Drug Discovery Today • Volume xxx, Number xx • xxxx 2021 REVIEWS

Artificial intelligence-enhanced drug design and development: Toward a computational precision medicine

Philippe Moingeon *, **Mélaine Kuenemann, Mickaël Guedj**

Servier, Research and Development, 50 rue Carnot, 92284 Suresnes Cedex, France

Artificial Intelligence (AI) relies upon a convergence of technologies with further synergies with life science technologies to capture the value of massive multi-modal data in the form of predictive models supporting decision-making. AI and machine learning (ML) enhance drug design and development by improving our understanding of disease heterogeneity, identifying dysregulated molecular pathways and therapeutic targets, designing and optimizing drug candidates, as well as evaluating *in silico* clinical efficacy. By providing an unprecedented level of knowledge on both patient specificities and drug candidate properties, AI is fostering the emergence of a computational precision medicine allowing the design of therapies or preventive measures tailored to the singularities of individual patients in terms of their physiology, disease features, and exposure to environmental risks.

Keywords: Artificial Intelligence; Big data; Computational precision medicine; Disease model; Drug discovery & development; Machine learning

Introduction

Drug development is a complex process that currently requires, on average, 12 years of development and a US$2.6 billion investment per individual drug made available to the patient.[1] Stringent regulatory requisites to demonstrate drug efficacy and safety result in a high attrition rate because of negative results during their evaluation in costly clinical studies, with an estimated 6.2% of drugs selected in the discovery phase eventually made available to patients.[2,3] In this context, **AI**-based predictive modeling (see Glossary) is emerging as a revolutionary solution to improve both the efficacy and speed of drug design and development, most particularly by optimizing early on the choice of therapeutic targets as well as drug candidates.[3–5] AI can be defined as a convergence of technologies recapitulating four dimensions of human intelligence (i.e., sensing, thinking, acting, and learning). As such, AI allows the integration of massive amounts of multi-modal data, both structured and unstructured, to build up probabilistic and dynamic models of a problem.

As it applies to drug development, AI-driven predictive models can be generated by using specific sets of data to inform a series of decisions taken throughout drug discovery, development, and registration steps (Fig. 1). These steps include selecting the right therapeutic target, the optimal drug candidate, the appropriate dosing and administration regimens, as well as the appropriate patients to include in clinical studies.[3,5] By providing a means to capture the value of data related to diagnosis, patient characterization, drug candidate attributes, and prediction of individual responses to therapy, AI enables a more personalized approach, termed 'precision medicine', that is proposing treatments better tailored to individual patient specificities.[6,7]

On this basis, we discuss herein four main applications of AI to support drug design and development: (i) the generation of disease models based on molecular profiling data from patients to represent disease heterogeneity; (ii) the identification of dysregulated molecular pathways and of candidate therapeutic targets predicted to contribute to disease causality; (iii) the design, synthesis, and optimization of drug candidates interacting with

* *Corresponding author.* Moingeon, P. (philippe.moingeon@servier.com)

https://doi.org/10.1016/j.drudis.2021.09.006

Please cite this article in press as: P. Moingeon et al., Drug Discovery Today (2021), https://doi.org/10.1016/j.drudis.2021.09.006

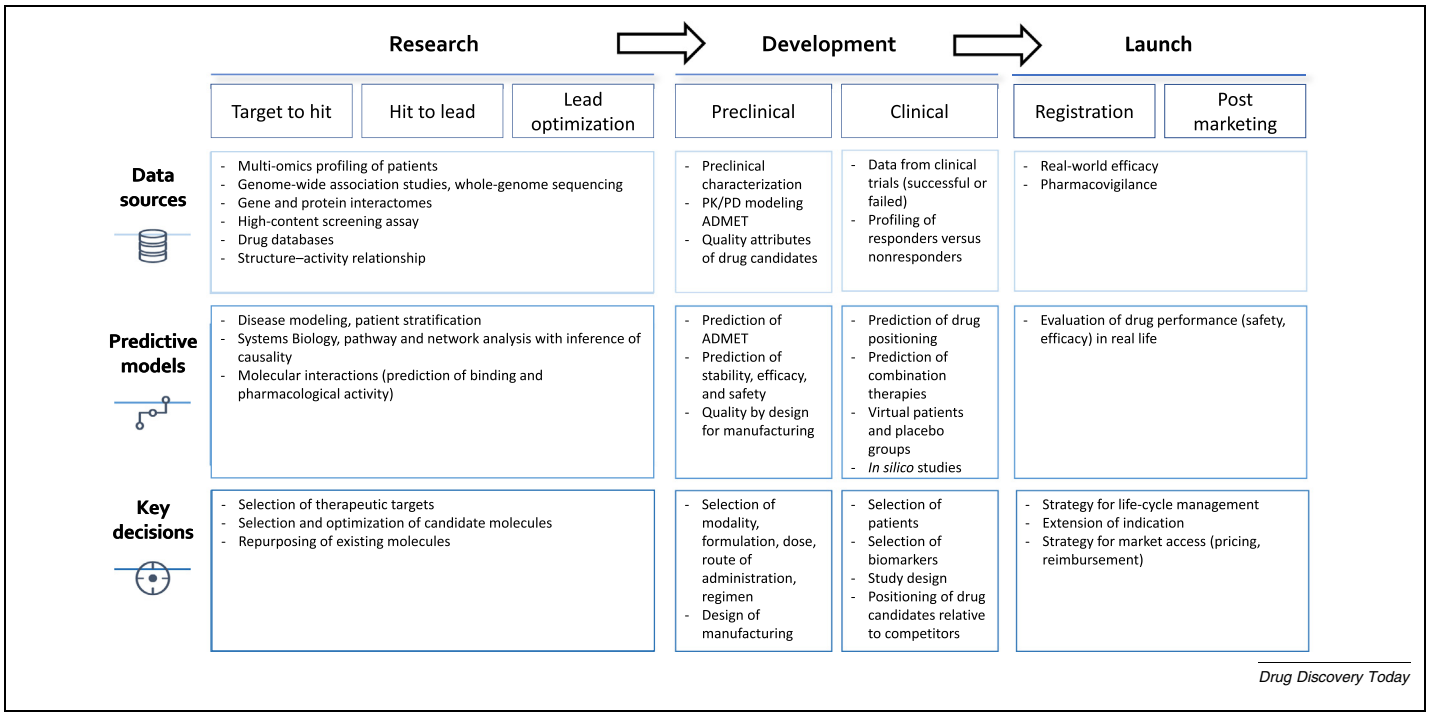

FIGURE 1

Decision-making during drug discovery, development, and registration. Key decisions to be taken as well as predictive models and examples of data sets supporting those models are provided for the drug discovery, development, and registration phases. Abbreviations: ADMET, absorption, distribution, metabolism, excretion, and toxicity; PD, pharmacodynamics; PK, pharmacokinetics.

these targets; and (iv) the evaluation of clinical efficacy by using virtual patients or real-world evidence data.

