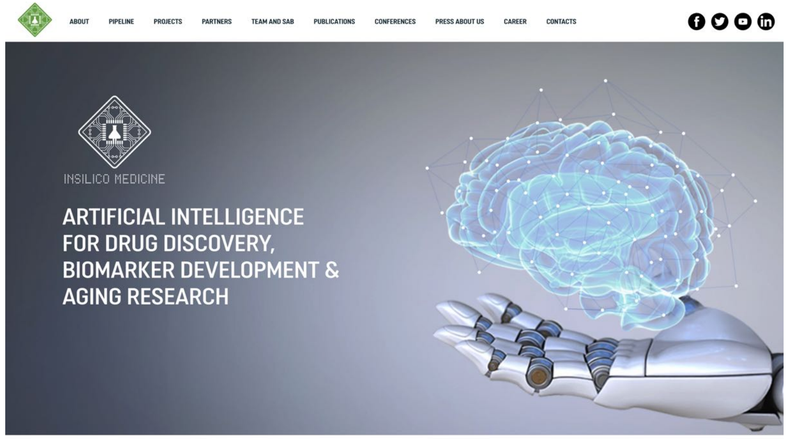
- Group I
  - 인터넷에 공개되어 있는 다양한 데이터 (유전체 정보 등) 을 가공하여 새로운 Drug Target을 발굴하거나 신약 후보 물질을 찾는 회사
- Group II
  - 기존 CADD 시스템의 정확도를 인공지능 기술을 이용해서 향상 시키는 기술을 개발하는 회사

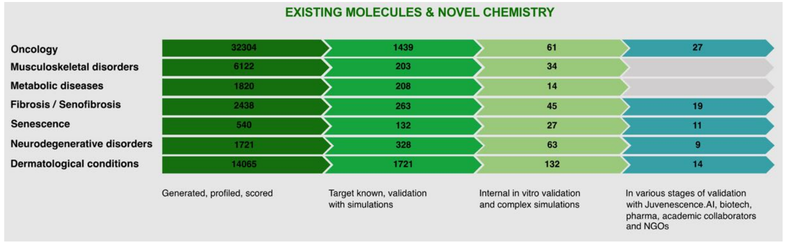
Mak, K.K., Pichika, M.R. Artificial intelligence in drug development: present status and future prospects, Drug Discov Today (2018)

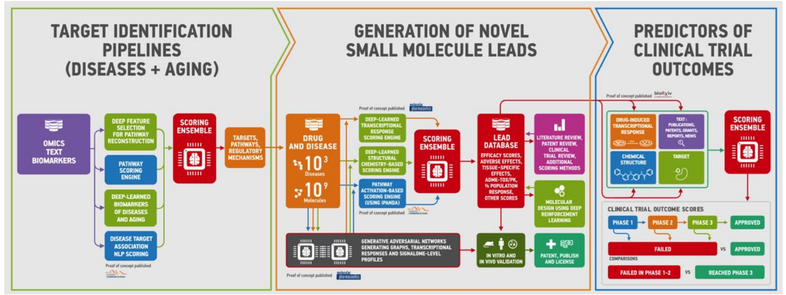
ARONTIER

AI 기반의 신약 개발 회사

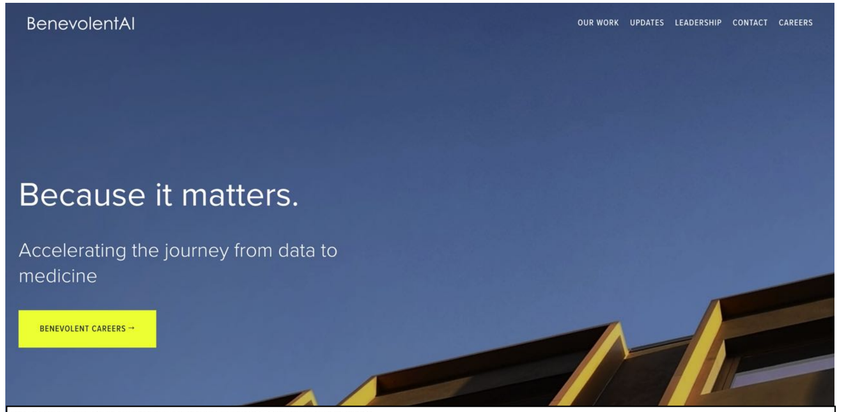
Insilico Medicine

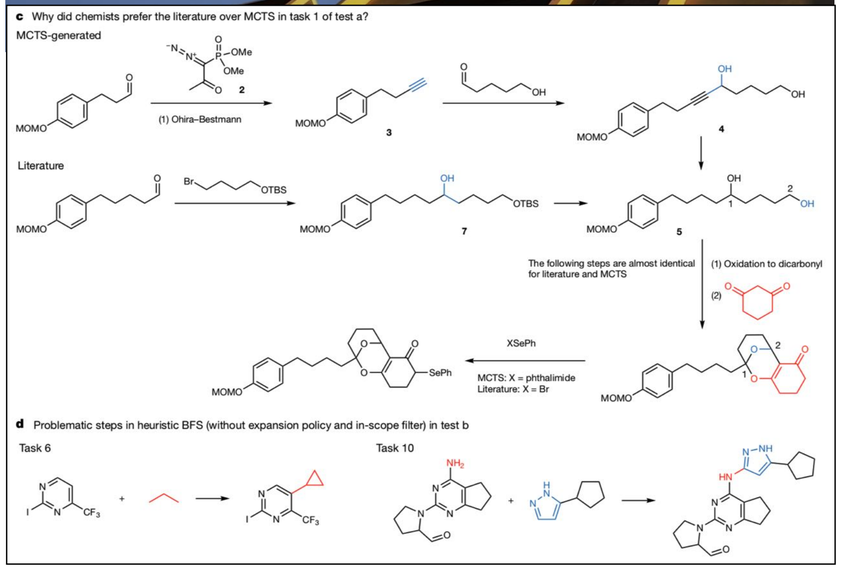






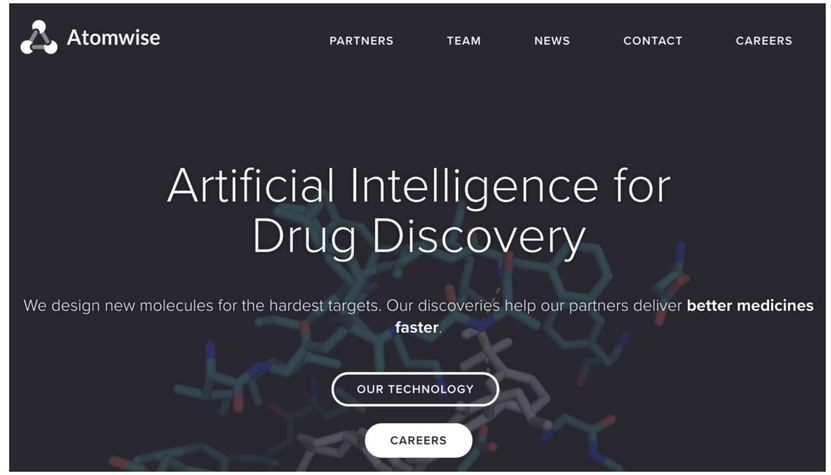
BenevolentAI

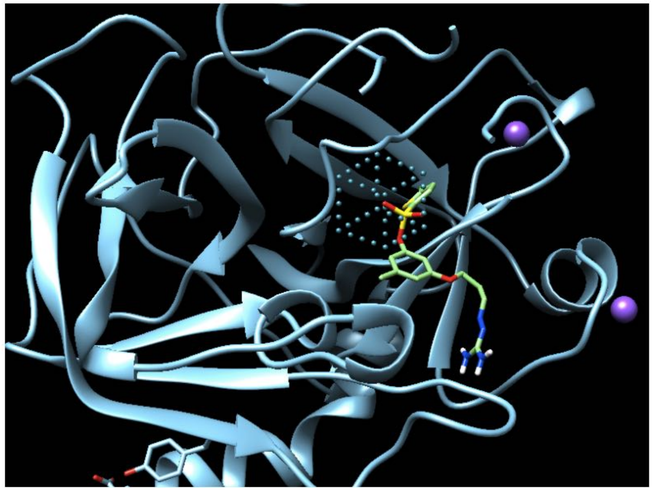




Segler, M. H. S., Preuss, M., & Waller, M. P. (2018)

AtomWise





Wallach, I., Dzamba, M., & Heifets, A. (2015, October 10)

ARONTIER

Drug design & optimization

This is an open access article published under an ACS AuthorChoice License, which permits copying and redistribution of the article or any adaptations for non-commercial purposes.

ACS central science

Research Article

Cite This: ACS Cent. Sci. 2018, 4, 120−131

Generating Focused Molecule Libraries for Drug Discovery with Recurrent Neural Networks

Marwin H. S. Segler,*[,†] Thierry Kogej,[‡] Christian Tyrchan,[§] and Mark P. Waller*[,∥]

[†]Institute of Organic Chemistry & Center for Multiscale Theory and Computation, Westfälische Wilhelms-Universität Münster, 48149 Münster, Germany

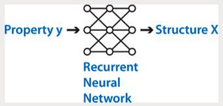
[‡]Hit Discovery, Discovery Sciences, AstraZeneca R&D, Gothenburg, Sweden

[§]Department of Medicinal Chemistry, IMED RIA, AstraZeneca R&D, Gothenburg, Sweden

[∥]Department of Physics & International Centre for Quantum and Molecular Structures, Shanghai University, Shanghai, China

Supporting Information

ABSTRACT: In *de novo* drug design, computational strategies are used to generate novel molecules with good affinity to the desired biological target. In this work, we show that recurrent neural networks can be trained as generative models for molecular structures, similar to statistical language models in natural language processing. We demonstrate that the properties of the generated molecules correlate very well with the properties of the molecules used to train the model. In order to enrich libraries with molecules active toward a given biological target, we propose to fine-tune the model with small sets of molecules, which are known to be active against that target. Against *Staphylococcus aureus*, the model reproduced 14% of 6051 hold-out test molecules that medicinal chemists designed, whereas against *Plasmodium falciparum* (Malaria), it reproduced 28% of 1240 test molecules. When coupled with a scoring function, our model can perform the complete *de novo* drug design cycle to generate large sets of novel molecules for drug discovery.



