Information Fusion 79 (2022) 263–278

Contents lists available at ScienceDirect

Information Fusion

journal homepage: www.elsevier.com/locate/inffus

Information fusion as an integrative cross-cutting enabler to achieve robust, explainable, and trustworthy medical artificial intelligence

Andreas Holzinger [a,b,*], Matthias Dehmer [c,d], Frank Emmert-Streib [e], Rita Cucchiara [g,h], Isabelle Augenstein [i], Javier Del Ser [n,o], Wojciech Samek [p], Igor Jurisica [j,k,l,m], Natalia Díaz-Rodríguez [f]

[a] Medical University Graz, Austria
[b] Alberta Machine Intelligence Institute, University of Alberta, Canada
[c] University of Medical Informatics Tyrol, Austria
[d] Swiss Distance University of Applied Sciences, Switzerland
[e] Predictive Society and Data Analytics Lab, Faculty of Information Technology and Communication Sciences, Tampere University, Tampere, Finland
[f] Andalusian Research Institute in Data Science and Computational Intelligence (DaSCI Institute). University of Granada, Spain
[g] University of Modena and Reggio Emilia, Modena, Italy
[h] Artificial Intelligence Research and Innovation Center, Modena, Italy
[i] Department of Computer Science, University of Copenhagen, Copenhagen, Denmark
[j] Osteoarthritis Research Program, Division of Orthopedic Surgery, Schroeder Arthritis Institute, University Health Network, Toronto, Canada
[k] Krembil Research Institute, Data Science Discovery Centre for Chronic Diseases, University Health Network, Toronto, Canada
[l] Departments of Medical Biophysics and Computer Science, University of Toronto, Canada
[m] Institute of Neuroimmunology, Slovak Academy of Sciences, Bratislava, Slovakia
[n] TECNALIA, Basque Research and Technology Alliance (BRTA), Derio, Spain
[o] University of the Basque Country (UPV/EHU), Bilbao, Spain
[p] Departments of Artificial Intelligence, Fraunhofer Heinrich Hertz Institute, Germany

ARTICLE INFO

Keywords:
Artificial intelligence
Information fusion
Medical AI
Explainable AI
Robustness
Explainability
Trust
Graph-based machine learning
Neural-symbolic learning and reasoning

ABSTRACT

Medical artificial intelligence (AI) systems have been remarkably successful, even outperforming human performance at certain tasks. There is no doubt that AI is important to improve human health in many ways and will disrupt various medical workflows in the future. Using AI to solve problems in medicine beyond the lab, in routine environments, we need to do more than to just improve the performance of existing AI methods. Robust AI solutions must be able to cope with imprecision, missing and incorrect information, and explain both the result and the process of how it was obtained to a medical expert. Using conceptual knowledge as a guiding model of reality can help to develop more robust, explainable, and less biased machine learning models that can ideally learn from less data. Achieving these goals will require an orchestrated effort that combines three complementary Frontier Research Areas: (1) Complex Networks and their Inference, (2) Graph causal models and counterfactuals, and (3) Verification and Explainability methods. The goal of this paper is to describe these three areas from a unified view and to motivate how information fusion in a comprehensive and integrative manner can not only help bring these three areas together, but also have a transformative role by bridging the gap between research and practical applications in the context of future trustworthy medical AI. This makes it imperative to include ethical and legal aspects as a cross-cutting discipline, because all future solutions must not only be ethically responsible, but also legally compliant.

1. Introduction and motivation

Artificial intelligence in medicine is on everyone's lips. Politicians around the world have declared its use a worthy goal. Industry sees its use as a huge driver of growth, and medicine sees it as a great opportunity for solving medical problems, providing new insights, and improving the quality of decision support. AI and machine learning will, and already are, transforming many workflows. As a result, society has recently seen an upsurge of success stories and use cases in which AI has played a major role in realizing unprecedented levels of diagnostic performance [1–9].

* Corresponding author at: Medical University Graz, Austria.
E-mail address: andreas.holzinger@medunigraz.at (A. Holzinger).

https://doi.org/10.1016/j.inffus.2021.10.007
Received 23 June 2021; Received in revised form 14 September 2021; Accepted 23 October 2021
Available online 12 November 2021
1566-2535/© 2021 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

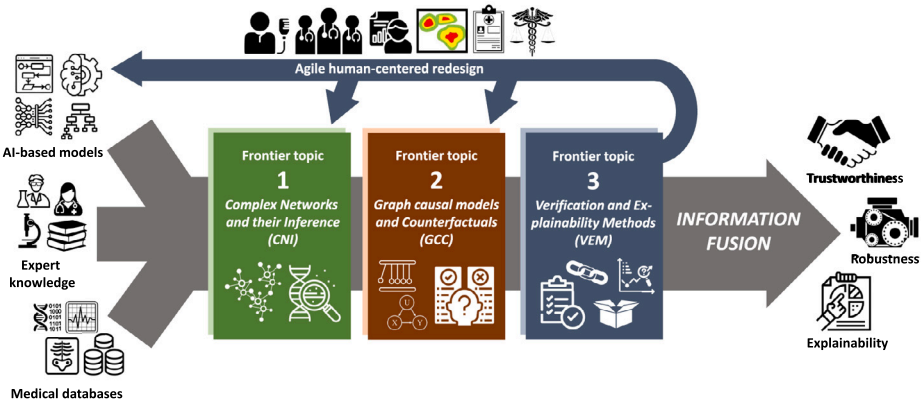


Fig. 1. Graphical abstract: Information Fusion as integrative cross-sectional topic.

In order to use AI to solve problems in medicine, biology, and life sciences outside of our labs and in routine settings, there is an urgent need to move beyond mere benchmarking and improve the performance of methods that work only with independent and identically distributed (i.i.d.) data. Even the best current machine learning models do not generalize well, have difficulty with small training datasets [10], and are sensitive to even small perturbations [11–13]. Moreover, the most promising approaches are difficult for human experts to interpret, however, most importantly, they are not able to infer causal relationships. Therefore, explainability and robustness have been declared by the European Union as the most important properties for successful medical AI. Robustness and explainability are also important prerequisites for discovering causal relationships and enabling the verifiability of machine decisions by a human expert in a given context. This cannot be achieved by a single approach, but requires concerted action from complementary disciplines.

Research and teaching are trying to keep up with these trends in order to meet these functional requirements for medical AI. A systematic and up-to-date treatment of the topic in research-led teaching is therefore not only necessary but crucial for the practical and effective implementation of AI in the future in order to secure the increasing demand for highly qualified specialists in Europe and worldwide. The task of this new generation of experts will be to bring the latest developments into daily application.

The goal of this position paper is to identify the most relevant pioneering frontier research areas and make the case for why and how they can contribute to a concerted integrative effort to make future medical AI efficient and effective in practice. Specifically, we discuss on three *Frontier Research Areas (FRA)*:

(1) Complex Networks and their Inference (CNI);
(2) Graph Causal models and Counterfactuals (GCC); and
(3) Verification and Explainability Methods (VEM).

All through the above FRA, we advocate for information fusion as the integrative cross-cutting catalyst that unleashes a great chance to unify and synergize these three FRA. The new "AI spring" is causing an exponential increase not only in interest in AI, but also an actual increase in the use of AI in all areas of life, including medicine. This inevitably raises questions of reliability, safety, fairness, as well as moral and ethical integrity [14], in addition to questions of robustness and explainability. Therefore, ethical and legal aspects must always be included. All future solutions must not only be ethically responsible [15], but also legally compliant [16]. The European Union has taken a clear stance on AI: AI must be human-centered and trustworthy. To be trustworthy, any AI must comply with applicable rules and regulations, adhere to ethical principles [17], and be implemented in a secure and robust manner, as defined by the EU High-Level Expert Group on AI.[1] To this end, and following 1, a cyclic, iterative, agile human-centered AI redesign process, based on agile user-centered design methods [18] is needed to intertwine the proposed frontier topics with respect to the proposed information fusion approach, eventually reaching the degrees of trustworthiness, robustness and explainability required to fully harness the potential of medical AI.