Capturing the value of big biomedical data

The recent and rapid advances in next-generation DNA, RNA, and exome sequencing, multi-omics molecular profiling, high-resolution medical imaging, and electronic capture technologies make it possible to characterize at an unprecedented level the specificities of individuals in terms of their physiology, pathophysiology of their disease as well as their environmental risk exposure. The Cancer Genome Atlas (TCGA), the Alzheimer's Disease Neuroimaging Initiative (ADNI), the Osteoarthritis Initiative (OAI), and the UK Biobank projects are all examples of this growing trend to integrate big data from large patient populations to support drug development (Table S1 in the supplemental information online). In the near future, such comprehensive molecular information will be available for millions of patients across multiple diseases, together with exponential data and knowledge compiled within hundreds of structured biomedical databases, such as those managed by the European Bioinformatics Institute (EBI) or the US National Center for Biotechnology Information (NCBI) (Table S1 in the supplemental information online).

When trying to capture the value of those ever-increasing amounts of data, the main challenges are linked to proper access and selection of standardized and machine-readable data, as well as to data complexity, heterogeneity, and **sparsity**. Integrating massive and multi-modal data generated from multiple technologies with proper quality attributes in terms of consistency and reliability remains a significant difficulty in data life-cycle management (Fig. 2). Access to accurate and curated data in large quantities is also key to improve **ML** repeatability. Solving those problems requires setting up computing hardware architectures adapted to life sciences specificities (Table S2 in the supplemental information online), often deported into the cloud. To this end, many initiatives, such as the Clinical Data Interchange Standards Consortium (CDISC)[8] or the FAIR guiding principles,[9] have emerged to enable the findability, accessibility, interoperability, reusability, and exchange of data.[10] In addition, the regulatory requirements imposed in terms of access, storage, sharing of confidential and sensitive health data by the European General Data Protection Regulation (GDPR)[11] and the US Health Information Technology for Economic and Clinical Act impose the implementation of clear and operational data governance strategies (Fig. 2).

In this context, precompetitive collaborative consortia between pharmaceutical companies or academic labs, such as MELLODDY[12] or the Drug Target Commons,[13] respectively, constitute innovative federated knowledge initiatives to assemble, curate, and share massive data of an appropriate quality for developing ML algorithms. The MELLODY consortium brings together several drug companies sharing their chemical libraries to train multitask predictive algorithms, subsequently applied by each individual partner in support of its own drug discovery program. In parallel, multiple crowd-source challenges, such as the Kaggle,[14] Dream,[15] and PrecisionFDA,[16] challenges propose data sets of reference to establish standards for benchmarking and testing novel algorithms to address complex biomedical problems.[6]

Please cite this article in press as: P. Moingeon et al., Drug Discovery Today (2021), https://doi.org/10.1016/j.drudis.2021.09.006

AI and disease modeling

The convergence of biotechnologies and AI provides an opportunity to create disease models to help positioning therapies in well-defined patient subpopulations. Such models are generated following extensive molecular profiling of patients compared with healthy controls using multi-omics technologies to represent diseases as endotypes defined based upon underlying pathophysiological mechanisms.[17] These data are classically produced during the follow-up of large cohorts of patients by public–private partnerships, with patient stratification being performed by using a combination of **unsupervised** and **supervised learning** approaches. The rationale for such a clustering, as a substitute to former classifications based solely upon clinical phenotypes, is that it better supports a precision medicine approach relying upon therapies targeted to well-defined subgroups of patients.[18] To this aim, molecular profiling data obtained in the blood and/or target organs of thousands of patients with a given disease are combined with detailed clinical information in terms of disease progression, severity, or response to treatments to stratify patients in homogeneous subgroups reflecting disease heterogeneity. Whereas the integration of such massive and multi-modal data is not possible with conventional bioinformatics, a comprehensive modeling of diseases can now be made by using AI.[19]

To do so, the main computational challenges still lie in the ability to: (i) integrate data coming from multi-omics technologies while reducing the multiplicity of their dimensions[20,21]; (ii) decipher disease mechanisms at a single cell level[22,23]; (iii) model the dynamic evolution of the disease[24]; and (iv) consolidate the findings through **consensus** and resampling approaches to support their validity and replication.[25] Following gene set enrichment analyses within each cluster, patient subgroups can then be further characterized in terms of molecular pathways being dysregulated.[26] Specific databases (e.g., Ingenuity Pathway Analysis and STRING) are used to regroup within established functional molecular pathways genes or proteins that are either up- or downregulated in patient samples compared with healthy controls. Given that a disease is defined in molecular terms in reference to normality, disease features need to be identified beyond molecular polymorphisms observed in association with the healthy state.

Overall, disease modeling can provide information on both the natural history of the disease and the relationship between pathophysiological mechanisms involved at both systemic and organ-specific levels. Furthermore, it sheds light on patient heterogeneity as well as on molecular signatures, which can be used to cluster patients in homogeneous groups to envision a precision medicine approach, taking into account patient specificities within clusters. Importantly, it also provides clues for further *in silico* identification of targets of therapeutic interest.

Identification, prioritization, and validation of therapeutic targets

Computational methods are being developed to identify disease-associated genes or proteins predicted to be involved in the causality of the disease, thus representing potential actionable therapeutic targets. As a first step, molecular pathways dysregulated in a given disease are represented in large-scale **networks** of interconnected genes or proteins, either established from protein–protein interactions (PPIs)[27] or reconstructed via inference techniques, such as correlation or Bayesian networks.[28,29] Such networks are also often referred to as **knowledge graphs**, representing knowledge as both concepts and relationships between them. This representation allows the delineation of disease-associated subnetwork modules serving as a basis for further computational analyses of their intrinsic topology to identify nodes predicted as 'causal' (including, for example, master regulators, hubs, and driver mutations).[19,30] In particular, **network propagation** algorithms (also referred to as diffusion) are commonly used to amplify the signal of nodes for which little or no direct evidence of disease association is available.[31] As above, main computational challenges involve the integration of multi**layer** networks obtained from different levels,[32] as well as the representation of large-scale dynamic information.[33]

Besides biological relevance, additional dimensions are considered to prioritize disease targets for further investigation, as