INTRODUCTION

Chemistry is the language of nature. Chemists speak it fluently and have made their discipline one of the true contributors to human well-being, which has "change[d] the way you live and die".[1] This is particularly true for medicinal chemistry. However, creating novel drugs is an extraordinarily hard and complex problem.[2] One of the many challenges in drug design is the sheer size of the search space for novel molecules. It has been estimated that 10^{60} drug-like molecules could possibly be synthetically accessible.[3] Chemists have to select and examine molecules from this large space to find molecules that are active toward a biological target. Active means for example that a molecule binds to a biomolecule, which causes an effect in the living organism, or inhibits replication of bacteria. Modern high-throughput screening techniques allow testing of molecules on the order of 10^6 in the lab.[4] However, larger experiments will get prohibitively expensive. Given this practical limitation of *in vitro* experiments, it is desirable to have computational tools to narrow down the enormous search space. *Virtual screening* is a commonly used strategy to search for promising molecules among millions of existing or billions of virtual molecules.[5] Searching can be carried out using similarity-based metrics, which provides a quantifiable numerical indicator of closeness between molecules. In contrast, in *de novo* drug design, one aims to directly create novel molecules that are active toward the desired biological target.[6,7] Here, like in any molecular design task, the computer has to

(i) create molecules,

(ii) score and filter them, and

(iii) search for better molecules, building on the knowledge gained in the previous steps.

Task i, the generation of novel molecules, is usually solved with one of two different protocols.[7] One strategy is to build molecules from predefined groups of atoms or fragments. Unfortunately, these approaches often lead to molecules that are very hard to synthesize.[8] Therefore, another established approach is to conduct virtual chemical reactions based on expert coded rules, with the hope that these reactions could then also be applied in practice to make the molecules in the laboratory.[9] These systems give reasonable drug-like molecules and are considered as "the solution" to the structure generation problem.[2] We generally share this view. However, we have recently shown that the predicted reactions from these rule-based expert systems can sometimes fail.[10−12] Also, focusing on a small set of robust reactions can unnecessarily restrict the possibly accessible chemical space.

Task ii, scoring molecules and filtering out undesired structures, can be solved with substructure filters for undesirable reactive groups in conjunction with established approaches such as docking[13] or machine learning (ML) approaches.[7,14,15] The ML approaches are split into two branches: Target prediction classifies molecules into active and inactive, and quantitative structure−activity relationships

Received: October 24, 2017
Published: December 28, 2017

 © 2017 American Chemical Society

DOI: 10.1021/acscentsci.7b00512
ACS Cent. Sci. 2018, 4, 120−131

Full Paper

www.molinf.com

molecular informatics

DOI: 10.1002/minf.201700111

Generative Recurrent Networks for *De Novo* Drug Design

Anvita Gupta,[a, b] Alex T. Müller,[a] Berend J. H. Huisman,[a] Jens A. Fuchs,[a] Petra Schneider,[a, c] and Gisbert Schneider*[a]

Abstract: Generative artificial intelligence models present a fresh approach to chemogenomics and *de novo* drug design, as they provide researchers with the ability to narrow down their search of the chemical space and focus on regions of interest. We present a method for molecular *de novo* design that utilizes generative recurrent neural networks (RNN) containing long short-term memory (LSTM) cells. This computational model captured the syntax of molecular representation in terms of SMILES strings with close to perfect accuracy. The learned pattern probabilities can be used for *de novo* SMILES generation. This molecular design concept eliminates the need for virtual compound library enumeration. By employing transfer learning, we fine-tuned the RNN's predictions for specific molecular targets. This approach enables virtual compound design without requiring secondary or external activity prediction, which could introduce error or unwanted bias. The results obtained advocate this generative RNN-LSTM system for high-impact use cases, such as low-data drug discovery, fragment based molecular design, and hit-to-lead optimization for diverse drug targets.

Keywords: Chemogenomics · deep learning · drug discovery · machine learning · medicinal chemistry

1 Introduction

Compound repositories of pharmaceutical companies contain up to a few million compounds. Even accounting for growth over time, these readily screenable libraries cover only a miniscule fraction of the synthetically accessible, druglike chemical space, which is estimated to contain $>10^{10}$ molecules.[1] Because chemical space is too large to be screened in its entirety for drugs active against a particular target, automated design and screening of selected compounds with desired properties and likelihood of activity presents itself as a complementary approach. Computational *de novo* drug design involves exploring this vast chemical space for such compounds which may not have been synthesized before, and "deep learning" methods present concepts for chemical space navigation.[2] Here, we present a generative deep learning model based on recurrent neural networks (RNNs) for *de novo* drug design. We demonstrate the model's efficacy in three main use cases of *de novo* design: generating libraries for high-throughput screening, hit-to-lead optimization, and fragment-based hit discovery.

RNNs successfully solve machine learning tasks, such as natural language processing[3] and translation,[4] and composing music,[5] to name only a few domains. In particular, much of this success has been achieved by the use of recurrent networks of LSTM (*long short-term memory*) cells, first introduced by Hochreiter and Schmidhuber in 1997.[6] In the field of molecular informatics, RNNs based on LSTMs have been used to predict protein function from sequence[7] and successfully predict aqueous solubility of drug-like compounds.[8] RNNs were used as autoencoders to provide a latent representation of molecular structure for sampling preferred regions of chemical space.[9] Importantly, several research groups have recently demonstrated that RNNs can be employed to generate canonical SMILES strings, and can be fine-tuned by transfer learning.[10,11] In transfer learning, the machine learning model tries to keep information from a previously learned task to solve a different but related, yet unseen task.[12] Researchers at AstraZeneca have extended SMILES-generating RNNs by using this concept for reinforcement learning. The model's parameters were optimized to produce strings that scored highly according to an external scoring function. They applied this approach to generate sets of structures with low sulfur content, high predicted target activity, and other desirable properties.[13]

We here present a new approach to *de novo* drug design using RNN deep learning methodology (Figure 1). In

[a] A. Gupta, A. T. Müller, B. J. H. Huisman, J. A. Fuchs, P. Schneider, G. Schneider
Swiss Federal Institute of Technology (ETH), Department of Chemistry and Applied Biosciences, Vladimir−Prelog−Weg 4, 8093 Zurich, Switzerland
E-mail: gisbert@ethz.ch
[b] A. Gupta
Stanford University, Department of Computer Science, 450 Sierra Mall, Stanford, CA, 94305, USA
[c] P. Schneider
inSili.com GmbH, 8049 Zurich, Switzerland

© 2017 The Authors. Published by Wiley-VCH Verlag GmbH & Co. KGaA.
This is an open access article under the terms of the Creative Commons Attribution Non-Commercial NoDerivs License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

 © 2018 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim

ACS central science

Research Article

Cite This: ACS Cent. Sci. 2018, 4, 268−276

Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules

Rafael Gómez-Bombarelli,[†,#] Jennifer N. Wei,[‡,#] David Duvenaud,[¶,#] José Miguel Hernández-Lobato,[§,#] Benjamín Sánchez-Lengeling,[‡] Dennis Sheberla,[‡] Jorge Aguilera-Iparraguirre,[†] Timothy D. Hirzel,[†] Ryan P. Adams,[∇,∥] and Alán Aspuru-Guzik*[,‡,⊥]

[†]Kyulux North America Inc., 10 Post Office Square, Suite 800, Boston, Massachusetts 02109, United States

[‡]Department of Chemistry and Chemical Biology, Harvard University, Cambridge, Massachusetts 02138, United States

[¶]Department of Computer Science, University of Toronto, 6 King's College Road, Toronto, Ontario M5S 3H5, Canada

[§]Department of Engineering, University of Cambridge, Trumpington Street, Cambridge CB2 1PZ, U.K.

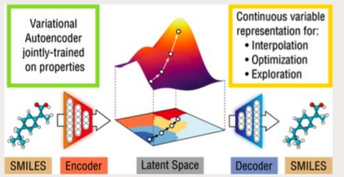
[∇]Google Brain, Mountain View, California, United States

[∥]Princeton University, Princeton, New Jersey, United States

[⊥]Biologically-Inspired Solar Energy Program, Canadian Institute for Advanced Research (CIFAR), Toronto, Ontario M5S 1M1, Canada

Supporting Information

ABSTRACT: We report a method to convert discrete representations of molecules to and from a multidimensional continuous representation. This model allows us to generate new molecules for efficient exploration and optimization through open-ended spaces of chemical compounds. A deep neural network was trained on hundreds of thousands of existing chemical structures to construct three coupled functions: an encoder, a decoder, and a predictor. The encoder converts the discrete representation of a molecule into a real-valued continuous vector, and the decoder converts these continuous vectors back to discrete molecular representations. The predictor estimates chemical properties from the latent continuous vector representation of the molecule. Continuous representations of molecules allow us to automatically generate novel chemical structures by performing simple operations in the latent space, such as decoding random vectors, perturbing known chemical structures, or interpolating between molecules. Continuous representations also allow the use of powerful gradient-based optimization to efficiently guide the search for optimized functional compounds. We demonstrate our method in the domain of drug-like molecules and also in a set of molecules with fewer that nine heavy atoms.