This paper is organized as follows: for each Frontier Research Area, we begin with a few selected specific problems to show *what* problems each FRA addresses. We proceed by describing *why* the topic under study by every FRA is a problem for medical AI, and the extent to which the current state of the art falls short of what is needed to solve the problem. We then describe *how* the problem can be addressed, and present some promising work in the literature that goes in the right direction for this FRA. In a subsequent section, which we refer to as *"Desiderata",* we list some general characteristics and features that future technical achievements in this FRA should have for this application domain. We conclude each section by highlighting the practical benefits of realizing this FRA and how it will help *bridge the gap* between scientific achievements and their practical implementation in the medical domain.

2. Frontier research area 1: Complex networks and their inference

2.1. What: Fighting complex diseases poses many problems in the integration and scalability of machine learning methods

Exploring and researching *complex diseases* such as arthritis, brain disorders, cancer, or infectious diseases such as COVID-19 requires novel medical decision support systems that are able to incorporate not only humans into the loop, but also integrative analyses combining diverse omics datasets along with clinical information from a wide variety of modalities [19], using scalable methods for data fusion and mining [20], machine learning, statistics, graph theory and graph visualization into low-dimensional representations because human cognition is not optimized to work well in high-dimensional spaces. Among the myriads of properties describing genome, epigenome, transcriptome, microbiome, phenotype, lifestyle, etc., no single data type, however, can capture the complexity of all the factors relevant to *understanding a phenomenon such as a disease.* A key challenge is the identification of effective models to provide a *relevant* systems view [21].

Additional insights can be gained and *in vivo* validations better planned by trying to understand the conservation of deregulated genes, networks, and pathways across organisms [22,23] – which is a major and, to date, unsolved problem.

[1] https://digital-strategy.ec.europa.eu/en/policies/expert-group-ai (access: June, 17, 2021).

Developing effective knowledge repositories to support the governance, processing, inference, analysis and interactive visualization of integrated omics [9] and network data is essential [24–38]. Integrating diverse assays and algorithms, in turn, helps to address both false positives and false negatives [39,40].

2.2. Why: Integrating data with networks makes it possible to identify novel relationships between data silos

Currently, explainable AI (XAI) developments are mostly uni-modal. However, enriched, more feasible explanations in the medical domain can be achieved if they consider multimodality. Integrating data with networks – protein interactions networks, transcription regulatory networks, microRNA-gene networks [41], metabolic and signaling pathways – enables to identify relationship among data silos [42]. Further analyzing these annotated networks with graph theory algorithms or knowledge engineering tools provides insights into their structure [43,44], which in turn, can characterize the function of these proteins, transcription factors and microRNAs [45]. Combining machine learning, data mining and graph theory is difficult, but critical to maximize the impact on translational research [46], enable more accurate and explainable modeling, increase our understanding of complex diseases [47,48] and, ultimately, support P4 medicine (precision, personalized, participatory, preventive) medicine [49–51].

Challenges at the intersection of machine learning and network biology for *Next-Generation Machine Learning* for *Biological Networks*, which could impact disease biology, drug discovery, microbiome research, and synthetic biology are discussed in Camacho et al. (2018) [52].

2.3. How: Quantitative graph theory can help interpret integrated omics data within diseases

Graphs have been used in life sciences for a long time. In recent years, there is a growing trend to combine elements of graph theory, machine learning, and statistical data analysis, which offers tremendous opportunities especially to support interactive knowledge discovery for personalized medicine [53]. In network analysis, complex biomedical graph data is examined, and the increasingly easy generation of large amounts of genomics, proteomics, metabolomics etc., and signaling data enables the construction of large networks that provide a framework for understanding the molecular basis of physiological and pathological conditions. Such complex networks have been investigated extensively for several purposes [54,55]. On the one hand, networks have been explored in the context of studying complex systems by means of graphs. Examples thereof are biological, linguistic, chemical and technical networks [56]. Other contributions in this area relate to study motifs and modules within complex networks [55]. On the other hand, lots of quantitative analyses on networks have been performed [57].

To shed light on this problem, we briefly sketch Quantitative Graph Theory, introduced by Dehmer and Emmert-Streib [58]. Quantitative Graph Theory can be divided into two major categories, namely *Comparative Network Analysis*, *Network Characterization* and *networks explainable by design*. Comparative Network Analysis relates to measuring the structural similarity between networks [59]. This can be done by using so-called *exact* or *inexact* graph matching, see [60,61]. Exact graph matching is based on the concept of graph isomorphism. Inexact graph matching relates to determining a gradual change on the similarity between graphs by utilizing graph invariants. Another approach for measuring the similarity between graphs is based on utilizing topological indices as an input when using distance or similarity measures for real numbers [62].

Next, *Network Characterization* using quantitative graph complexity measures can be employed. A network measure is a function that maps network instances to positive real numbers. In mathematical chemistry, they are often referred to as topological indices [63]. Many complexity measures for graphs have been developed, e.g., based on distances, vertex degrees, graph automorphism and so forth. We refer to [63–65] for more details. One promising domain for the future is the emerging field of geometric deep learning, which is an umbrella term for new techniques that attempt to generalize (structured) deep neural models to non-Euclidean domains, such as graphs and manifolds [66]. Machine learning of networks is promising and has recently been used very successfully to fight Covid-19 [67].

With respect to networks explainable by design, compositional part-based object detecting and classifying neural symbolic explainable models [68] can aid the explanations based on not only on coarse grained labels, but more fine grained findings, and provide a wider provenance that traces the explanation to the very source, i.e., at the data acquisition stage. This goes beyond current XAI techniques that limit their explanations to provide rationale only for a given input and output sample data [69–71]. Going beyond uni-modal explanations makes the information fusion aspect to be of paramount importance in the explanation process, to allow traceability from the data collection, to the output explanation interfaces with a diverse set of audience profiles that participate in the medical and clinical processes characterized by different backgrounds and expertise.

Apart from the methods sketched above, networks have also been used in other areas including data mining, machine learning, lexical semantics, information fusion [72–75] and integrative computational biology, such as cell differentiation [44].

Despite inherent noise present in interaction datasets, systematic analyses of these networks uncover biologically relevant information, such as lethality [76,77], functional organization [78–81], hierarchical structure [44,82,83], modularity [45,84–87] and network-building motifs [43,88,89], even across time [50]. This suggests that networks have a strong structure-function relationship [43], which can be used to help interpret integrated omics data within diseases [42,90], across diseases [91] and across organisms [36,37], understand drug mechanism of action and toxicity [92], and performing causal inference on big data [93].

2.4. Desiderata: Fusing machine learning with systematic graph theory promotes the knowledge gain of multi-modal data and their interrelationships

Many interactions are transient, so networks change in different tissues or under different stimuli [37,94–96]. Studying the dynamics of these networks is an exponentially complex task. Many stable complexes show strong co-expression of corresponding genes, whereas transient complexes lack this support [97,98]. These contextual network dynamics must be considered when linking interactions to phenotypes and when studying the networks topology. Analyzing such insights on the network dynamics towards the identification and minimization of different biases of individual detection methods, the simple intersection of results achieves high precision at the cost of low recall.

Systematic graph theory analyses of dynamic changes in interaction networks, combined with probabilistic modeling [99], and integrated with gene and protein cancer profiles enable comprehensive analyses of complex diseases such as cancer [100–102], generating new insights [42,51], robust biomarkers [90,91,103] and models that explain causal relationships through network inference [104,105]. Implementing algorithms using heuristics fine-tuned for interaction networks [106–108] will ensure their scalability. Finally, we also highlight achievements reported lately on the use of Deep Learning methods to undertake modeling problems formulated over interaction networks, which have so far elicited promising results [7,109].