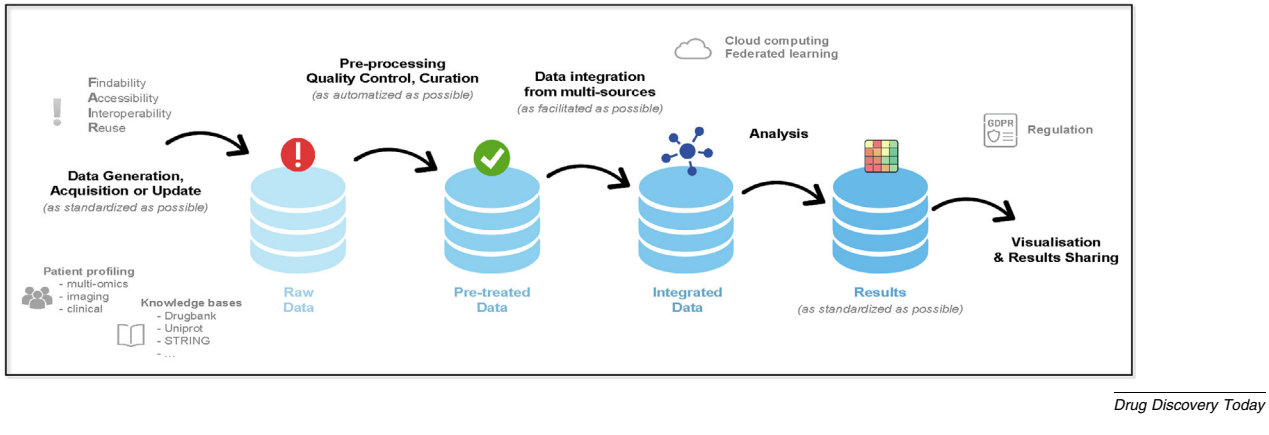

FIGURE 2

Biomedical data life-cycle management. Representation of a general biomedical data life cycle from data generation to the sharing of results, with emphasis on the needs for more standardization and automation in data governance.

Please cite this article in press as: P. Moingeon et al., Drug Discovery Today (2021), https://doi.org/10.1016/j.drudis.2021.09.006

illustrated by the Open Targets initiative.[34] This includes: (i) druggability (i.e., the likelihood of being able to modulate the function of a target with a small synthetic or biological drug, or any other therapeutic modality)[35]; (ii) potential safety implications when interfering with this target[36]; (iii) innovativeness documented from patent and literature mining by using **natural language processing** (NLP)[37]; and (iv) feasibility of drug development.[38] The selection of targets can be facilitated by the identification of features based on protein structure or sequence suggesting that they can bind small molecules.[4] Thus, the confirmation of target druggability significantly benefits from advances in 3D-structure modeling,[39] including the recent improved prediction of protein structures by DeepMind's AlphaFold algorithm based on primary amino acid sequences.[40]

Candidate targets identified with inferences of causality in the disease using network computing approaches need to be validated on the basis of empirical evidence generated in wet lab experiments. This validation step, which includes, for example, CRISPR-Cas9 gene deletion or siRNA gene silencing, phenotyping assessment of target expression on cells or tissues from patients relative to healthy controls or functional assays conducted in animal models, can be substantially simplified when using computational predictive models. As a result, both cost and timelines associated with drug discovery are reduced, while strengthening the rationale for selecting the candidate target before entering clinical development.

AI-enhanced drug design, selection, and optimization

Network-based proximity analyses allow the prediction of drug–target interactions, which can be applied to the repurposing of existing drugs in new indications.[41,42] For instance, the deepDTnet algorithm is based on a network-based **deep learning** methodology for *in silico* identification of new molecular targets for known drugs.[43] DeepDTnet embeds 15 types of chemical, genomic, phenotypic, and cellular networks to generate biologically and pharmacologically relevant features. AI is also generating considerable interest in the design or identification of new compounds with desirable properties from virtual drug target screens. Computational chemistry has been broadly used to document quantitative structure–activity relationships (QSAR) with the goal to predict activities in a chemical space potentially encompassing millions of molecules. The QSAR field benefited over the past decade from the combined application of deep learning to neural networks with higher computational power and better algorithms addressing the **overfitting** and **gradient problems**.[44,45] ML methods are now applied to train neural networks on ligand-based virtual screens to identify and optimize drugs interacting with candidate therapeutic targets, predict their absorption, distribution, metabolism, excretion, and toxicity (ADMET) characteristics, or repurpose existing molecules.[4,42,46]

Interestingly, deep learning allows multitask prediction by developing models encompassing more than one activity, such as bioactivity and ADME properties. Whereas the prediction of multiple activities can be trained in parallel, because they share the same input and hidden layers, each activity is associated with a specific output node (Fig. 3a). A Kaggle contest evaluating various ML approaches to improve the prediction performance of QSAR methods was won by a multitask deep network yielding a 15% improvement over baseline.[47] Besides improving the accuracy of the prediction, multitask prediction based on deep learning further enhances drug discovery compared with classical ML methods (such as Random Forrest or Support Vector Machine) because the latter only predict a single property at a time. Instead of solely relying upon on-the-shelf and expert-derived chemical features, deep learning also allows the identification of novel molecular descriptors. Whereas previous ML methods used expert-compiled molecular descriptors to train the algorithms, deep learning uses such features generated without any human intervention with a form of image processing called graph convolution.[48] Combining new molecular representations with multitask prediction results in models outperforming classical QSAR models.[49] To better predict molecular activities, **multitask deep learning** can also be applied to data from image analyses generated during high-content screen (HCS) assays involving the molecule itself. Such HCSs are a rich source of information, which can be used in combination with molecular descriptors to predict biological activities, while avoiding the need for customized assays.[50]

Deep learning has also been applied to *de novo* molecule generation, with the molecule being designed by the model as opposed to by the chemist. Whereas manual approaches were previously used for evolving existing molecules by adding chemical *R* groups or changing atoms, deep learning can be used to train neural networks and generate new candidates based on previously known molecules. By adapting methods commonly applied to image analysis or language translation, a first model for *de novo* molecule generation with deep learning was built up using a variational **autoencoder** encompassing both an encoder and a decoder network (Fig. 3b). The role of the encoder is to translate the chemical structure represented as a chain of characters (e.g., SMILES) into a vector called latent space. The decoder network then translates back from the latent space vectors into SMILES to obtain refined chemical structures. A random variation can be applied to the latent space or combined with model prediction to identify a decoded molecule slightly different from the input that fits the model criteria. Multiple applications of autoencoders and derivatives have been reported in combination or not with the use of recurrent neural networks (RNNs).[51–53] Additional approaches to *de novo* molecule design are being applied in computational chemistry, such as reinforcement learning (RL), in which the network is trained step by step to reach a specific output to maximize the notion of cumulative reward.[54] Another approach is to use generative adversarial networks (GAN) associating two neural networks that both compete and collaborate in a zero-sum game to perform molecular feature extraction from very large data sets. When applied to drug development, the first 'generative' network generates candidate molecules evaluated by the second 'discriminative' network.[55–57] Despite many successes obtained in drug design by using *de novo* molecule generation and multitask prediction, some of the models obtained still produce molecules that are difficult to synthesize. In this context, computational approaches have been developed to support retrosynthesis, as a substitute of expert-derived rules or knowledge-based systems built from chemical