INTRODUCTION

The goal of drug and material design is to identify novel molecules that have certain desirable properties. We view this as an optimization problem, in which we are searching for the molecules that maximize our quantitative desiderata. However, optimization in molecular space is extremely challenging, because the search space is large, discrete, and unstructured. Making and testing new compounds are costly and time-consuming, and the number of potential candidates is overwhelming. Only about 10^8 substances have ever been synthesized,[1] whereas the range of potential drug-like molecules is estimated to be between 10^{23} and 10^{60}.[2]

Virtual screening can be used to speed up this search.[3−6] Virtual libraries containing thousands to hundreds of millions of candidates can be assayed with first-principles simulations or statistical predictions based on learned proxy models, and only the most promising leads are selected and tested experimentally.

However, even when accurate simulations are available,[7] computational molecular design is limited by the search strategy used to explore chemical space. Current methods either exhaustively search through a fixed library,[8,9] or use discrete local search methods such as genetic algorithms[10−15] or similar discrete interpolation techniques.[16−18] Although these techniques have led to useful new molecules, these approaches still face large challenges. Fixed libraries are monolithic, costly to fully explore, and require hand-crafted rules to avoid impractical chemistries. The genetic generation of compounds requires manual specification of heuristics for mutation and crossover rules. Discrete optimization methods have difficulty

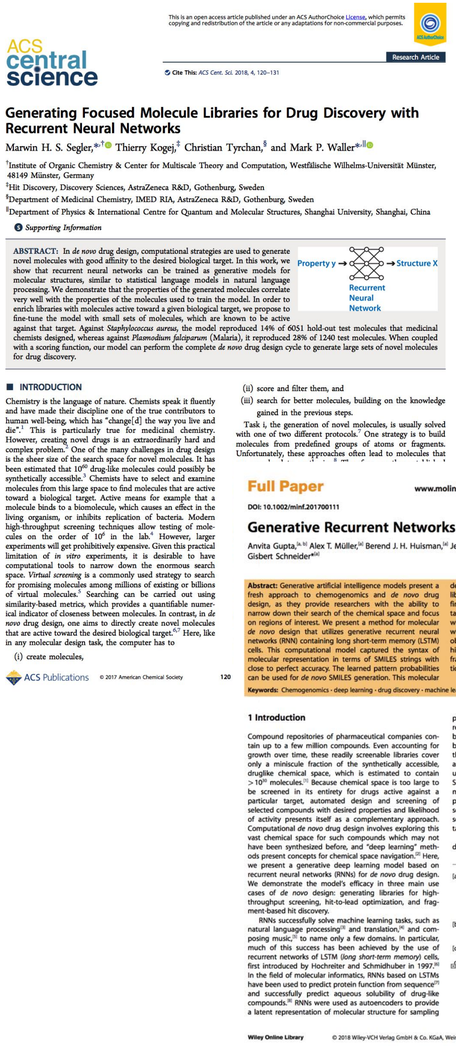
Received: December 2, 2017
Published: January 12, 2018

 © 2018 American Chemical Society

DOI: 10.1021/acscentsci.7b00572
ACS Cent. Sci. 2018, 4, 268−276

ARONTIER

Drug design & optimization



ACS central science

This is an open access article published under an ACS AuthorChoice License, which permits copying and redistribution of the article or any adaptations for non-commercial purposes.

Research Article

Cite This: ACS Cent. Sci. 2018, 4, 120−131

Generating Focused Molecule Libraries for Drug Discovery with Recurrent Neural Networks

Marwin H. S. Segler,*[,†] Thierry Kogej,[‡] Christian Tyrchan,[§] and Mark P. Waller*[,∥]

[†]Institute of Organic Chemistry & Center for Multiscale Theory and Computation, Westfälische Wilhelms-Universität Münster, 48149 Münster, Germany
[‡]Hit Discovery, Discovery Sciences, AstraZeneca R&D, Gothenburg, Sweden
[§]Department of Medicinal Chemistry, IMED RIA, AstraZeneca R&D, Gothenburg, Sweden
[∥]Department of Physics & International Centre for Quantum and Molecular Structures, Shanghai University, Shanghai, China

Supporting Information

ABSTRACT: In *de novo* drug design, computational strategies are used to generate novel molecules with good affinity to the desired biological target. In this work, we show that recurrent neural networks can be trained as generative models for molecular structures, similar to statistical language models in natural language processing. We demonstrate that the properties of the generated molecules correlate very well with the properties of the molecules used to train the model. In order to enrich libraries with molecules active toward a given biological target, we propose to fine-tune the model with small sets of molecules, which are known to be active against that target. Against *Staphylococcus aureus*, the model reproduced 14% of 6051 hold-out test molecules that medicinal chemists designed, whereas against *Plasmodium falciparum* (Malaria), it reproduced 28% of 1240 test molecules. When coupled with a scoring function, our model can perform the complete *de novo* drug design cycle to generate large sets of novel molecules for drug discovery.

Property y → Recurrent Neural Network → Structure X

■ INTRODUCTION

Chemistry is the language of nature. Chemists speak it fluently and have made their discipline one of the true contributors to human well-being, which has "change[d] the way you live and die".[1] This is particularly true for medicinal chemistry. However, creating novel drugs is an extraordinarily hard and complex problem.[2] One of the many challenges in drug design is the sheer size of the search space for novel molecules. It has been estimated that 10^{60} drug-like molecules could possibly be synthetically accessible.[3] Chemists have to select and examine molecules from this large space to find molecules that are active toward a biological target. Active means for example that a molecule binds to a biomolecule, which causes an effect in the living organism, or inhibits replication of bacteria. Modern high-throughput screening techniques allow testing of molecules on the order of 10^6 in the lab.[4] However, larger experiments will get prohibitively expensive. Given this practical limitation of *in vitro* experiments, it is desirable to have computational tools to narrow down the enormous search space. *Virtual screening* is a commonly used strategy to search for promising molecules among millions of existing or billions of virtual molecules.[5] Searching can be carried out using similarity-based metrics, which provides a quantifiable numerical indicator of closeness between molecules. In contrast, in *de novo* drug design, one aims to directly create novel molecules that are active toward the desired biological target.[6,7] Here, like in any molecular design task, the computer has to

(i) create molecules,
(ii) score and filter them, and
(iii) search for better molecules, building on the knowledge gained in the previous steps.

Task i, the generation of novel molecules, is usually solved with one of two different protocols.[7] One strategy is to build molecules from predefined groups of atoms or fragments. Unfortunately, these approaches often lead to molecules that

ACS Publications © 2017 American Chemical Society 120

Full Paper www.molinf.com molecular informatics

DOI: 10.1002/minf.201700111

Generative Recurrent Networks for *De Novo* Drug Design

Anvita Gupta,[a,b] Alex T. Müller,[a] Berend J. H. Huisman,[a] Jens A. Fuchs,[a] Petra Schneider,[a,c] and Gisbert Schneider*[a]

Abstract: Generative artificial intelligence models present a fresh approach to chemogenomics and *de novo* drug design, as they provide researchers with the ability to narrow down their search of the chemical space and focus on regions of interest. We present a method for molecular *de novo* design that utilizes generative recurrent neural networks (RNN) containing long short-term memory (LSTM) cells. This computational model captured the syntax of molecular representation in terms of SMILES strings with close to perfect accuracy. The learned pattern probabilities can be used for *de novo* SMILES generation. This molecular design concept eliminates the need for virtual compound library enumeration. By employing transfer learning, we fine-tuned the RNN's predictions for specific molecular targets. This approach enables virtual compound design without requiring secondary or external activity prediction, which could introduce error or unwanted bias. The results obtained advocate this generative RNN-LSTM system for high-impact use cases, such as low-data drug discovery, fragment based molecular design, and hit-to-lead optimization for diverse drug targets.

Keywords: Chemogenomics · deep learning · drug discovery · machine learning · medicinal chemistry

1 Introduction

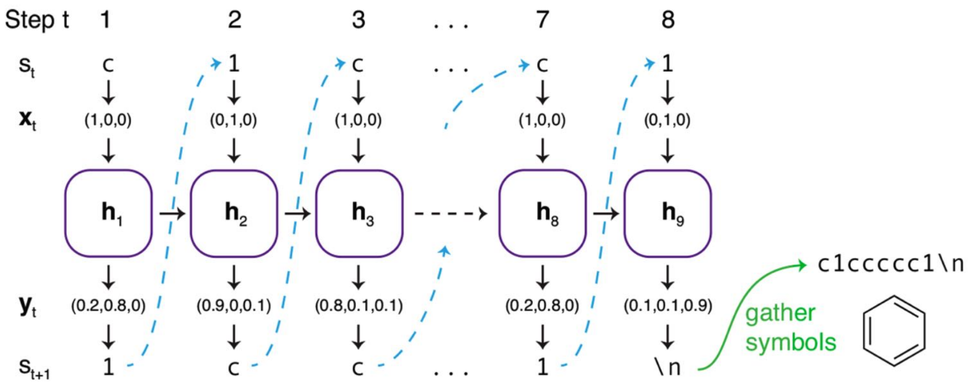
Compound repositories of pharmaceutical companies contain up to a few million compounds. Even accounting for growth over time, these readily screenable libraries cover only a miniscule fraction of the synthetically accessible, druglike chemical space, which is estimated to contain $>10^{60}$ molecules.[1] Because chemical space is too large to be screened in its entirety for drugs active against a particular target, automated design and screening of selected compounds with desired properties and likelihood of activity presents itself as a complementary approach. Computational *de novo* drug design involves exploring this vast chemical space to such compounds which may not have been synthesized before, and "deep learning" methods present concepts for chemical space navigation.[2] Here, we present a generative deep learning model based on recurrent neural networks (RNNs) for *de novo* drug design. We demonstrate the model's efficacy in three main use cases of *de novo* design: generating libraries for high-throughput screening, hit-to-lead optimization, and fragment-based hit discovery.