2.5. What for: Pushing the boundaries in this FRA will help understanding complex diseases

There are many benefits emerging from early steps taken along this FRA. For instance, some of the most successful network-based methods

of gene group identification for class prediction have been the score-based sub-network markers [110–113]. Sub-networks identified using these approaches were recently shown to be highly conserved across studies and to perform better than individual genes or pre-defined gene groups at predicting breast cancer metastasis [111]. Improving these methods by considering network modularity results in better prediction of aging [45]. Combining existing known and predicted interactions with novel local co-expression annotation of existing edges will elucidate disease-specific dynamics and identify local network structures (graphlets, [107,114]) that are the most aberrant components in the cancer network, as compared to a normal control case. Network dynamics [94], in turn, enable explainable modeling of healthy and disease signaling cascades [115], or modeling cancer progression [50].

3. Frontier research area 2: Graph causal models and counterfactuals

3.1. What: Causal learning from observational data is a central problem relevant to many application domains

Causal learning from pure observational data and predictive modeling is a general problem relevant for many application domains. It is gaining much interest recently and has been largely tackled by the AI community [116,117]. There are a number of fundamental problems that have existed for a long time and have not yet been solved. The renowned American philosopher Charles S. Peirce argued that human induction must be guided by special aptitudes for guessing right, which led to the challenge of simplicity or parsimony, which is even going back to Occam's razor. Alone, the concept of simplicity poses a lot of problems for both causal machine learning [118] and causal human learning [119]. If causal inference has a rational basis, we would expect the resulting causal knowledge to allow the formulation of coherent answers to a variety of causal questions.

Two main problems about causal relationships can be distinguished in the literature: (1) "What is the probability that a cause causes (or prevents) an effect?" and (2) "What is the probability that a causal relationship exists between these two variables?" Or, put another way, "Does the cause have a nonzero probability of producing (or preventing) the effect?" [120]. The generality and wide spectrum of practical scenarios in which such questions can be formulated makes the discovery of causal relationships from data a subject under vibrant study in diverse fields and disciplines. AI-based medicine is not an exception, with specific tasks such as diagnosis and treatment calling for further advances in causality inference that unveil novel interventional and prescriptional strategies from medical data.

3.2. Why: Typically, the underlying causal model that accounts for all factors affecting an outcome variable of interest is missing

A common challenge in applying causal analysis is the lack of an underlying causal model that can account for all factors influencing an outcome variable of interest. Recent progress has been done on causal signal extraction from images [121,122]. Causality has also been applied to generative neural networks and proxy variables in an attempt to better deal with the kind of data used by Deep Learning [123,124]. Nevertheless, the international research community agrees that there are a lot of shortcomings and many open problems to be solved, for instance, dealing with the all possible underlying, and often unknown, factors of variation and variables on which causality is feasible to be studied in practice [117,125,126].

3.3. How: XAI with counterfactual explanations and causal algorithmic recourse can help determine what is causally related

Formal reasoning about causal relations between features $\mathbf{X} = [X_1, \dots, X_d]$ can be done by using a structural causal model, i.e. a non-parametric model with independent errors according to Judea Pearl [127,128]. In the following we introduce some basics to show how this can be helpful. For more extensive introductions, please refer to [120,129]. The data-generating process of $\mathbf{X}$ is described by an (unknown) underlying structural causal model $\mathcal{M}$ of the general form:

$$\mathcal{M} = (\mathbf{S}, P_{\mathbf{U}}), \quad \mathbf{S} = [X_r := f_r(\mathbf{X}_{pa(r)}, U_r)]_{r=1}^{d}, \quad P_{\mathbf{U}} = P_{U_1} \times \cdots \times P_{U_d}. \tag{1}$$

The structural equations $\mathbf{S}$ are a set of assignments generating each observed variable X_r as a deterministic function f_r of its causal parents $\mathbf{X}_{pa(r)} \subseteq \mathbf{X} \setminus X_r$ and an *unobserved* noise variable U_r. Here it is important to note that $P_{\mathbf{U}}$ is a factorizing joint distribution over background variables which introduces uncertainty due to the lack of observations. The assumption of mutually independent noises (i.e., a fully factorized $P_{\mathbf{U}}$) entails that there is no hidden confounding and is referred to as *causal sufficiency*. For an experimental proof, we refer to Karimi et al. (2020) [129].

Structural causal models are often represented by a so-called causal graph $\mathcal{G}$. Such causal graphs can be obtained by drawing a directed edge from each node in $\mathbf{X}_{pa(r)}$ to X_r for $r \in \{1, \dots, d\}$.

Figs. 2(b) and 2(c) show a typical textbook example. We assume henceforth that $\mathcal{G}$ is acyclic. In this case, the data-generating process $\mathcal{M}$ implies a unique observational distribution $P_{\mathbf{X}}$, which factorizes over $\mathcal{G}$, defined as the push-forward of $P_{\mathbf{U}}$ via $\mathbf{S}$.

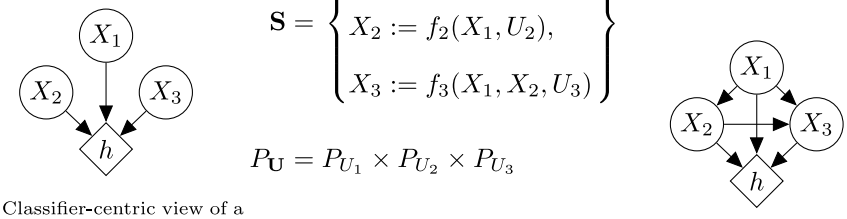
The structural causal model framework allows for the study of interventional distributions, describing a situation in which some variables are manipulated externally. The structural causal model also implies distributions over *counterfactuals*, i.e. statements about (hypothetical) interventions that were *all else being equal* (*Ceteris Paribus*, namely, the analysis of the effect of one variable on another, assuming that all other variables remain the same).

When formulated in the context of classification via a model h, a popular approach to the study of counterfactuals is to find so-called (nearest) *counterfactual explanations* [130] where the term "counterfactual" is meant in the sense of the closest possible "fact" with a different outcome. Counterfactual predictions consist of asking ourselves what would have been the effect of something if we had not taken an action, i.e., alternative scenarios [131], or modifications of the input data that could eventually alter the original prediction of the model h, and help the user understand the performance boundaries of the model for improved trust and informed criticism. Interventional clinical predictive models require the calculation of counterfactuals, apart from the correct specification of cause and effect [131]. Just to give an example, to analyze counterfactuals based on the structural causal model $\mathcal{M}$, an intervention (also known as *do* operator) can be used to indicate that a set of variables $\mathbf{X}' \subseteq \mathbf{X}$ is fixed to $\boldsymbol{\gamma}$, which is often denoted as $do(\mathbf{X}' = \boldsymbol{\gamma})$. The corresponding distribution of the remaining variables $\mathbf{X} \setminus \mathbf{X}'$ can be computed from $\mathcal{M}$ by replacing the structural equations for $\mathbf{X}' \in \mathbf{S}$ to obtain the new set of equations $\mathbf{S}(do(\mathbf{X}' = \boldsymbol{\gamma}))$. The interventional distribution $P_{\mathbf{X}'|do(\mathbf{X}'=\boldsymbol{\gamma})}$ is then given by the observational distribution implied by the manipulated structural causal model $(\mathbf{S}do(\mathbf{X}' = \boldsymbol{\gamma}), P_{\mathbf{U}})$.

Given observations $\mathbf{x}_{obs}$, the definition of the $do(\cdot)$ interventional operator permits, for example, to ask *what would have happened* if $\mathbf{X}'$ had instead taken the value $\boldsymbol{\gamma}$.