Please cite this article in press as: P. Moingeon et al., Drug Discovery Today (2021), https://doi.org/10.1016/j.drudis.2021.09.006

reaction databases, by decomposing the newly generated molecule using reverse reactions to design its chemical synthesis.[58,59] Deep learning has also been recently applied to support retrosynthesis analysis using a sequence-to-sequence-based model, in which the chemical structure is described as SMILES for RNN, and the reactant and product are linked as a pair in an encoder decoder.[60] Other studies reported the use in this application of either a reaction graph[61] or a combination of three deep neural networks with a Monte Carlo tree search.[62]

Toward virtual clinical studies

AI can be used in support of the design, implementation, and monitoring of clinical trials evaluating the efficacy and safety of drug candidates, with the aim to improve success rates.[63,64] For example, the selection of patient recruited in the trials is facilitated by the understanding of disease and patient heterogeneity based on models previously discussed in the section on AI and disease modeling. In addition, NLP is being used to mine real-world evidence (RWE) data or health records to assess patient eligibility in clinical studies.[65] In this approach, automated text mining is used to identify and select patients precisely fulfilling the inclusion criteria proposed in the study design, such as level of disease severity, involvement of specific target organs, and exposure to authorized background therapies. AI is also useful to inform the design of innovative trials in a precision-medicine approach by integrating massive biological, medical imaging, and clinical data to document patient specificities. During trial monitoring, AI helps to capture in a remote fashion patient-reported measurements and outcomes generated by wearable sensors or devices. It is also applied to mine such digital biomarkers providing useful information regarding symptoms, pain, cognitive function, motricity, or sleep patterns, to support diagnostic or therapeutic decisions made by the physician.[65,66] AI and ML are also being used to analyze data from successful, but also failed studies to generate models capable of predicting simultaneously the evolution of multiple and multimodal clinical parameters.[67] Those analyses can provide hypotheses regarding candidate biomarkers predictive of progression, severity, response to treatment, or even survival in the form of genome-wide polygenic scores or multi-omics signatures.[65,68,69]

A mid-term perspective generating considerable interest is to predict the efficacy of drug candidates from virtual trials. Currently, virtual representations of the characteristics of a patient are assembled in the form of a 'synthetic' patient.[70] Those models are particularly useful as substitutes for real patients when assembling placebo control groups to test drug candidates in a life-threatening or rare disease indication. The evolution of such a virtual placebo group can be modeled from RWE clinical data obtained from real patients affected by the condition, when receiving the standard of care. Furthermore, with the aim of testing the clinical efficacy of an experimental drug, *in silico* models based on quantitative system pharmacology (QSP) are also being developed, with some encouraging results.[71,72] QSP models of a

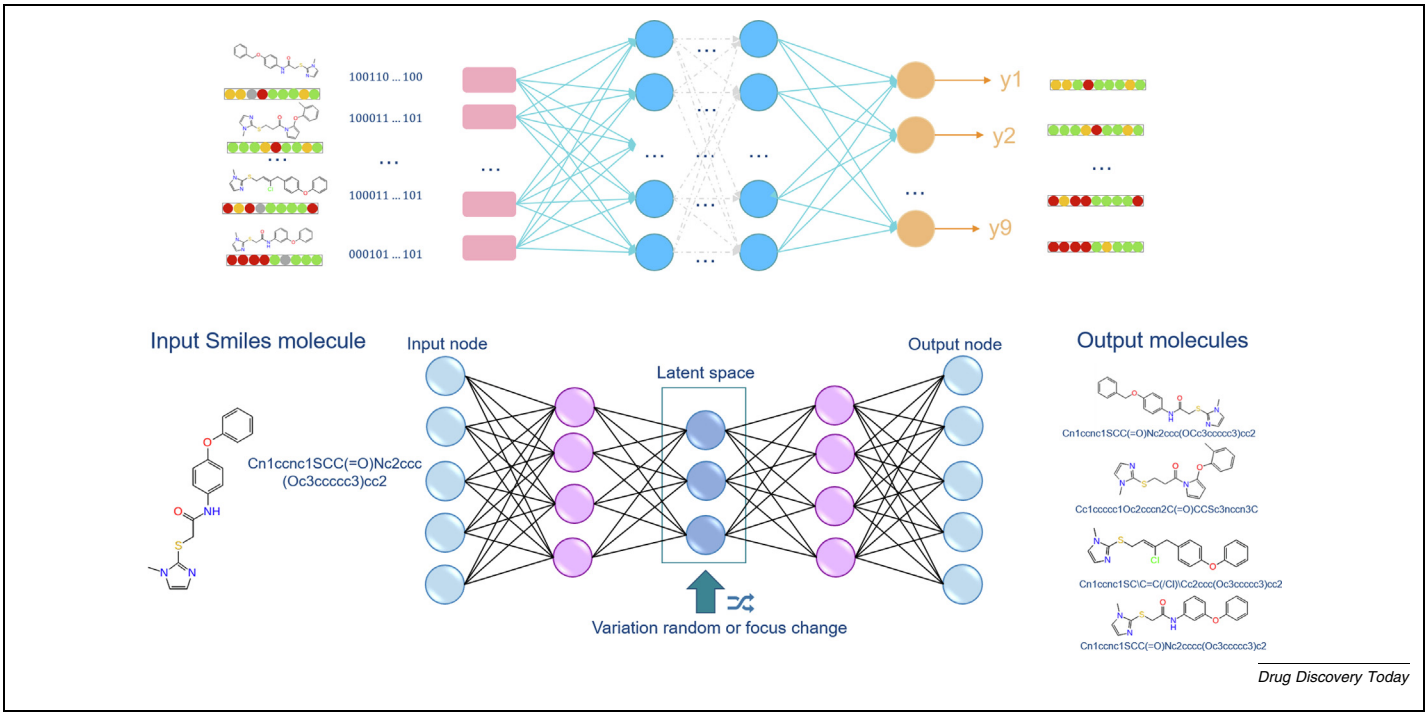

FIGURE 3

Examples of deep learning networks used in molecular modeling and drug design. **(a)** Schematic representation of a multitask prediction deep learning algorithm, on the left, with compounds used for the training and their associated data. Once the network has been trained and the best hyperparameters found, the algorithm yields as an output the full matrix prediction for all endpoints on which it has been trained. **(b)** Schematic representation of an autoencoder, with the encoder on the left, the latent space in the middle, and the decoder in the right. Once the autoencoder has been trained on millions of molecules, the latent space can be modified (through random or focus variation) to generate molecules close to the input, albeit with small changes. The autoencoder takes a SMILES as an input and produces a SMILES as the output.