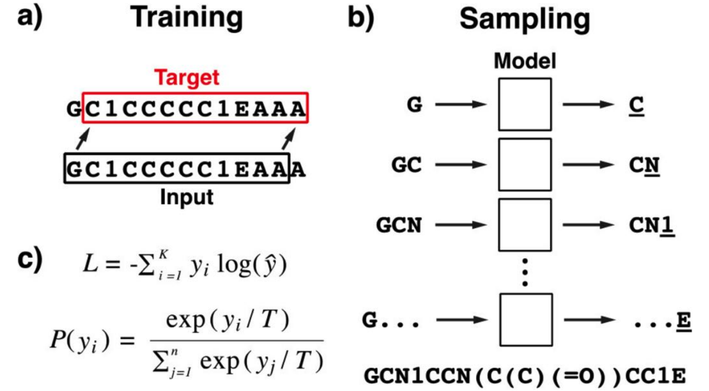
RNNs successfully solve machine learning tasks, such as natural language processing[3] and translation,[4] and composing music,[5] to name only a few domains. In particular, much of this success has been achieved by the use of recurrent networks of LSTM (*long short-term memory*) cells, first introduced by Hochreiter and Schmidhuber in 1997.[6] In the field of molecular informatics, RNNs based on LSTMs have been used to predict protein function from sequence[7] and successfully predict aqueous solubility of drug-like compounds.[8] RNNs were used as autoencoders to provide a latent representation of molecular structure for sampling preferred regions of chemical space.[9] Importantly, several research groups have recently demonstrated that RNNs can be employed to generate canonical SMILES strings, and can be fine-tuned by transfer learning.[10,11] In transfer learning, the machine learning model tries to keep information from a previously learned task to solve a different but related, yet unseen task.[12] Researchers at AstraZeneca have extended SMILES-generating RNNs by using this concept for reinforcement learning. The model's parameters were optimized to produce strings that scored highly according to an external scoring function. They applied this approach to generate sets of structures with low sulfur content, high predicted target activity, and other desirable properties.[13]

We here present a new approach to *de novo* drug design using RNN deep learning methodology (Figure 1). In

[a] A. Gupta, A. T. Müller, B. J. H. Huisman, J. A. Fuchs, P. Schneider, G. Schneider
Swiss Federal Institute of Technology (ETH), Department of Chemistry and Applied Biosciences, Vladimir−Prelog−Weg 4, 8093 Zurich, Switzerland
E-mail: gisbert@ethz.ch
[b] A. Gupta
Stanford University, Department of Computer Science, 450 Sierra Mall, Stanford, CA, 94305, USA
[c] P. Schneider
inSilicam GmbH, 8049 Zurich, Switzerland
© 2017 The Authors. Published by Wiley-VCH Verlag GmbH & Co. KGaA.
This is an open access article under the terms of the Creative Commons Attribution Non-Commercial NoDerivs License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

- Language models based on characters or letters.
  - Swap words or letters with atoms, or characters in the SMILES.
- RNN + LSTM
  - A network with three stacked LSTM layers, using the Keras library
    - Each 1024 dimensions, and each one followed by a dropout layer, with a dropout ratio of 0.2.
    - Batch size 128
    - The RNN was unrolled for 64 steps. (21.3x10^6 parameters)
- The model was trained with back-propagation through time, using the ADAM optimizer at standard settings. A gradient norm clipping of 5 is applied