An answer to this question departs from the definition of the counterfactual variable by $\mathbf{X}(do(\mathbf{X}' = \boldsymbol{\gamma}))|\mathbf{x}_{obs}$, and the distribution of this counterfactual variable can be computed in three steps [128]:

1. *Abduction*: first compute the posterior distribution over background variables given $\mathbf{x}_{obs}$, $P_{\mathbf{U}|\mathbf{x}_{obs}}$.



(a) Classifier-centric view of a factor graph. (b) Causal model $\mathcal{M} = (\mathbf{S}, P_{\mathbf{U}})$. (c) Causal graph $\mathcal{G}$ for $\mathcal{M}$.

Fig. 2. Counterfactual explanations (a) treat features as independently manipulable inputs to a given fixed and deterministic classifier $h : \mathcal{X} \to \{1, \ldots, L\}$ trained to make decisions about i.i.d. samples from the data distribution $P_{\mathbf{X}}$. In the causal approach to algorithmic recourse taken in this work, we instead view variables as causally related to each other through a structural causal model $\mathcal{M}$ (in (b)) with associated causal graph $\mathcal{G}$ (c) [129].

2. *Action:* perform the intervention to obtain the new structural equations $\mathbf{S}^{do(\mathbf{X}'=\boldsymbol{\gamma})}$; and
3. *Prediction:* then compute the counterfactual distribution $P_{\mathbf{X}(do(\mathbf{X}'=\boldsymbol{\gamma}))|\mathbf{x}_{obs}}$ induced by the resulting structural causal model $\mathbf{S}^{do(\mathbf{X}'=\boldsymbol{\gamma})}, P_{\mathbf{U}|\mathbf{x}_{obs}}$.

Causal inference and counterfactual prediction for actionable healthcare are discussed in Prosperi et al. (2020) [131]. In medical applications, some of the tests for measuring robustness of estimated effects on non pharmaceutical interventions include intervention models doing different structural assumptions and validation of such assumptions when they do not hold. An example of such interventions against COVID-19 includes generalization over countries presented in [132]. In cases where causal effect estimation is aimed at individual-level recommendations, alerting decision makers when predictions are not to be trusted is crucial. Therefore, identifying failure with uncertainty-aware models (e.g., when covariate shift makes training and test datasets vary), as proposed in [133], facilitates uncertainty communication to decision-makers. Generally, uncertainty enables deep learning methods to be adopted into clinical workflows [134].

A different but intuitively similar concept related to the characterization of counterfactuals is that of *contrastive* explanation [135], which consists of explaining not only why an event occurred, but also why it occurred as opposed to some alternative event. They are considered necessary for agents to achieve moral responsibility, although a debate exists on contrastive explanations entailing causal determinism [136,137]. Approaches producing contrastive explanations serve to learn more efficiently from data. For example, using pertinent negatives [138] is one among such approaches, and relates to learning structural descriptions from examples. Another example is using active learning, which can help select the most informative pairs of labels to elicit contrastive natural language explanations from experts, while dynamically changing the model [139].

Equally important is the integration of "Big Data" methods with explanations that involve a causal analysis. This integrated analysis is key, especially in omics and imaging for causal inference [93]. An example of such tight integration is the use of deep feature selection for causal analysis in Alzheimer's Disease [140]. Other example is the alignment of domain expert knowledge with Deep Learning models in order to achieve more expert-compatible explainability. Neural-symbolic learning and reasoning systems can be used for this purpose with different kinds of integration schemes [68,141].

3.4. Desiderata: Disentangling influential factors from multivariate observations and plausible yet diverse counterfactuals

A concern with causal AI in medicine is how to disentangle correlated factors of influence in high-dimensional settings. One way to deal with the independent manipulation of as set of correlated factors is to disentangle the influence of correlated factors from multivariate observations with interventions. An example of such is Back-to-Back regression [142], to help identifying the causal contributions of co-linear factors in multi-variate and multi-dimensional magnetic resonance imaging observations. Back-to-Back regression produces an interpretable scalar estimate for each factor from a set of correlated factors to estimate those that most plausibly account for multidimensional observations. As a result, this method disentangles respective contributions of collinear factors to identify the causal contribution of covarying factors.

In regards to counterfactual explanations, the plausibility, feasibility, and diversity of the obtained counterfactual explanations (whether they are contrastive or not) are particularly relevant aspects that should be considered in the medical domain. In this regard we advocate for an increasing prevalence of modern generative learning approaches applied to the discovery of counterfactuals. The capability of such methods to model the distribution of existing multi-dimensional data yields a proxy generator of plausible hypothesis that can be of utmost help to ensure that counterfactual instances can occur in reality. Further along this line, the diversity of counterfactuals can be a conflicting objective with their plausibility as per $P_{\mathbf{X}}$, hence counterfactual generation methods should also properly balance among such objectives [143].

3.5. What for: Causality and counterfactual generation may reduce diagnostic results, increase quality of care and life, reduce overall costs, and free up clinicians' time

As in other fields with strong human interaction, in designing a medical AI system it is critical to consider *who* will use it. Furthermore, when the system is used for diagnostics, it is also crucial to ensure proper balance between sensitivity and specificity, and to optimize the user interface and workflow integration. There are numerous examples that support these claims from pathology, radiology and dermatology, e.g. a smartphone based melanoma classifier would likely be used by general public as a first step in screening for skin diseases.

Here the main goal – specially when the treatment for the disease to be diagnosed is invasive or has serious side effects for the health of the patient – is to maintain a low false negative rate. On the other hand, a system for radiologists should automatically classify common cases, and leave the decision on more complex cases for the expert, aiming at a high true positive rate. Properly using such systems would reduce false negatives and false positives, increase quality of treatments and quality of life of patients, decrease the overall cost and free-up clinicians' time, which becomes more critical as decision-making situations become more patient-centered [144].

Advances on graph causal modeling and counterfactuals can be a major step towards realizing such objectives. On one hand, interventional clinical studies can be driven by the results of causal analysis of multi-dimensional medical data, thereby eliciting new diagnostic and treatment criteria that in turn, produces data from such new cases

that can be fed back to the AI-based models. On the other hand, counterfactuals can increase the trustworthiness of the medical expert on the decisions issued by the AI model, discerning when it must not be fully relied as a result of a counterfactual being *close* to the case to be diagnosed/treated. This augmented information offered to the expert could reduce the amount of false positives, thereby favoring the aforementioned decrease of costs and efforts.

4. Frontier research area 3: Verification and explainability methods

4.1. What: The use of AI requires the ability to verify correctness and causal accuracy

In the medical domain, the use of AI and machine learning models that are explainable and verifiable by human medical experts is an absolute necessity, primarily for legal reasons [145]. The central problem is that no AI method will be deployed if its results cannot pass a *verification process* for correctness and causal accuracy by a human expert on demand. Making these assessments is difficult if the AI methods in question do not provide explanations to users. The problem becomes clear when we consider the classic problem described by Caruana et al. (2015) [146], where an AI system trained to predict a person's risk of pneumonia came to incorrect conclusions, and applying this model would have increased, not reduced, the number of patient deaths. At the same time, this is also a good example of the usefulness of having a human-in-the-loop [147], because physicians can easily verify the results based on their experience – namely, that such results of an AI system are not correct after all. Moreover, a human in-the-loop approach can bring in contextual understanding, implicit knowledge and experience to statistical machine learning methods, and consequently provide prior knowledge. However, one core open problem remains, namely, how to integrate this knowledge into the machine learning pipeline.