Please cite this article in press as: P. Moingeon et al., Drug Discovery Today (2021), https://doi.org/10.1016/j.drudis.2021.09.006

disease of interest are built up from data related to biological processes in the blood or in tissues in association with clinical symptoms. The obtained biological system is then modeled as ordinary differential mathematical equations to represent dynamic interactions between components, further incorporating some main characteristics of the drug candidate (e.g., affinity for the target, pharmacokinetics, and biodistribution) to assess how the latter will perturb the system. QSP is used not only to predict how the drug could alleviate symptoms as it relates to specific organs, but also to identify potential biomarkers to categorize or monitor patients, select dosing and administration regimens as well as clinical endpoints to be used in the confirmatory real-world trial.[71,72] A remaining hurdle foreseen in implementing successfully a ML-powered precision medicine is related to the difficulty in establishing causal inferences, that is, to predict from a data-driven model a causal effect of drug exposure on clinical outcomes.[7] However, the future availability of AI-generated models of various diseases in the form of interactomes of genes or proteins with inferences of causality in the pathophysiology might considerably increase the capacity of in *silico* analyses to predict both the efficacy and safety of drug candidates.[73]

A remaining challenge to the broad application of AI to clinical studies remains the acceptance by major regulatory agencies of such virtual placebo groups, synthetic patients, and digital endpoints, as well as the validation of AI-based decision algorithms. Obviously, irrespective of advances in this field, real-world clinical studies will still be needed, likely fewer, simpler, and better designed with the help of AI.

Concluding remarks

Considering drug development as a succession of important decisions to be made to select the right target, drug, dosing regimen, and patient, it appears obvious that AI can support each of those decisions by capturing the value of massive and multimodal data into useful predictive models. Thus, AI and ML will undoubtfully produce an unprecedented revolution in drug development by making this complex and costly process ultimately cheaper and more effective, with both an anticipated shortening of the discovery phase and a reduction in failure rates during drug development. The health industry is now integrating those new technologies at a fast pace, as reflected by the exponential increase in the number of companies dedicated to AI applications to drug development (Table S2 in the supplemental information online). In 2020, a first AI-designed drug in the field of immuno-oncology entered Phase I clinical evaluation after only 12 months of research, compared with the 5–7 years commonly required in drug discovery. A new antibiotic, named halicin, has also been identified in record time using AI mining of existing molecules.[74] Numerous opportunities for drug repurposing generated by network computing have also been identified with first applications to cancers, neurological diseases, and Coronavirus 2019 (COVID-19).[4,42] Noteworthy, whereas ML has been mostly applied to the design of chemical molecules, those methods are also being considered for the design and selection of biologicals, including synthetic oligonucleotides, monoclonal antibodies, or peptides with predicted pharmacological properties.

Drug design and development encompass a range of existing human expertise, and the synergy between human and machine intelligences is vital for the successful enhancement of drug design and development. Intelligent machines can provide tremendous computing memory and power to conduct non-supervised analyses from massive multimodal data. Whereas deep learning methods are assimilated to black boxes, by contrast, humans are skilled at extracting features and providing transparency on the rationale underlying classification tasks or interpretability from the outputs of predictive models. Human expertise is needed to design and perform validation experiments in wet lab and real-world clinical studies. Importantly, human intelligence and judgement are required to consider ethical implications when implementing AI. The ultimate responsibility of diagnostic or therapeutic decisions informed by algorithms lies in healthcare professionals.

By helping to provide an unprecedented understanding of patient characteristics, AI is paving the way for a highly personalized medicine offering the perspective of future therapies and preventive measures precisely tailored to the needs of each individual patient based on their physiology and disease specificities. AI and ML also support the development of a medicine increasingly more predictive through access to multidimensional models encompassing the disease, patient, and drug candidate, and further participative by engaging patients and healthy individuals in managing their health. As such, we foresee the impact of AI and ML in the form of a rapid evolution towards an integrated computational precision medicine.

Declaration of interest

All authors are employees at SERVIER. The authors have no competing interests related to this manuscript.

Acknowledgment

The authors thank Dorothée Piva for excellent secretarial assistance.

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.drudis.2021.09.006.

References

1 J.A. DiMasi, H.G. Grabowski, R.W. Hansen, Innovation in the pharmaceutical industry: new estimates of R&D costs, J Health Econ 47 (2016) 20–33.

2 M.J. Waring, J. Arrowsmith, A.R. Leach, P.D. Leeson, S. Mandrell, R.M. Owen, G. Pairaudeau, W.D. Pennie, S.D. Pickett, J. Wang, O. Wallace, A. Weir, An analysis of the attrition of drug candidates from four major pharmaceutical companies, Nat Rev Drug Discov 14 (7) (2015) 475–486.

3 K.-K. Mak, M.R. Pichika, Artificial intelligence in drug development: present status and future prospects, Drug Discovery Today 24 (3) (2019) 773–780.

4 J. Vamathevan, D. Clark, P. Czodrowski, I. Dunham, E. Ferran, G. Lee, B. Li, A. Madabhushi, P. Shah, M. Spitzer, S. Zhao, Applications of machine learning in drug discovery and development, Nature Reviews Drug Discovery 18 (6) (2019) 463–477.

Please cite this article in press as: P. Moingeon et al., Drug Discovery Today (2021), https://doi.org/10.1016/j.drudis.2021.09.006