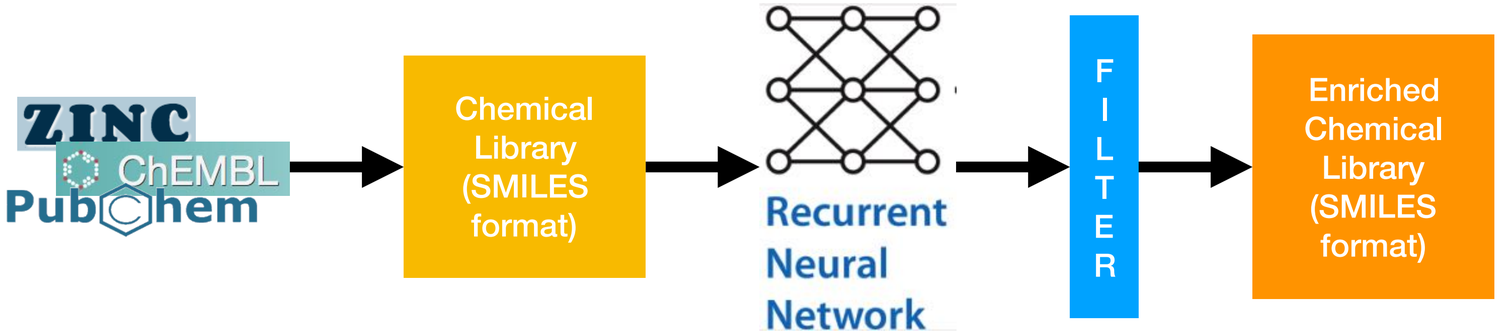




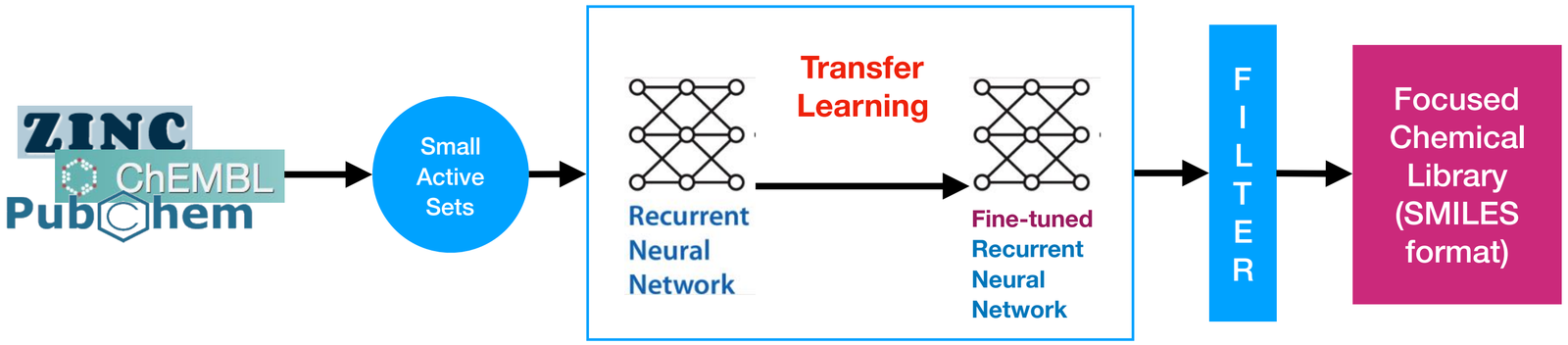
ARONTIER

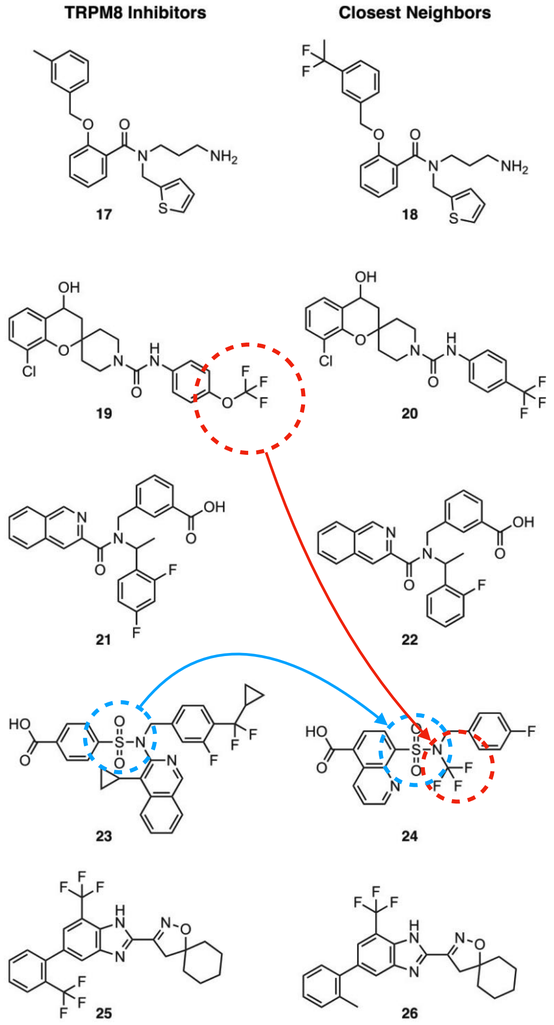
Drug design & optimization

To generate large sets of diverse molecules for virtual screening campaigns



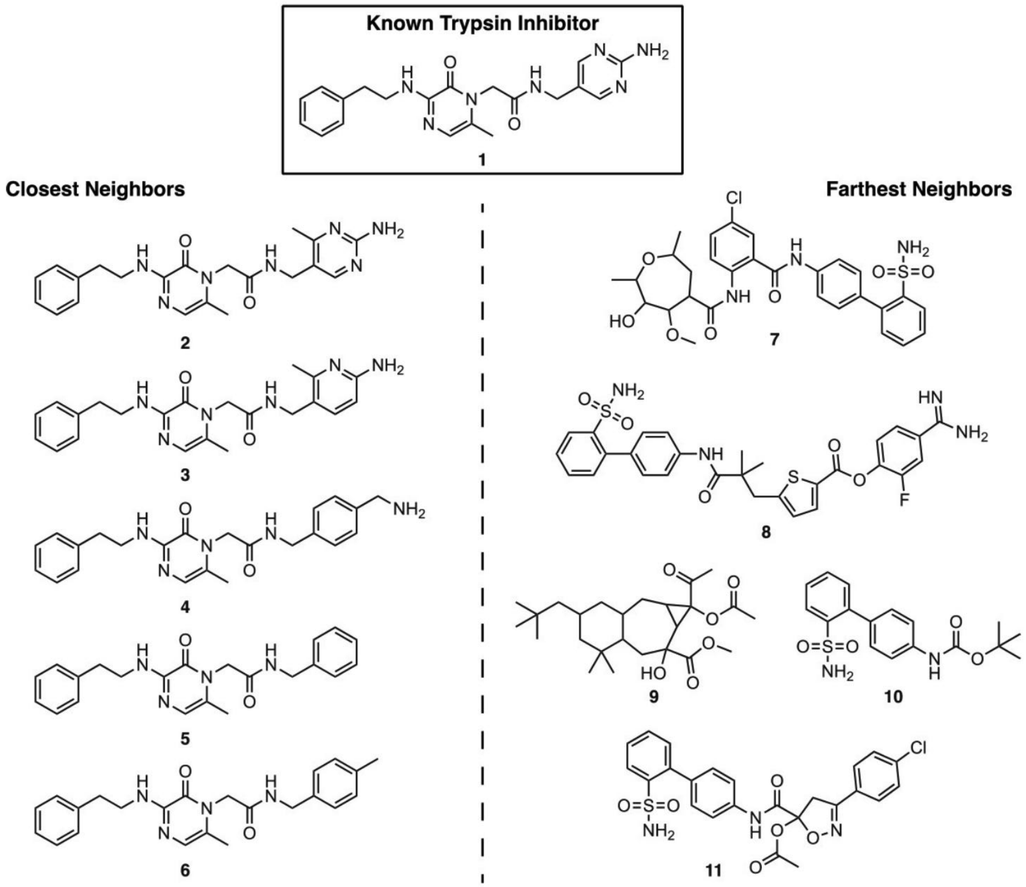
To generate smaller, focused libraries enriched with possibly active molecules for specific target



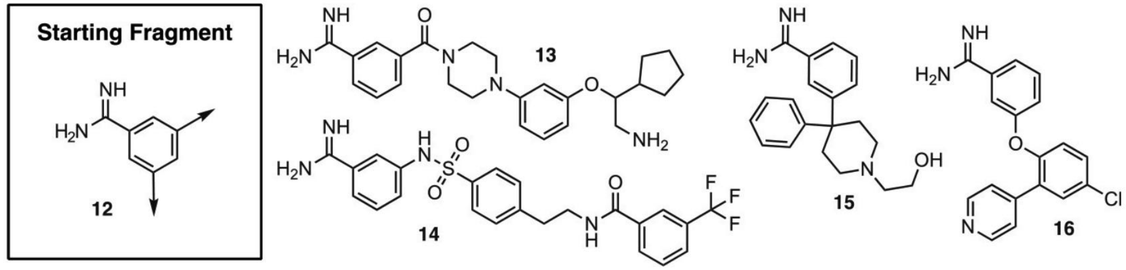


ARONTIER

Drug design & optimization



Modifications to known inhibitors

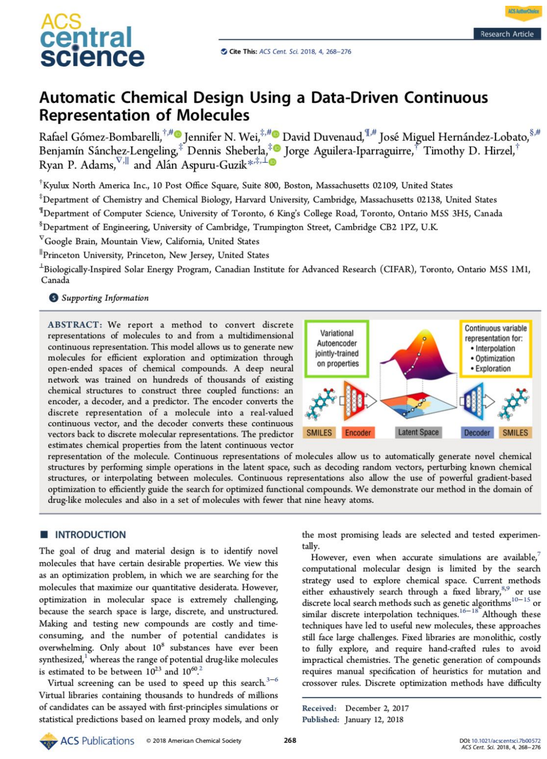


Fragment-growing

ARONTIER

최적화

- AutoEncoder를 이용한 화합물의 Encoding & Decoding
  - Compounds -> 116 dimension vector
- Predictor를 이용한 최적화
  - LogP, QED



ACS central science

Research Article

Cite This: ACS Cent. Sci. 2018, 4, 268−276

Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules

Rafael Gómez-Bombarelli,[†,#] Jennifer N. Wei,[‡,#] David Duvenaud,[¶,#] José Miguel Hernández-Lobato,[§,#] Benjamín Sánchez-Lengeling,[‡] Dennis Sheberla,[‡] Jorge Aguilera-Iparraguirre,[†] Timothy D. Hirzel,[†] Ryan P. Adams,[∇,∥] and Alán Aspuru-Guzik*[,⊥,□]

[†]Kyulux North America Inc., 10 Post Office Square, Suite 800, Boston, Massachusetts 02109, United States
[‡]Department of Chemistry and Chemical Biology, Harvard University, Cambridge, Massachusetts 02138, United States
[§]Department of Computer Science, University of Toronto, 6 King's College Road, Toronto, Ontario M5S 3H5, Canada
[¶]Department of Engineering, University of Cambridge, Trumpington Street, Cambridge CB2 1PZ, U.K.
[∇]Google Brain, Mountain View, California, United States
[∥]Princeton University, Princeton, New Jersey, United States
[⊥]Biologically-Inspired Solar Energy Program, Canadian Institute for Advanced Research (CIFAR), Toronto, Ontario M5S 1M1, Canada

Supporting Information

ABSTRACT: We report a method to convert discrete representations of molecules to and from a multidimensional continuous representation. This model allows us to generate new molecules for efficient exploration and optimization through open-ended spaces of chemical compounds. A deep neural network was trained on hundreds of thousands of existing chemical structures to construct three coupled functions: an encoder, a decoder, and a predictor. The encoder converts the discrete representation of a molecule into a real-valued continuous vector, and the decoder converts these continuous vectors back to discrete molecular representations. The predictor estimates chemical properties from the latent continuous vector representation of the molecule. Continuous representations of molecules allow us to automatically generate novel chemical structures by performing simple operations in the latent space, such as decoding random vectors, perturbing known chemical structures, or interpolating between molecules. Continuous representations also allow the use of powerful gradient-based optimization to efficiently guide the search for optimized functional compounds. We demonstrate our method in the domain of drug-like molecules and also in a set of molecules with fewer that nine heavy atoms.

INTRODUCTION

The goal of drug and material design is to identify novel molecules that have certain desirable properties. We view this as an optimization problem, in which we are searching for the molecules that maximize our quantitative desiderata. However, optimization in molecular space is extremely challenging, because the search space is large, discrete, and unstructured. Making and testing new compounds are costly and time-consuming, and the number of potential candidates is overwhelming. Only about 10^8 substances have ever been synthesized,[1] whereas the range of potential drug-like molecules is estimated to be between 10^{23} and 10^{60}.[2]

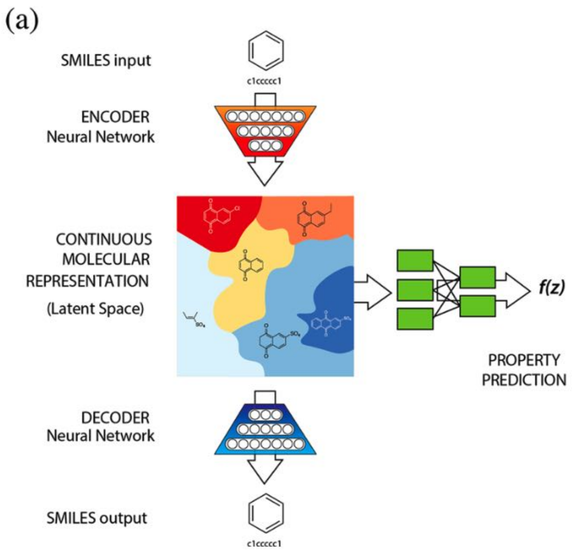
Virtual screening can be used to speed up this search.[3−6] Virtual libraries containing thousands to hundreds of millions of candidates can be assayed with first-principles simulations or statistical predictions based on learned proxy models, and only the most promising leads are selected and tested experimentally.