The term verification comes from both software engineering and medicine and was used in AI as well [148], the term explainability is used to technically highlight decision-relevant parts of machine representations, i.e., parts that contributed to the accuracy of a particular prediction. However, such a technical explanation does not refer to a human model. For this, explainability must be extended to include the concept of *causability* [149], which refers to a human model. Causability was introduced in reference to the well-known term of *usability* [150]. While explainability is about implementing transparency and traceability, causability is about measuring the quality of explanations, i.e., the measurable extent to which an explanation of a statement achieves a certain level of causal understanding for a user with effectiveness, efficiency, and satisfaction in a given context of use [151]. In other words, causability measures whether an explanation achieves a given level of causal understanding for a human. This is a major challenge in the medical field, as many different modalities contribute to a single outcome, requiring multimodal causability [19].

4.2. Why: The best machine learning methods to date lack robustness and are difficult to interpret

Currently, the most important and most lacking aspect of AI in general, and in medical AI in particular, is robustness. Recent success in machine learning has led to an explosion of AI applications, resulting in high expectations being placed in autonomous systems, such as autonomous vehicles [152,153], medical diagnosis [154,155], industrial prognosis [156,157], or cybersecurity [158]. These developments require that we recognize and understand the fundamental limitations of current intelligent systems, which often apply across many different application areas. This crucial deficit of robustness of current systems concretely relates to their lack of ability to adapt to changes in the environment. In medicine, this is even more profound, as data changes because of changes in patient cohorts, due to advancements of instruments and assays that generate images and omics data, and as a result of changes of treatment modalities and our understanding of health and disease states at physiological and molecular levels.

The field of machine learning deals with the development of successful adaptation strategies and attempts to enable machines to recognize or respond to changing conditions for which they have not been specifically programmed or trained. So far, however, most work in machine learning has been based on the "independent identically distributed" assumption. That is, the machine must be able to process new input data that have not been seen during training, but that they conform to the same statistical distribution. As the i.i.d. assumption is a strong assumption that is rarely met in practice, the field of machine learning is currently working extensively on theoretical and empirical approaches to develop learning strategies that do not require this assumption to hold. These efforts are particularly related to the concepts of "transfer learning" [159,160], "domain adaptation" [161–163], "adversarial training" [164–167] and "lifelong" or "continual learning" [168,169].

Even if non-i.i.d. issues are circumvented or simply do not occur, an obstacle to reach fully actionable medical AI is the lack of explainability. In particular, modern Deep Learning models that nowadays monopolize modeling approaches for medical imaging usually remain "black-boxes" [69,170,171] that are unable to explain the reasons for their predictions or recommendations. This property largely precludes the diagnosis and correction of defects, and only favors conservative safety assessments of the behavior of a learning model. Both problems are very much related to a lack of understanding of cause–effect relationships. This hallmark of human cognition is a necessary (though not sufficient) component for machine learning methods achieving human-like intelligence, which would provide the basis for a much broader application of AI in industry and business. A grand issue in the task of learning from a set of observed samples is to estimate the generalization error of learning algorithms. The problem with these typical measurements, e.g., the training error, is that they are biased, particularly if the available amount of data is small. Traditionally this is measured by *complexity measures* such as the Vapnik–Chervonenkis (VC) dimension [172,173], or stability [174].

In the race towards properly characterizing and understanding medical AI-based models, one cannot ignore the importance of providing important features for explainable models, which becomes particularly essential for image processing algorithms [140]. Furthermore, these systems need to be integrated with existing research and clinical workflows. Importantly, proper independent verification and explainability methods may highlight that well-performing AI systems are reportedly superior to humans in some clinical systems (or e.g., radiologist-level [175]), and unveil the reasons why their outperforming behavior can degrade severely in other healthcare systems as a result of potentially non-identically distributed data resulting from a context-induced bias [176].

4.3. How: Causal approaches and explainability methods can contribute to achieving target trials, transportability, and predictive invariance

From the previous section it is clear that robustness is a key aspect to be addressed in medical AI-based systems. Performance guarantees can only be given if models are proven to be robust against different phenomena that compromise their generalization capability. An interesting approach to study generalization of learning algorithms from the perspective of robustness was presented in [177], which derived generalization bounds for learning algorithms based on their *algorithmic robustness*. The assumption is that if a testing sample is "similar" to a training sample, then the testing error is close to the training error, which is different from the traditional complexity or stability arguments mentioned earlier that concentrate on solely optimizing pure performance measurements.

Indeed, in the machine learning community the overall trending goal seems to be maximizing standard accuracy, and many papers from the biomedical domain report increasing accuracy levels for different medical diagnostic tasks by virtue of models of increasing complexity and sophistication. However, such models still yield erroneous cases, which should motivate doctors to retrace and find the rooting cause of such errors. However, a non-automated inspection and verification of such cases is often unfeasible due to the multi-modality of data and the efforts it requires from the medical expert. At this point a new opportunity arises for causality and explainability as enablers to automate this medical verification process.

Unfortunately, observational biomedical studies are affected by confounding and selection biases among other biases [178], which makes causal inference infeasible unless robust assumptions are made. These require *a priori* domain knowledge, as data-driven predictive models can be used to infer causal effects. However, neither their parameters nor their predictions necessarily have a causal interpretation.

Consequently, we firmly call for the use of causal approaches and learning causal structures by using certain *linchpins* to develop and test intervention models [131], namely: (1) target trials, (2) transportability, and (3) prediction invariance. To begin with, target trials refer to algorithmic emulation of randomized studies. Transportability [179] is a *license* to "transfer causal effects learned in experimental studies to a new population, in which only observational studies can be conducted". Akin to transportability is prediction invariance, where a "true causal model is contained in all prediction models whose accuracy does not vary across different settings". When a causal structure is available or a target trial design can be devised, the evaluation of model transportability for a given set of action queries (e.g., treatment options or risk modifiers) is recommended; while for exploratory analyses where causal structures are to be discovered, prediction invariance could be used. In this way, as advocated by Prosperi et al. (2020) [131], transportability and prediction invariance could become guideline core tools and part of reporting protocols for intervention models, for a better alignment with the standards for prognostic and diagnostic models of medicine and biomedical practice today.

Another phenomenon placing at risk the trustworthiness and verification of medical AI models is their robustness to adversarial attacks. Technically, we assume a model processing unseen examples from the underlying distribution $P_{\mathbf{X}}$. In general, the goal of model training is to reach a minimum of a expected loss function [180]. However, many machine learning models, particularly deep neural networks [181], are susceptible to be deceived by the presence of adversarial examples [182]. Adversarial examples can be conceived as modified data instances resulting from small yet intelligently tailored perturbations made to original examples. Even if they are not even visible to the human eye, such perturbations yield dramatic effects when processed through the machine learning model, provoking a wrong output with high confidence.

Fig. 3 depicts a schematic diagram showing the different reasons by which model verification and robustness assessment are of utmost necessity in the medical domain. XAI methods can help determining what a model observes in an input when predicting its output, ascertaining the presence of biases inherited from data or purposely inserted by adversarial attacks. Likewise, counterfactual explanations can also benefit for stronger input–output causal relationships discovered from data, stepping beyond the production of correlation-based counterfactuals to the generation of interventional what-if stories. This might be a major step in the medical AI field to trascend from verifiable models for diagnosis towards verifiable AI-based solutions for medical prescription and treatment.

A pause must be done before proceeding further to highlight, once again, the importance of having a human-in-the-loop as the ultimate stakeholder to decide whether an AI-based model is robust enough [147]. Even if the verification process can be partly automated by XAI and causality inference methods, trustworthiness always requires a qualitative assessment of the overall verification process, both in terms of their starting assumptions (e.g. *is a certain adversarial attack strategy for a medical AI-based model plausible and likely to occur in the context in which data are produced?*) and the results it conveys (corr. *is the detected bias inherited from data? Can we reduce this bias by preprocessing or improving the data collection process anyhow?*). All in all, humans, even if we make mistakes, can be considered a robust proxy in decision making when informed with quantitative and well-summarized measures of algorithmic robustness.