5 D. Paul, G. Sanap, S. Shenoy, D. Kalyane, K. Kalia, R.K. Tekade, Artificial intelligence in drug discovery and development, Drug Discov Today 26 (1) (2021) 80–93.
6 J. Xu, P. Yang, S. Xue, B. Sharma, M. Sanchez-Martin, F. Wang, K.A. Beaty, E. Dehan, B. Parikh, Translating cancer genomics into precision medicine with artificial intelligence: applications, challenges and future perspectives, Hum Genet 138 (2) (2019) 109–124.
7 J. Wilkinson, K.F. Arnold, E.J. Murray, M. van Smeden, K. Carr, R. Sippy, M. de Kamps, A. Beam, S. Konigorski, C. Lippert, M.S. Gilthorpe, P.W.G. Tennant, Time to reality check the promises of machine learning-powered precision medicine, The Lancet Digital Health 2 (12) (2020) e677–e680.
8 CDISC | Clear Data. Clear Impact. https://www.cdisc.org/ [Accessed September 14, 2021].
9 FAIR Principles. GO FAIR. https://www.go-fair.org/fair-principles/ [Accessed September 14, 2021].
10 M.D. Wilkinson, M. Dumontier, I.J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L.B. da Silva Santos, P.E. Bourne, J. Bouwman, A.J. Brookes, T. Clark, M. Crosas, I. Dillo, O. Dumon, S. Edmunds, C.T. Evelo, R. Finkers, A. Gonzalez-Beltran, A.J.G. Gray, P. Groth, C. Goble, J.S. Grethe, J. Heringa, P.A.C. 't Hoen, R. Hooft, T. Kuhn, R. Kok, J. Kok, S.J. Lusher, M.E. Martone, A. Mons, A.L. Packer, B. Persson, P. Rocca-Serra, M. Roos, R. van Schaik, S.-A. Sansone, E. Schultes, T. Sengstag, T. Slater, G. Strawn, M.A. Swertz, M. Thompson, J. van der Lei, E. van Mulligen, J. Velterop, A. Waagmeester, P. Wittenburg, K. Wolstencroft, J. Zhao, B. Mons, The FAIR Guiding Principles for scientific data management and stewardship, Scientific Data 3 (1) (2016), https://doi.org/10.1038/sdata.2016.18.
11 General Data Protection Regulation (GDPR) – Official Legal Text. General Data Protection Regulation (GDPR). https: //gdpr-info.eu/ [Accessed September 14, 2021].
12 MELLODDY. https://www.melloddy.eu [Accessed September 14, 2021].
13 J. Tang, Z.-u.-R. Tanoli, B. Ravikumar, Z. Alam, A. Rebane, M. Vähä-Koskela, G. Peddinti, A.J. van Adrichem, J. Wakkinen, A. Jaiswal, E. Karjalainen, P. Gautam, L. He, E. Parri, S. Khan, A. Gupta, M. Ali, L. Yetukuri, A.-L. Gustavsson, B. Seashore-Ludlow, A. Hersey, A.R. Leach, J.P. Overington, G. Repasky, K. Wennerberg, T. Aittokallio, Drug target commons: a community effort to build a consensus knowledge base for drug-target interactions, Cell Chem Biol 25 (2) (2018) 224–229.e2.
14 Kaggle: Your Machine Learning and Data Science Community. https://www.kaggle.com/ [Accessed July 12, 2021].
15 DREAM Challenges. https://dreamchallenges.org/ [Accessed September 14, 2021].
16 PrecisionFDA Truth Challenge – precisionFDA. https://precision.fda.gov/challenges/truth [Accessed September 14, 2021].
17 E.P. García del Valle, G. Lagunes García, L. Prieto Santamaría, M. Zanin, E. Menasalvas Ruiz, A. Rodríguez-González, Disease networks and their contribution to disease understanding: a review of their evolution, techniques and data sources, J Biomed Inform 94 (2019) 103206, https://doi.org/10.1016/j.jbi.2019.103206.
18 S.A. Dugger, A. Platt, D.B. Goldstein, Drug development in the era of precision medicine, Nat Rev Drug Discov 17 (3) (2018) 183–196.
19 L.-H. Lee, J. Loscalzo, Network medicine in pathobiology, Am J Pathol 189 (7) (2019) 1311–1326.
20 Y. Hasin, M. Seldin, A. Lusis, Multi-omics approaches to disease, Genome Biol 18 (1) (2017) 83.
21 G. Tini, L. Marchetti, C. Priami, M.-P. Scott-Boyer, Multi-omics integration-a comparison of unsupervised clustering methodologies, Brief Bioinform 20 (4) (2019) 1269–1279.
22 E. Becht, L. McInnes, J. Healy, C.-A. Dutertre, I.W.H. Kwok, L.G. Ng, F. Ginhoux, E.W. Newell, Dimensionality reduction for visualizing single-cell data using UMAP, Nature Biotechnology 37 (1) (2019) 38–44.
23 I.C. Macaulay, C.P. Ponting, T. Voet, Single-cell multiomics: multiple measurements from single cells, Trends Genet 33 (2) (2017) 155–168.
24 J.C.B. Gamboa, Deep learning for time-series analysis, arXiv 2017 (1887) 17010.
25 J. Guinney, R. Dienstmann, X. Wang, A. de Reyniès, A. Schlicker, C. Soneson, L. Marisa, P. Roepman, G. Nyamundanda, P. Angelino, B.M. Bot, J.S. Morris, I.M. Simon, S. Gerster, E. Fessler, F. De Sousa E Melo, E. Missiaglia, H. Ramay, D. Barras, K. Homicsko, D. Maru, G.C. Manyam, B. Broom, V. Boige, B. Perez-Villamil, T. Laderas, R. Salazar, J.W. Gray, D. Hanahan, J. Tabernero, R. Bernards, S.H. Friend, P. Laurent-Puig, J.P. Medema, A. Sadanandam, L. Wessels, M. Delorenzi, S. Kopetz, L. Vermeulen, S. Tejpar, The consensus molecular subtypes of colorectal cancer, Nature Medicine 21 (11) (2015) 1350–1356.
26 C. Ai, L. Kong, CGPS: a machine learning-based approach integrating multiple gene set analysis tools for better prioritization of biologically relevant pathways, Journal of Genetics and Genomics 45 (9) (2018) 489–504.
27 U. Stelzl, U. Worm, M. Lalowski, C. Haenig, F.H. Brembeck, H. Goehler, M. Stroedicke, M. Zenkner, A. Schoenherr, S. Koeppen, J. Timm, S. Mintzlaff, C. Abraham, N. Bock, S. Kietzmann, A. Goedde, E. Toksöz, A. Droege, S. Krobitsch, B. Korn, W. Birchmeier, H. Lehrach, E.E. Wanker, A human protein-protein interaction network: a resource for annotating the proteome, Cell 122 (6) (2005) 957–968.
28 N. Friedman, Inferring cellular networks using probabilistic graphical models, Science 303 (5659) (2004) 799–805.
29 C.J. Needham, J.R. Bradford, A.J. Bulpitt, D.R. Westhead, Inference in Bayesian networks, Nature Biotechnology 24 (1) (2006) 51–53.
30 A.-L. Barabási, N. Gulbahce, J. Loscalzo, Network medicine: a network-based approach to human disease, Nature Reviews Genetics 12 (1) (2011) 56–68.
31 L. Cowen, T. Ideker, B.J. Raphael, R. Sharan, Network propagation: a universal amplifier of genetic associations, Nature Reviews Genetics 18 (9) (2017) 551–562.
32 L. Cantini, E. Medico, S. Fortunato, M. Caselle, Detection of gene communities in multi-networks reveals cancer drivers, Scientific Reports 5 (1) (2015) 17386.
33 Raue A, Schilling M, Bachmann J, Matteson A, Schelker M, Kaschek D, et al. Lessons learned from quantitative dynamical modeling in systems biology. PLoS ONE 2013; 8(9): e74335.
34 Ochoa D, Hercules A, Carmona M, Suveges D, Gonzalez-Uriarte A, Malangone C, et al. Open Targets Platform: supporting systematic drug–target identification and prioritisation. Nucleic Acids Research 2021; 49(D1): D1302-D1310.
35 Owens J. Determining druggability. Nature Reviews Drug Discovery 2007; 6(3): 187-187.
36 W. Muster, A. Breidenbach, H. Fischer, S. Kirchner, L. Müller, A. Pähler, Computational toxicology in drug development, Drug Discovery Today 13 (7-8) (2008) 303–310.
37 L.J. Jensen, J. Saric, P. Bork, Literature mining for the biologist: from information retrieval to biological discovery, Nature Reviews Genetics 7 (2) (2006) 119–129.
38 V. Vergetis, D. Skaltsas, V.G. Gorgoulis, A. Tsirigos, Assessing drug development risk using big data and machine learning, Cancer Res 81 (4) (2021) 816–819.
39 Skalic M, Varela-Rial A, Jiménez J, Martínez-Rosell G, De Fabritiis G. LigVoxel: inpainting binding pockets using 3D-convolutional neural networks. Bioinformatics 2019; 35(2): 243-250.
40 A.W. Senior, R. Evans, J. Jumper, J. Kirkpatrick, L. Sifre, T. Green, C. Qin, A. Žídek, A.W.R. Nelson, A. Bridgland, H. Penedones, S. Petersen, K. Simonyan, S. Crossan, P. Kohli, D.T. Jones, D. Silver, K. Kavukcuoglu, D. Hassabis, Improved protein structure prediction using potentials from deep learning, Nature 577 (7792) (2020) 706–710.
41 E. Guney, J. Menche, M. Vidal, A.-L. Barábasi, Network-based in silico drug efficacy screening, Nature Communications 7 (1) (2016) 10331.
42 F. Cheng, R.J. Desai, D.E. Handy, R. Wang, S. Schneeweiss, A.-L. Baraba´si, J. Loscalzo, Network-based approach to prediction and population-based validation of in silico drug repurposing, Nat Commun 9 (1) (2018), https://doi.org/10.1038/s41467-018-05116-5.
43 X. Zeng, S. Zhu, W. Lu, Z. Liu, J. Huang, Y. Zhou, J. Fang, Y. Huang, H. Guo, L. Li, B.D. Trapp, R. Nussinov, C. Eng, J. Loscalzo, F. Cheng, Target identification among known drugs by deep learning from heterogeneous networks, Chem Sci 11 (7) (2020) 1775–1797.
44 Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: a simple way to prevent neural networks from overfitting. J Mach Learn Res 2014; 15(1): 1929-1958.]
45 Nair V, Hinton GE. Rectified linear units improve restricted Boltzmann machines. In: Fürnkranz J, Joachims T, eds. Proceedings of the 27th International Conference on International Conference on Machine Learning. ICML'10. Madison: Omnipress, 2010: 807-814
46 A.S. Rifaioglu, H. Atas, M.J. Martin, R. Cetin-Atalay, V. Atalay, T. Doğan, Recent applications of deep learning and machine intelligence on in silico drug discovery: methods, tools and databases, Brief Bioinform 20 (5) (2019) 1878–1912.
47 B. Ramsundar, B. Liu, Z. Wu, A. Verras, M. Tudor, R.P. Sheridan, V. Pande, Is multitask deep learning practical for pharma?, J Chem Inf Model 57 (8) (2017) 2068–2076
48 Wieder O, Kohlbacher S, Kuenemann M, Garona A, Ducrota P, Seidel T et al. A compact review of molecular property prediction with graph neural networks. Drug Discovery Today: Technologies. Published online December 17, 2020. http://dx.doi.org/10.1016/j.ddtec.2020.11.009
49 Walters WP, Barzilay R. Applications of deep learning in molecule generation and molecular property prediction. Acc Chem Res 2021; 54(2): 263-270.