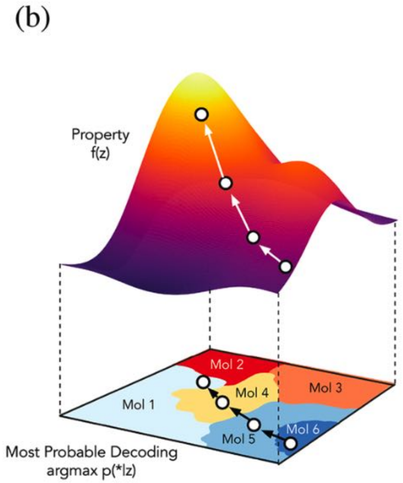
However, even when accurate simulations are available,[7] computational molecular design is limited by the search strategy used to explore chemical space. Current methods either exhaustively search through a fixed library,[8,9] or use discrete local search methods such as genetic algorithms[10−13] or similar discrete interpolation techniques.[14−19] Although these techniques have led to useful new molecules, these approaches still face large challenges. Fixed libraries are monolithic, costly to fully explore, and require hand-crafted rules to avoid impractical chemistries. The genetic generation of compounds requires manual specification of heuristics for mutation and crossover rules. Discrete optimization methods have difficulty

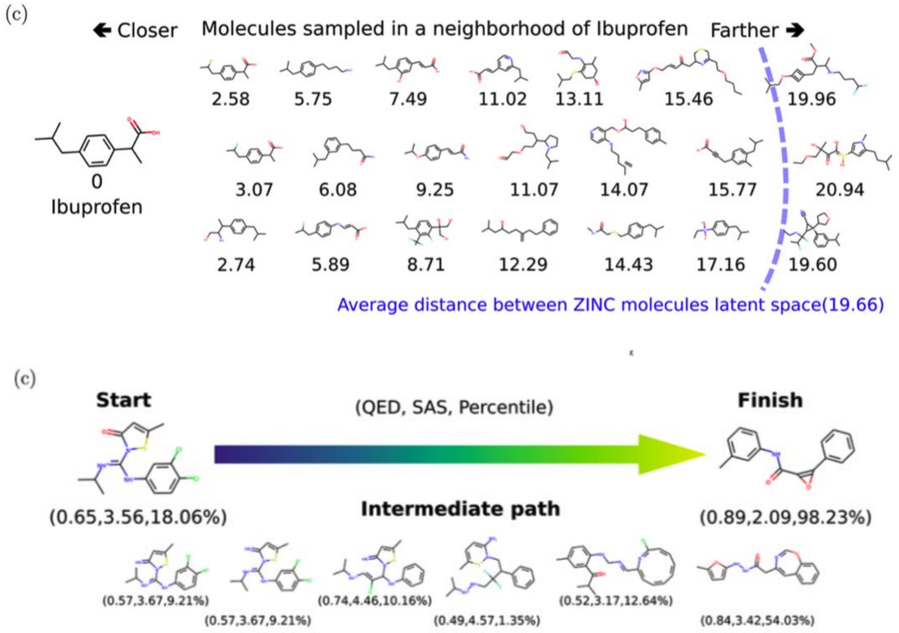
Received: December 2, 2017
Published: January 12, 2018

 © 2018 American Chemical Society

DOI: 10.1021/acscentsci.7b00572
ACS Cent. Sci. 2018, 4, 268−276

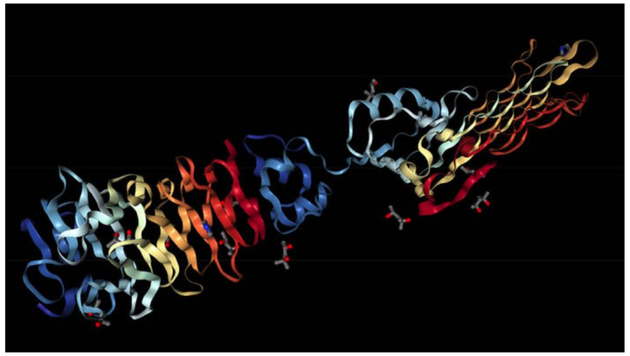






ARONTIER

구조 예측 : Deepmind AlphaFold



Complex of bacteria-infecting viral proteins modeled in CASP 13. The complex contains four separate subunits that were modeled individually. PROTEIN DATA BANK

Google's DeepMind aces protein folding

By **Robert F. Service** | Dec. 6, 2018 , 12:05 PM

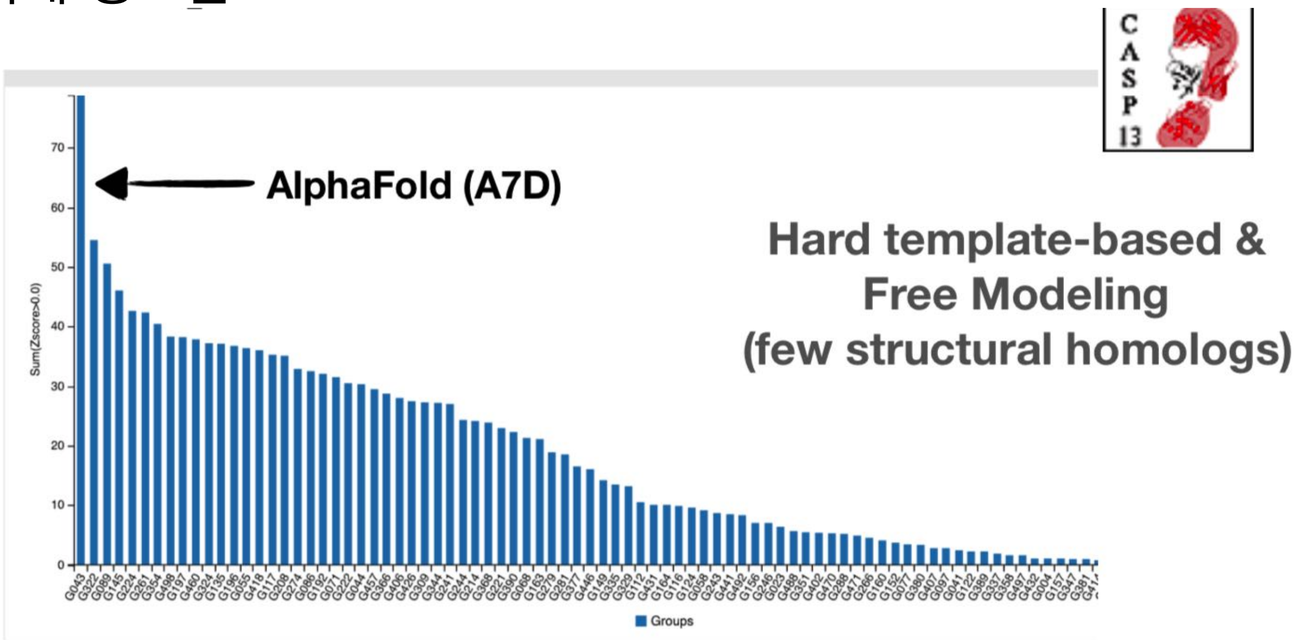
Turns out mastering chess and Go was just for starters. On 2 December, the Google-owned artificial intelligence firm DeepMind took top honors in the 13th Critical Assessment of Structure Prediction (CASP), a biannual competition aimed at predicting the 3D structure of proteins.

The contest worked like this: Competing teams were given the linear sequence of amino acids for 90 proteins for which the 3D shape is known but not yet published. Teams then computed how those sequences would fold. Though London-based DeepMind had not previously joined this competition, the predictions of its AlphaFold software were, on average, more accurate than those of its 97 competitors.

How close was the race? By one metric, not very. For protein sequences for which no other information was known—43 of the 90—AlphaFold made the most accurate prediction 25 times. That far outpaced the second place finisher, which won three of the 43 tests.

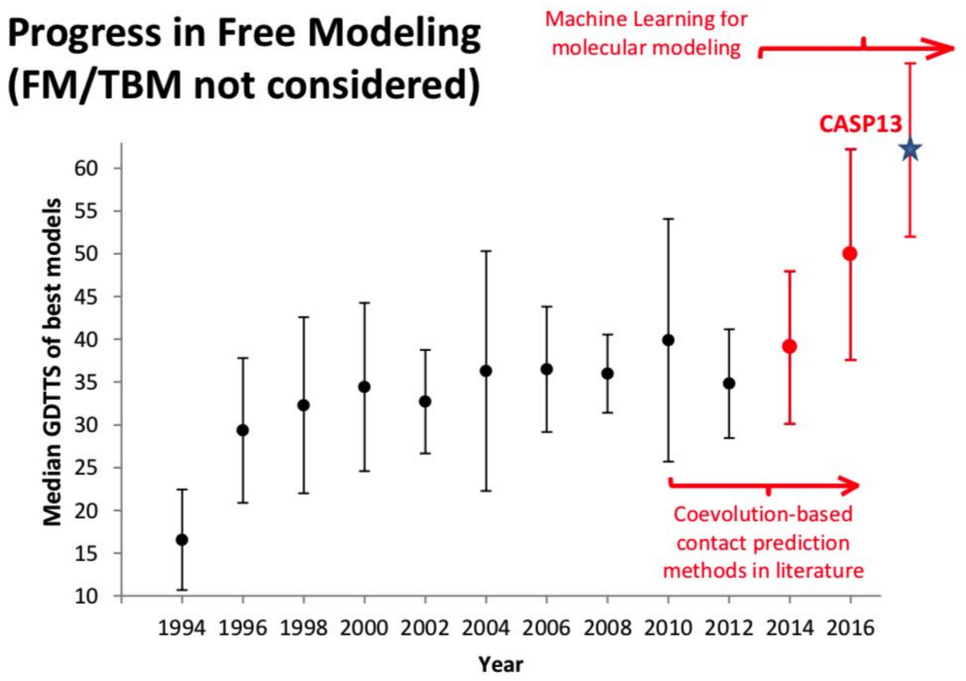
SIGN UP FOR OUR DAILY NEWSLETTER
Get more great content like this delivered right to you!