4.4. Desiderata: Adversarial training can contribute to better robustness and explainability

A very different use of adversarial training is to make models more robust and interpretable. The work in [183] shows that adversarial training improves the interpretability of gradient-based saliency maps in medical imaging diagnosis of skin cancer. In particular, adversarially trained convolutional neural networks are significantly sharper and more visually coherent than non-adversarial traditionally trained CNNs. What many of these robustness tests highlight is the needs for verification and validation methods for deep learning techniques beyond academic toy datasets. It is clear that much of the research efforts have focused on overfitting deep learning models with ever-increasing numbers of parameters to a small selection of research benchmark datasets [184].

Even results reported in carefully curated international challenges such as PASCAL VOC [185] later turned out to be largely based on spurious correlations (e.g., ships were classified by the presence of water, or horses were linked to copyright watermarks). In a similar vein, popular text classification datasets have been shown to contain biases, meaning that only parts of the input are needed to make the correct predictions [186]. This type of cheating is also referred to as "Clever Hans effect" [187].

In spite of permitting the incremental improvement and incredible advances in the field, natural image datasets can normally be very different from real life datasets, which are more sparse, noisy and in uncontrolled settings. Language differences aside, similar conclusions can be derived from medical text data collected in diverse environments, which ground on cultural, geographical or individually-induced biases present in such data. Generalizing to real life datasets is thus a part of the desiderata of having robust machine learning models for medical application. For this to occur, we envision that explainability tools will become increasingly relevant, becoming a core part of prospective studies reporting successful real-world cases.

4.5. What for: The most important practical benefit of implementing this FRA theme is maintaining trust

In the medical domain, the use of AI methods that are verifiable, comprehensible and interpretable by human experts will not only be mandatory for legal reasons in the future, but also offers a number of other technical and non-technical advantages. Advantages from the technical point of view include that developers get a better understanding of the medical system endowed with AI-based functionalities, thus are able to improve existing methods (e.g., by reducing complexity or model size) with increased knowledge about the niches and directions along which such improvements can be attained. Bias identification [188] or adversarial attack detection [189] can be arguably the most evident examples of technical advantages granted by XAI methods for model verification.

Above all, the big advantage for the medical expert and the end user affected by decisions issued by verified medical AI models lies in the increased trust on their outcomes, the remaining responsibility of the human being (human-in-control) and the avoidance of bias and discrimination. Medical decisions can pose a turning point in the life of a patient, so trustworthiness on the suitability of decisions issued by

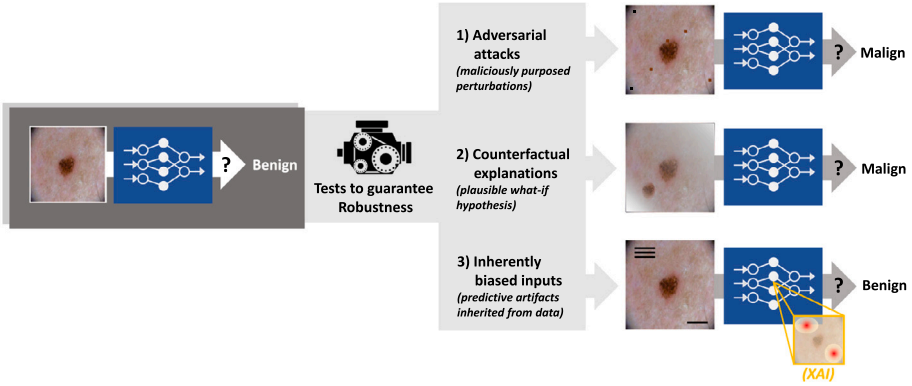


Fig. 3. Schematic diagram exemplifying the different circumstances under which robustness of a medical AI-based systems (in this case, for diagnosing a melanoma) must be verified: adversarial attacks, counterfactual explanations and biases. Causality inference and explainability methods can enable automated means to perform such a verification procedure.

such models is a must at many different levels of the medical workflow, from the diagnosis (confidence of predictions), to the design of the treatment (suitability of prescribed therapies/medication by a model) and the acceptability of the patient (causability to ensure that he/she understand that the AI-informed decisions are the best ones for his/her disease). When understanding this need for trustworthiness at multiple levels of the medical workflow, one can realize the enormous relevance of AI verification and explainability in the medical realm.

5. A unified view on the integrative role of information fusion in medical AI

Encoding multidimensional data, but also tabular data and data of temporal sequential nature, is an open challenge for the latest DL models to assimilate incomplete and irregular healthcare data. Reinforcement learning and explainable models to fully control this family of AI black-box models [190] can better use this data for sequential decision making from observational multi-modal data if meaningful representations are learned and used to represent a patient state [191].

In this context, local and global explanations are equally important, i.e., assessing machine learning model output with respect to a single input data point, also called "decision understanding" (e.g., as done by methods such as Local Interpretable Model-Agnostic Explanations — LIME [192] or Layer-wise Relevance Propagation — LRP [193]), but also verifying and certifying the full model at a global scale, also called "model understanding" [194]. Likewise, [195] advocates for explanations in cooperative decision making in medicine to be mutual, implicitly implying a continual fusion of explanations. Mutual explanations [196] are introduced in a context of transparent expert companions towards medical decision support systems where interactive and explainable HRI [197] machine learning plays a key role. Mutual explanations naturally provide the understanding of verbal explanations, i.e., based on dialog incremental processes to provide human machine learning users with trust and deeper involvement in the learning process. When explanations are not accepted, the human cannot only ask for them but also correct them. This way, expert domain knowledge is used in learning and inference through explanation sketches that are applied as constraints for the inductive logic programming system Aleph.

Verbal interpretability perspective [198] is achieved by ensuring that the model is capable of providing humanly understandable statements, e.g., logical relations, showing positive words drawing to a conclusion, verbal chunks or sentences [199] that indicate causality, and that the model produces explanations which are non-contradictory, non-redundant, fluent and cover all important aspects related to the prediction [200].

Also related to human expert alignment are the needs for developing models for clinical acceptance. An example of such good practice is shown in [201], where such acceptance test is done through ratings by ophthalmologists on the correlation of the attribution method scores with diagnostic features. In this context, in addition to local explainable models of a single sample, approaches to test global explanations such as TCAV (Testing with Concept Activation Vector) [202] or SpRAy (Spectral Relevance Analysis) [187] are desired in order to explain beyond a single data point example. However, they may not be fully considered as global method, as they only consider the set of all training examples from a given class [198]. Another critique of current Natural Language Processing (NLP) models provided with verbal interpretability is the lack of provision of the actual underlying mechanisms to generate texts. Generating free text explanations is often framed as a summarization task — either as extractive settings, where salient sentences from provided evidence documents are selected as explanations [200], or abstractive settings, where, given evidence documents, the explanation is produced from scratch using a generative model [203]. While the latter can result in more fluent explanations and incorporate further background knowledge not explicitly present in the evidence documents, it is known that, as for example used for EHR generation from conversations in [204], fake facts are hallucinated by neural generators [205]. Yet other works rely on hybrid approaches, where extractive summarization is followed by abstractive summarization [206,207]. However, as also advocated by [198], further work on providing explanations of the process and shape of the embedding optimization is needed.

The role of natural language in information fusion and XAI is twofold: on the one hand, language is one of the data modalities, in which complex facts and relationships are expressed, e.g. in electronic health records (EHRs) or medical literature. On the other hand, language is the prime channel of explanation: verbalizing the algorithmic reasoning enables the health practitioner to easily detect whether the reason for the algorithmic decision is acceptable.

For both variants, the use of cross-modal representations that link, e.g., textual, image and omics data will be crucial for AI in multimodal data as present widely in the medical domain. Challenges lie in the harmonization and curation of cross-modal datasets aligned across two or more modalities enabling the cross-modal transfer, either by learning a common subspace via methods such as DCCA [208] or by projection learning [209]. While suitable datasets are becoming available in the public domain, they are yet to be constructed for medical data.