Please cite this article in press as: P. Moingeon et al., Drug Discovery Today (2021), https://doi.org/10.1016/j.drudis.2021.09.006

50 J. Simm, G. Klambauer, A. Arany, M. Steijaert, J.K. Wegner, E. Gustin, V. Chupakhin, Y.T. Chong, J. Vialard, P. Buijnsters, I. Velter, A. Vapirev, S. Singh, A. E. Carpenter, R. Wuyts, S. Hochreiter, Y. Moreau, H. Ceulemans, Repurposing high-throughput image assays enables biological activity prediction for drug discovery, Cell Chem Biol 25 (5) (2018) 611–618.e3.

51 A. Kadurin, S. Nikolenko, K. Khrabrov, A. Aliper, A. Zhavoronkov, druGAN: an advanced generative adversarial autoencoder model for de novo generation of new molecules with desired molecular properties in silico, Mol Pharm 14 (9) (2017) 3098–3104.

52 T. Blaschke, M. Olivecrona, O. Engkvist, J. Bajorath, H. Chen, Application of generative autoencoder in de novo molecular design, Mol Inform 37 (1-2) (2018) 1700123, https://doi.org/10.1002/minf.201700123.

53 M.H.S. Segler, T. Kogej, C. Tyrchan, M.P. Waller, Generating focused molecule libraries for drug discovery with recurrent neural networks, ACS Cent Sci 4 (1) (2018) 120–131.

54 M. Olivecrona, T. Blaschke, O. Engkvist, H. Chen, Molecular de-novo design through deep reinforcement learning, J Cheminform 9 (1) (2017) 48.

55 Guimaraes GL, Sanchez-Lengeling B, Outeiral C, Farias PLC, Aspuru-Guzik A. Objective-reinforced generative adversarial networks (ORGAN) for sequence generation models. arXiv 2018: 170510843.

56 De Cao N, Kipf T. MolGAN: an implicit generative model for small molecular graphs. arXiv 2018: 180511973.

57 E. Putin, A. Asadulaev, Y. Ivanenkov, V. Aladinskiy, B. Sanchez-Lengeling, A. Aspuru-Guzik, A. Zhavoronkov, Reinforced adversarial neural computer for de novo molecular design, J Chem Inf Model 58 (6) (2018) 1194–1204.

58 O. Engkvist, P.-O. Norrby, N. Selmi, Y.-H. Lam, Z. Peng, E.C. Sherer, W. Amberg, T. Erhard, L.A. Smyth, Computational prediction of chemical reactions: current status and outlook, Drug Discov Today 23 (6) (2018) 1203–1218.

59 J. Law, Z. Zsoldos, A. Simon, D. Reid, Y. Liu, S.Y. Khew, A.P. Johnson, S. Major, R. A. Wade, H.Y. Ando, Route Designer: a retrosynthetic analysis tool utilizing automated retrosynthetic rule generation, J Chem Inf Model 49 (3) (2009) 593–602.

60 B. Liu, B. Ramsundar, P. Kawthekar, J. Shi, J. Gomes, Q. Luu Nguyen, S. Ho, J. Sloane, P. Wender, V. Pande, Retrosynthetic reaction prediction using neural sequence-to-sequence models, ACS Cent Sci 3 (10) (2017) 1103–1113.