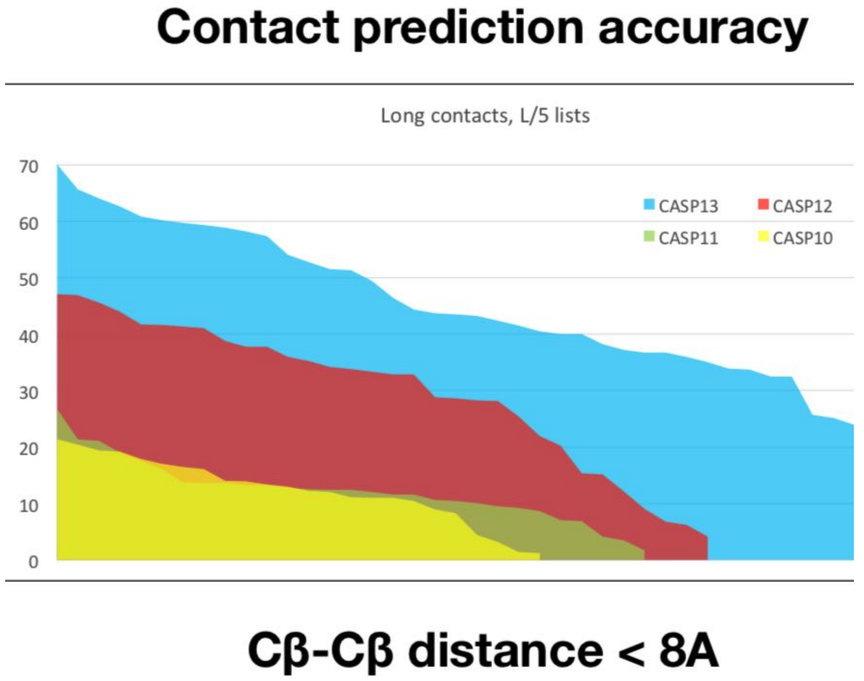
- AlphaGo를 만들었던 Google Deepmind에서 개발한 인공지능 기반의 단백질 구조 예측 프로그램
- CASP13에서 최초로 참가하여 압도적인 정확도로 많은 예측 그룹들 중 1위에 랭크됨



#	GR code	GR name	Domains Count	SUM Zscore (>-2.0)	Rank SUM Zscore (>-2.0)	AVG Zscore (>-2.0)	Rank AVG Zscore (>-2.0)	SUM Zscore (>0.0)	Rank SUM Zscore (>0.0)
1	043	A7D	43	78.8266	1	1.8332	1	78.8266	1
2	322	Zhang	43	53.9247	2	1.2541	2	54.5247	2
3	089	MULTICOM	43	50.0762	3	1.1646	3	50.5987	3

AlphaFold



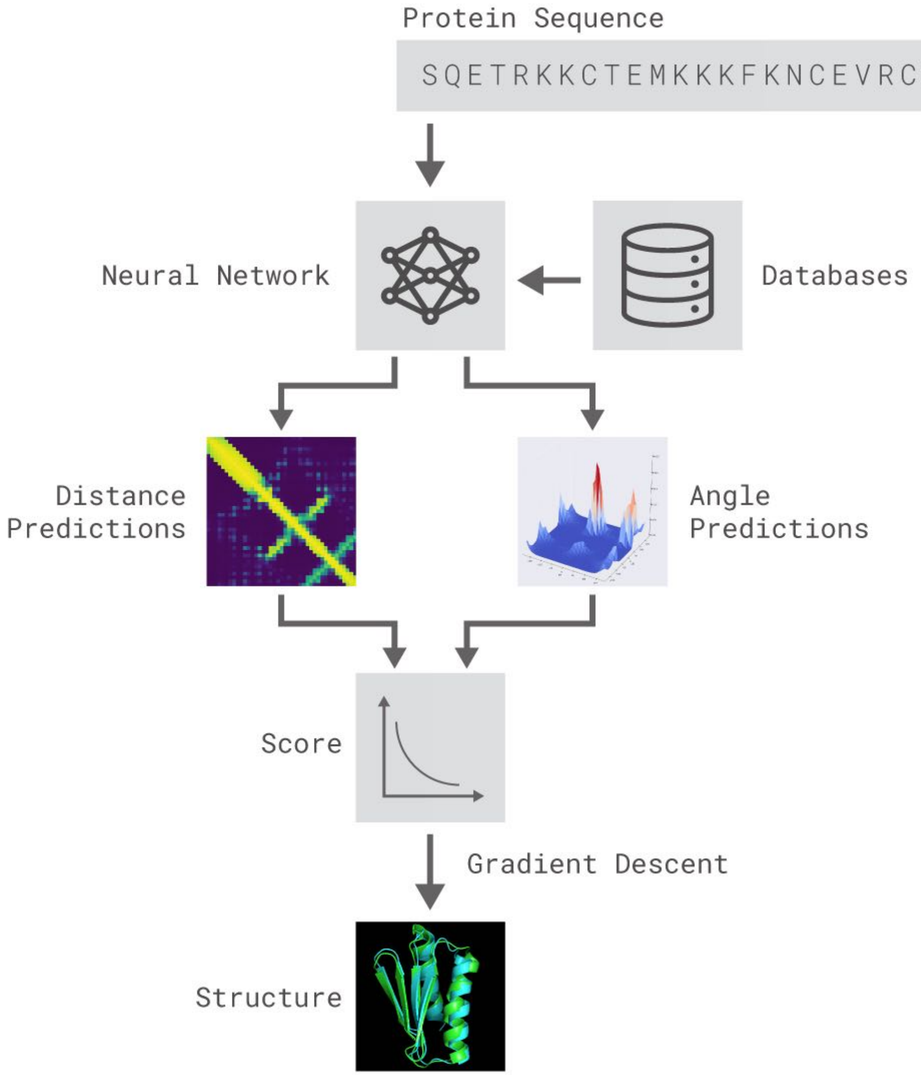


- During the last two CASPs, there were significant improvements in protein modeling, especially in contact prediction

강원대학교 화학과 이주용 교수님 발표자료 (2018, 아론티어)

ARONTIER

AlphaFold 구조



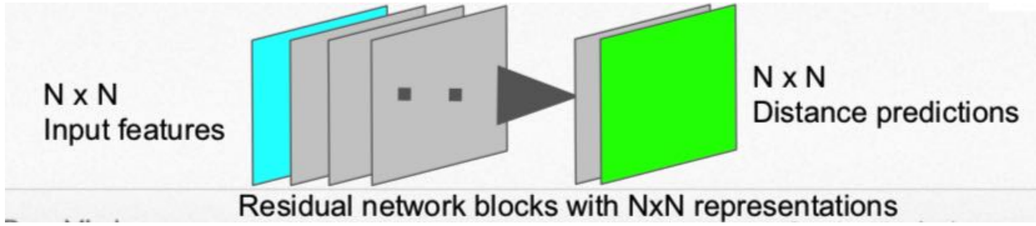
- Alignment(MSA)를 얻음
- MSA에서 co-evolution 정보를 얻어냄
- Co-evolution 정보와 PDB에 있는 단백질의 3차원 구조에서 얻은 거리 정보를 기반으로 residue-residue distance를 예측하는 인공 지능을 학습시킴
- 이를 바탕으로 구조를 모르는 단백질의 서열이 주어졌을 때 residue사이의 거리를 예측
- 예측된 거리를 바탕으로 단백질의 구조를 생성

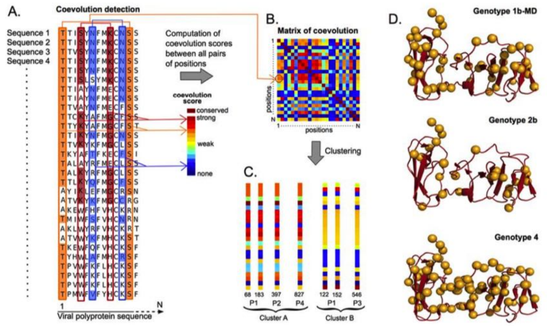
강원대학교 화학과 이주용 교수님 발표자료 (2018, 아론티어)