For processing and generating language in a transparent way, future work will have to concentrate on NLP models with provenance, i.e., models that provide the data on which their output is based on. In the case of automatic summarization, for example, this would be the statements that lead to the formulation of a summarizing sentence; for semantic processing it could be the use of hybrid models that combine sparse representations [74] with dense representations, e.g., [210]. For

Transformer-based architectures (e.g., [159]), in the absence of human rationales to train a model to generate explanations, this could be realized with attention scores, although they only loosely correspond to human-acceptable explanations [13,211,212]. An alternative could be to investigate the utility of diagnostic properties, such as Faithfulness, Dataset Consistency and Confidence Indication [71]. These have been shown to be useful for automatically evaluating the quality of explanations, and might be suitable as objectives for generating explanations in an unsupervised way. Another option is the use of (intransparent) NLP technologies to identify and extract information with provenance, as for example done in [213] for metadata extraction from biomedical literature to increase reproducibility of studies.

Metrics worth assessing beyond model understanding through subspace explanation (MUSE) induce fidelity (based on instances disagreement between model and explanation), unambiguity (in terms of rule overlap and cover), or interpretability (in terms of triple rule set size, width, and predicate size) [214].

One strand of future methods strives for high quality data in order to produce better predictions, the requirements to deploy AI systems in medicine advocate as well for natural handling of noisy and incomplete data, which is much more realistic in healthcare, where many information silos due to the distributed nature of domain expert knowledge bases and respective EHR. In this line, techniques to complete partial data from missing sensor readings through data level- and feature level information fusion to improve the overall data quality include, for instance, kernel random forests in fog computing for heart disease prediction [215]. Another example showing improved results with extra fused data includes the use of self-attention architectures for CT-image and non visual features for immunotherapy treatment response prediction [216]. In fog computing, a similar approach to federated learning in terms of data decentralization, the ability to access all data at once is not possible. However, fusing the different sensors available for different users makes all data actionable [131], and the full set richer, and of better quality. Recent work showed that it is even possible to train largely personalized models in such distributed settings [217]. Other strand of ideology advocates for approaches that incorporate a natural handling for anomalies and outliers [218], as well as incomplete, dirty and irregular datasets, as a common feature of medical AI systems [219]. The latter work also warns for the potentially large impact of unintended consequences of machine learning in medicine from an empirical and technical viewpoint. These and other pitfalls in data-driven decision making [131] are to be considered in the development of the frontier topics discussed in this paper, hand in hand with experts-in-the-loop.

Integrative computational biology and AI algorithms play a central role in precision medicine. Individual analyses can be combined using multiple networks, including transcription regulatory, microRNA-gene, physical protein interactions, metabolic and signaling pathways [220]. Such analyses help identify better prognostic and predictive signatures, drug mechanism of action, combination therapies, and possible novel drug targets. These networks can be further annotated with tissues and diseases to form richly-annotated typed graphs, which in turn can be analyzed with graph theory algorithms to form explainable models. For example, Bhattacharyya and colleagues integrated a pathway-based patient model with multi-scale Bayesian network to predict specific treatment options [221]. Similarly, exploring the possible links between AKT1 (Akt is a Protein kinase B that plays a key role in glucose metabolism, apoptosis, cell proliferation, transcription and cell migration) and BTK (Bruton's tyrosine kinase that plays a crucial role in B cell development and signaling), we obtain 1,862 proteins connected by 2,324 edges (i.e., direct physical protein interactions, 437 unidirectional, 84 bi-directional and the rest non-directional), as shown in Fig. 4. The network in this figure highlights which of the interactions are relevant to arthritis, neuro-degenerative diseases, or cognitive disorders.

Importantly, once a hypothesis and model are created from an integrative analysis, such as the one highlighted in Fig. 4, one would need to select the most appropriate – and ideally, the least costly – organism to act as the model for further functional studies and validation. Considering this network, the mouse would be the best model organism, as about 98% of all interactions in the network are conserved from human to mouse, while the rabbit has only 33% of the network conserved, and fly, worm and yeast have none of these interactions present (Fig. 4, a). Using analogous selection, the most relevant tissues for functional validation include adipose, lung, spleen and bone (81%–85%), falling to just around 50% for heart and brain (Fig. 4, b). Considering diseases, only cancer has a substantial set of annotated interactions in this network with almost 60% of the network being annotated to diverse cancers (Fig. 4, c). (See Fig. 5.)

6. Discussion

As we have seen in previous sections, for AI models in medicine, there are several concerns with respect to the development of these frontier topics. Besides them, another dimension with large concerns in medicine whose importance can be exacerbated upon the fusion of multimodal data is the privacy and confidentiality awareness of medical AI-based models. Indeed, the compliance with patient privacy normally hinders medical AI methods from excelling in practical settings due to a diversity of reasons, such as the increased difficulty of collecting data, restrictions to their use following ethical and legal constraints, or the potential performance penalty obtained when data are encoded prior to modeling. Ideas using the concept of differential privacy [222], privacy-preserving representations [223] or along the lines of privacy distillation [224] are key to further develop this line of work. Privacy distillation [225] allows patients to decide the type and amount of information they disclose to healthcare information systems while retaining the model accuracy under a sufficient subset of original privacy-relevant features. The idea behind this model-agnostic mechanism is to balance accuracy of the model with the redacted inputs of users. An example of application in a DL regression setting for dose prediction is in [225]; it demonstrates to reduce the amount of over-prescriptions and under-prescriptions of warfarin. To sum up, we foresee that the growing amount and diversity of patient, medical and clinical information combined and flowing together into medical processes relying on AI-based models will give rise to unprecedented challenges in what relates to the privacy of sensitive data, calling for overarching strategies that maintain the confidentiality of protected information of the patient all over the process.

One size does not fit all. While AI can solve standard cases with similar accuracy to human experts, it cannot yet beat human specialists. However, we rather stand with the synergy that flourishes when AI and the specialist collaborate together, feeding each other with knowledge that allow them performing better, more robustly and reliably in their respective tasks. Human-in-the-loop systems would benefit from AI approaches, and even more from an ensemble of AI systems, implemented using different approaches and algorithms, and trained and validated on different patient cohorts. Conversely, AI-based systems can leverage the qualitative verification of the knowledge captured from data, as well as the conformity of explanations with the medical expertise and the evidence recorded over the medical workflow.

Ignoring the implications of improper usability planning may lead to incorrect results and reduced applicability. This requires one to weigh up sensitivity with specificity to ensure specialist vs general use cases or screening vs treatment planning. It is also important to ensure clear understanding of limitations based on validation — which patient cohorts may or may not be appropriate for a given trained model. Besides explicitly acknowledging and recognizing the limitations of these AI models and resulting systems, patient-centric medicine requires models to provide specific confidence and uncertainty estimates

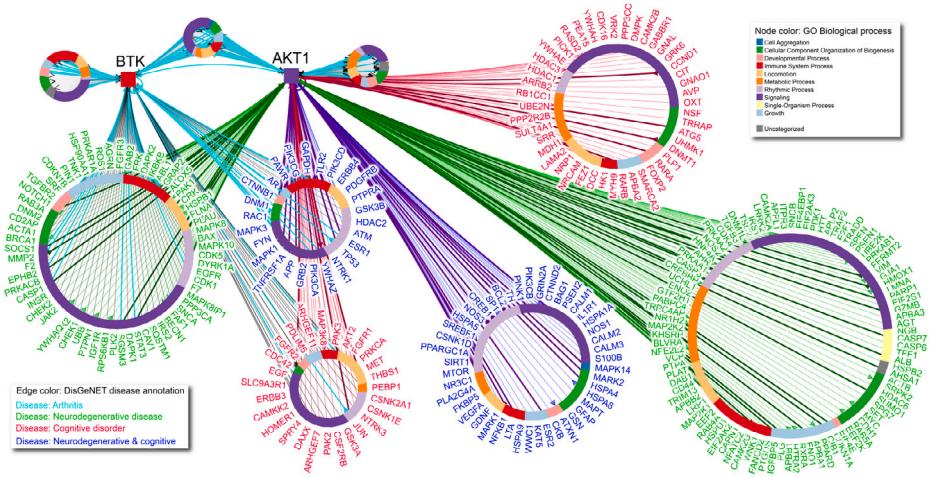


Fig. 4. Exploring the connection between AKT1 and BTK. The physical protein interaction network from the Integrated Interactions Database (IID v.2020-05) [37] highlights the Gene Ontology biological process (node color) and disease annotation from DisGeNET (edge color); specifically, arthritis, neurodegenerative diseases, cognitive disorders, and their overlap (thicker, darker color edges).