61 Savage J, Kishimoto A, Buesser B, Diaz-Aviles E, Alzate C. Chemical reactant recommendation using a network of organic chemistry. In: Cremonesi P, Ricci F, eds; RecSys '17: Proceedings of the Eleventh ACM Conference on Recommender Systems. New York: Association for Computing Machinery, 2017: 210-214.

62 M.H.S. Segler, M. Preuss, M.P. Waller, Planning chemical syntheses with deep neural networks and symbolic AI, Nature 555 (7698) (2018) 604–610.

63 M. Hay, D.W. Thomas, J.L. Craighead, C. Economides, J. Rosenthal, Clinical development success rates for investigational drugs, Nat Biotechnol 32 (1) (2014) 40–51.

64 S. Harrer, P. Shah, B. Antony, J. Hu, Artificial intelligence for clinical trial design, Trends in Pharmacological Sciences 40 (8) (2019) 577–591.

65 P. Shah, F. Kendall, S. Khozin, R. Goosen, J. Hu, J. Laramie, M. Ringel, N. Schork, Artificial intelligence and machine learning in clinical development: a translational perspective, NPJ Digital Medicine 2 (1) (2019), https://doi.org/10.1038/s41746-019-0148-3.

66 P. Boehme, A. Hansen, R. Roubenoff, J. Scheeren, M. Herrmann, T. Mondritzki, J. Ehlers, H. Truebel, How soon will digital endpoints become a cornerstone for future drug development?, Drug Discov Today 24 (1) (2019) 16–19

67 D.B. Fogel, Factors associated with clinical trials that fail and opportunities for improving the likelihood of success: a review, Contemp Clin Trials Commun 11 (2018) 156–164.

68 A.V. Khera, M. Chaffin, K.G. Aragam, M.E. Haas, C. Roselli, S.H. Choi, P. Natarajan, E.S. Lander, S.A. Lubitz, P.T. Ellinor, S. Kathiresan, Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations, Nat Genet 50 (9) (2018) 1219–1224.

69 A.J. Steele, S.C. Denaxas, A.D. Shah, H. Hemingway, N.M. Luscombe, T.R. Singh, Machine learning models in electronic health records can outperform conventional survival models for predicting patient mortality in coronary artery disease, PLoS ONE 13 (8) (2018) e0202344, https://doi.org/10.1371/journal.pone.020234410.1371/journal.pone.0202344.g00110.1371/journal.pone.0202344.g00210.1371/journal.pone.0202344.g00310.1371/journal.pone.0202344.g00410.1371/journal.pone.0202344.g00510.1371/journal.pone.0202344.t00110.1371/journal.pone.0202344.t002.

70 A. Tucker, Z. Wang, Y. Rotalinti, P. Myles, Generating high-fidelity synthetic patient data for assessing machine learning healthcare software, NPJ Digital Medicine 3 (1) (2020) 1–13.

71 Sorger P, Allerheiligen SRB. Quantitative and Systems Pharmacology in the Post-genomic Era: New Approaches to Discovering Drugs and Understanding Therapeutic Mechanisms. 2011. https://www.nigms.nih.gov/training/documents/systemspharmawpsorger2011.pdf [Accessed September 14, 2021]

72 C.M. Friedrich, A model qualification method for mechanistic physiological QSP models to support model-informed drug development, CPT Pharmacometrics Syst Pharmacol 5 (2) (2016) 43–53.

73 M. Danhof, Systems pharmacology - towards the modeling of network interactions, Eur J Pharm Sci 94 (2016) 4–14.

74 J.M. Stokes, K. Yang, K. Swanson, W. Jin, A. Cubillos-Ruiz, N.M. Donghia, C.R. MacNair, S. French, L.A. Carfrae, Z. Bloom-Ackermann, V.M. Tran, A. Chiappino-Pepe, A.H. Badran, I.W. Andrews, E.J. Chory, G.M. Church, E.D. Brown, T.S. Jaakkola, R. Barzilay, J.J. Collins, A deep learning approach to antibiotic discovery, Cell 180 (4) (2020) 688–702.e13.

Glossary

Artificial Intelligence (AI): any type of machine presenting a simple intelligence. In computer science, AI is defined as a machine able to perform tasks requiring human intelligence, such as visual perception, speech recognition, decision-making, and language translation.

Autoencoder: neural network technique that performs dimensionality reduction. It comprises an encoder part compressing and encoding data efficiently, and a decoder part, which learns how to reconstruct the data back as close as possible to the original.

Consensus: convergence between predictions obtained from different models, each of them generated from either different data sources (e.g., multi-omics) or computational approaches (e.g., hierarchical, Gaussian, and k-means clusterings).

Data sparsity: computational challenge linked to the completeness of the observations in a data set.

Deep learning: advanced algorithm mimicking the human brain by using artificial neurons in a complex network.

Gradient problem: a gradient measures how much the output of a model changes when inputs are modified. In gradient-based approaches, such as neural networks, gradients are used during training to update some parameter weights. When the magnitudes of the gradients accumulate, an unstable model is likely to occur, which can lead to poor prediction results. Methods to manage exploding gradients include gradient clipping and weight regularization.

Layer: structure in the architecture of the model, which enables information to be taken from a previous layer and be passed on to the next one. Different types of layer exist, such as fully connected or convolutional layers.

Machine learning (ML): use and development of algorithms based on sample data (e.g., experimental data), which can learn and adapt without human instructions to analyze and draw inferences from patterns in data.

Multitask deep learning: process to solve multiple learning tasks simultaneously, while exploiting commonalities and differences across each of them. Multitasking has been used successfully across all applications of ML.

Natural language processing (NLP): treatment of human language by intelligent machines to understand and extract relevant information from the content of data sources, such as publications or patents.

Networks/knowledge graphs: set of entities or nodes (e.g., genes, proteins, drugs, or diseases), connected to each other by relationships or links (e.g., gene–disease association, protein–protein interaction, or drug–target interactions).

Network propagation: algorithms (e.g., random walk or information diffusion) used to propagate data into the topology of a given network to amplify colocalized high signals and support functional interpretation.

Overfitting: predictive model corresponding too closely to the data set on which it has been trained, which therefore might fail to be validated in other data sets. Cross-validation or bootstrapping resampling approaches are generally proposed to reduce the overfitting effect.

Supervised learning: aims to create a prediction function based on labeled data (as opposed to unsupervised learning); encompasses both classification learning based on qualitative data and regression learning trained by using quantitative data.

Unsupervised learning: a mode of ML in which data are not labeled; aims to discove r the underlying structures to label or group those unlabeled data.

Please cite this article in press as: P. Moingeon et al., Drug Discovery Today (2021), https://doi.org/10.1016/j.drudis.2021.09.006