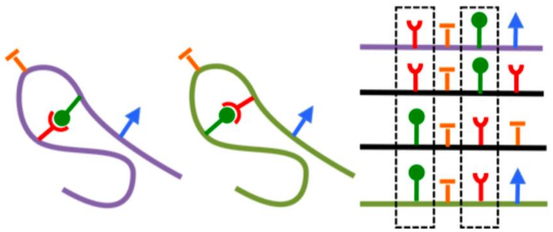
ARONTIER

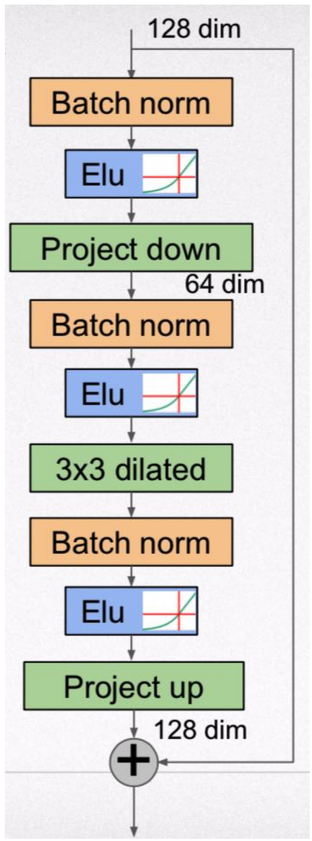
AlphaFold : Co-evolution 기반의 Res.-Res. Contact Prediction

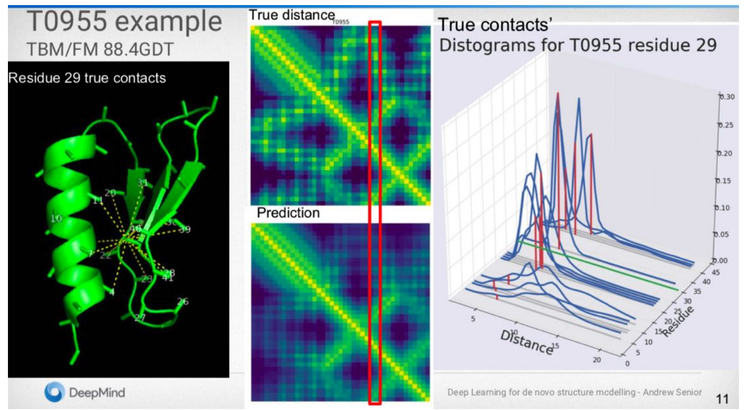
- 기존의 단백질 구조 예측 파이프라인 중 Residue-Residue Contact Prediction 부분을 CNN 기반의 예측 시스템을 구성하여 높은 정확도의 인공지능 모델을 만드는데 성공함.
  - 660 layers 모델 사용.
  - 4개의 독립적으로 학습된 모델을 이용

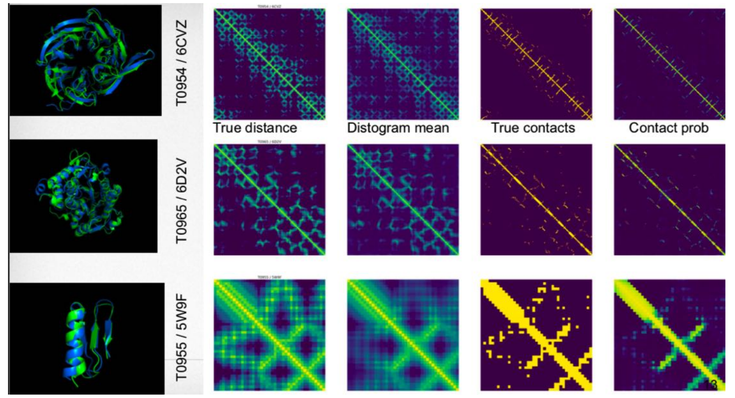












강원대학교 화학과 이주용 교수님 발표자료 (2018, 아론티어)

A.W. Senior, et al, (2018, CASP13 Conference)

ARONTIER

다른 SBDD 기술들

Deep learning 기반의 Binding Affinity 예측

K_{DEEP}: Protein–Ligand Absolute Binding Affinity Prediction via 3D-Convolutional Neural Networks

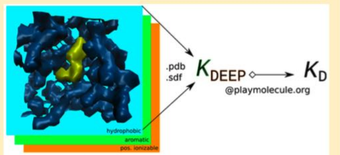
José Jiménez,[†] Miha Škalič,[†] Gerard Martínez-Rosell,[†] and Gianni De Fabritiis*[,†,‡]

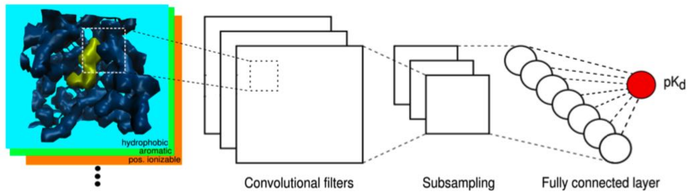
[†]Computational Biophysics Laboratory, Universitat Pompeu Fabra, Parc de Recerca Biomèdica de Barcelona, Carrer del Dr. Aiguader 88, Barcelona 08003, Spain

[‡]Institució Catalana de Recerca i Estudis Avançats (ICREA), Passeig Lluis Companys 23, 08010 Barcelona, Spain

Supporting Information

ABSTRACT: Accurately predicting protein–ligand binding affinities is an important problem in computational chemistry since it can substantially accelerate drug discovery for virtual screening and lead optimization. We propose here a fast machine-learning approach for predicting binding affinities using state-of-the-art 3D-convolutional neural networks and compare this approach to other machine-learning and scoring methods using several diverse data sets. The results for the standard PDBbind (v.2016) core test-set are state-of-the-art with a Pearson's correlation coefficient of 0.82 and a RMSE of 1.27 in pK units between experimental and predicted affinity, but accuracy is still very sensitive to the specific protein used. K_{DEEP} is made available via PlayMolecule.org for users to test easily their own protein–ligand complexes, with each prediction taking a fraction of a second. We believe that the speed, performance, and ease of use of K_{DEEP} makes it already an attractive scoring function for modern computational chemistry pipelines.





Deep learning 기반의 Pocket 특성 분석

Bioinformatics

ISCB INTERNATIONAL SOCIETY FOR COMPUTATIONAL BIOLOGY

Article Navigation

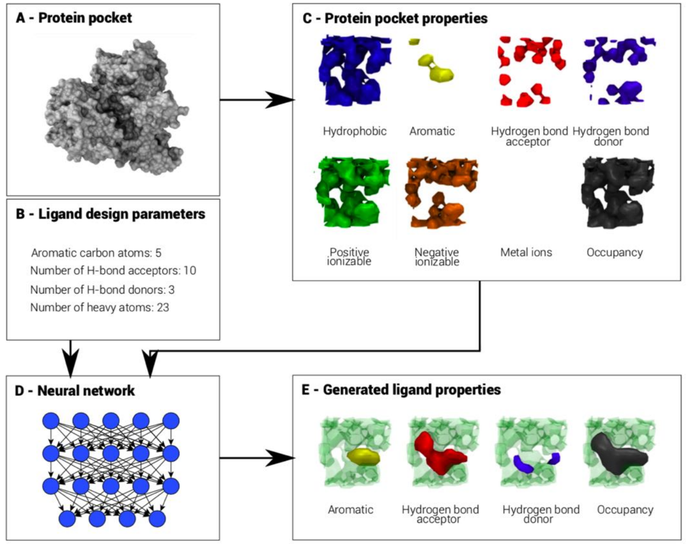
LigVoxel: inpainting binding pockets using 3D-convolutional neural networks

Miha Skalic, Alejandro Varela-Rial, José Jiménez, Gerard Martínez-Rosell, Gianni De Fabritiis

Bioinformatics, Volume 35, Issue 2, 15 January 2019, Pages 243–250, https://doi.org/10.1093/bioinformatics/bty583

Published: 06 July 2018 **Article history**

Views Cite Permissions Share



독성 예측

ORIGINAL RESEARCH ARTICLE

Front. Environ. Sci., 02 February 2016 | https://doi.org/10.3389/fenvs.2015.00080

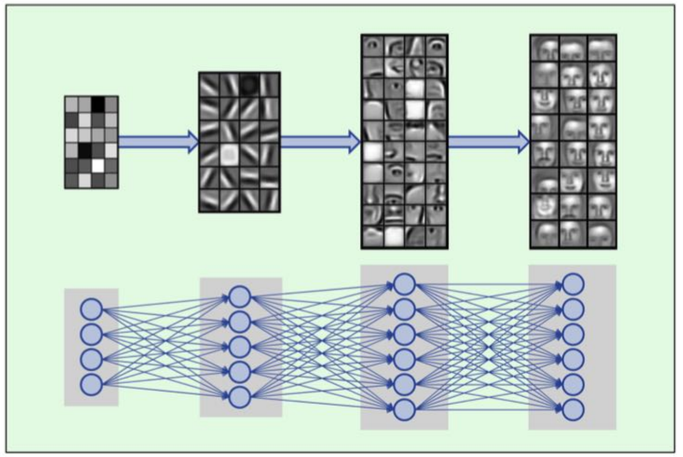
DeepTox: Toxicity Prediction using Deep Learning

Andreas Mayr[1,2†], Günter Klambauer[1†], Thomas Unterthiner[1,2†] and Sepp Hochreiter[1*]

[1]Institute of Bioinformatics, Johannes Kepler University Linz, Linz, Austria

[2]RISC Software GmbH, Johannes Kepler University Linz, Hagenberg, Austria

The Tox21 Data Challenge has been the largest effort of the scientific community to compare computational methods for toxicity prediction. This challenge comprised 12,000 environmental chemicals and drugs which were measured for 12 different toxic effects by specifically designed assays. We participated in this challenge to assess the performance of Deep Learning in computational toxicity prediction. Deep Learning has already revolutionized image processing, speech recognition, and



ARONTIER

END