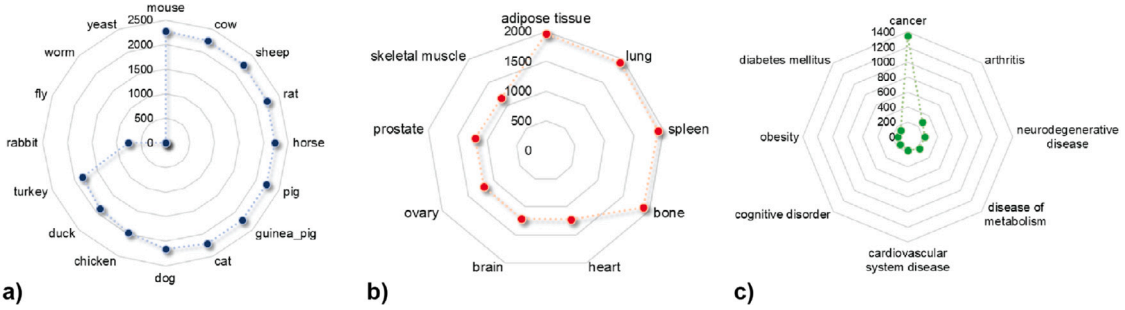


Fig. 5. Conservation of the physical protein interaction network from Fig. 4 across (a) non-human species, (b) tissues, and (c) diseases.

on the recommendation for each patient, rather than simply provide broad accuracy measures across cohorts.

To realize this holistic vision, it is important that ongoing studies dealing with medical AI are verified swiftly, providing informed evidence that AI-based models for medical practice can be trusted. On the other side of the coin, research retractions should be managed and resolved quickly, as done in recent COVID-19 related research contributions (e.g., Mehra et al. (2020) [226] in *The New England Journal of Medicine* and *Lancet*, Mulvey et al. (2020) [227] in *Annals of Diagnostic Pathology*, and Zeng et al. (2018) [228] in *Lancet-Global Health*.

However, the process takes a long time — mistakes are usually detected and retracted within months, but fraud often takes years [229]. This has direct, negative implication for evidence-based medicine, and a significant impact on computational biology and AI. Considering requirements for training and validation of AI systems, data from retracted papers may affect large number of workflows and analyses, leading to incorrect models and interpretations. Training or validating AI systems on flawed data may not be obvious immediately, and even when the paper is retracted, data will likely exist in multiple forms on the Web for years after.

To circumvent this latter issue, online data repositories are crucial, but stringent curation processes are essential to ensure high quality, reliable and properly annotated data. For example, the IMEx consortium [24,25,230–232] curates interaction data from published literature to enable integrative computational biology analyses, and ensure the implementation of data-driven medicine and the correct analysis and interpretation of model results. The availability of such curated repositories, and evidences of real-world AI-based models that largely rely on advances over the frontier topics reviewed in this position paper would free-up specialists by solving straightforward cases automatically, and comprehensively characterizing complex cases for further consideration and inspection.

Finally, the important element of visualization must be included, because it is ultimately what is presented to the expert end user [233]. The findings and knowledge from the long-established domain of visual analytics [234] must therefore be comprehensively taken into account and integrated into future overall solutions [235] to build new human–AI interfaces supporting explainability and causability [236].

7. Conclusion

From our experience, we have identified and outlined three key Frontier Research Areas that need to be developed hand-in-hand within AI and the application fields. These frontier topics would benefit enormously from a Frontier Development Lab, an example of successful implementation being the SETI-NASA-ESA FDL program for AI, Space and Earth Sciences, which benefits from a catalytic environment for tackling some of the most challenging interdisciplinary research problems. In similar synergy, future biomedical AI would benefit from cross-domain research teams solving challenges in the context of cross-science problems. New additional PhD schools that take such a research-based approach can be of help.

Experts at the seam between AI and medicine are urgently needed worldwide. In the European Union, there is a dramatic shortage of qualified experts who understand both domains. Industry is desperate for properly trained professionals [237]. Additionally, these experts need also to understand ethical and legal issues, and how and by whom AI is used. This requires that future experts are not only theoretically educated in ethical and legal aspects, but also given the opportunity to

put them into practice in both healthcare institutions and industry. This is where agile, human-centered AI design methods can be beneficial (refer to Fig. 1).

In our holistic vision of medical AI, we highlight the cohesive role of information fusion as a technology to transport all medical data modalities through the frontier research areas. New challenges around multi-modal explanations, causality (cause–effect) and causability (quality of explanations) analysis are still to be addressed by the research community for achieving full trustworthy and robust medical AI-based systems and the use of new types of human-AI interfaces and supportive visualizations.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

We are very grateful for the valuable comments of the anonymous reviewers and the editor, which helped us a lot in revising this paper. Furthermore, we thank our colleagues from the international research community for their critical comments and suggestions for improvement.

Andreas Holzinger acknowledges funding support from the Austrian Science Fund (FWF), Project: P-32554 explainable Artificial Intelligence and from the European Union's Horizon 2020 research and innovation program under grant agreement 826078 (Feature Cloud). This publication reflects only the authors' view and the European Commission is not responsible for any use that may be made of the information it contains; Natalia Díaz-Rodríguez is supported by the Spanish Government Juan de la Cierva Incorporación contract (IJC2019-039152-I); Isabelle Augenstein's research is partially funded by a DFF Sapere Aude research leader grant; Javier Del Ser acknowledges funding support from the Basque Government through the ELKARTEK program (3KIA project, KK-2020/00049) and the consolidated research group MATHMODE (ref. T1294-19); Wojciech Samek acknowledges funding support from the European Union's Horizon 2020 research and innovation program under grant agreement No. 965221 (iToBoS), and the German Federal Ministry of Education and Research (ref. 01IS18025 A, ref. 01IS18037I and ref. 0310L0207C); Igor Jurisica acknowledges funding support from Ontario Research Fund (RDI 34876), Natural Sciences Research Council (NSERC 203475), CIHR Research Grant (93579), Canada Foundation for Innovation (CFI 29272, 225404, 33536), IBM, Ian Lawson van Toch Fund, the Schroeder Arthritis Institute via the Toronto General and Western Hospital Foundation.

References

[1] J.H. Thrall, X. Li, Q. Li, C. Cruz, S. Do, K. Dreyer, J. Brink, Artificial intelligence and machine learning in radiology: Opportunities, challenges, pitfalls, and criteria for success, J. Am. Coll. Radiol. 15 (3) (2018) 504–508, http://dx.doi.org/10.1016/j.jacr.2017.12.026.

[2] K.-H. Yu, A.L. Beam, I.S. Kohane, Artificial intelligence in healthcare, Nat. Biomed. Eng. 2 (10) (2018) 719–731, http://dx.doi.org/10.1038/s41551-018-0305-z.

[3] X. Mei, H.-C. Lee, K.-y. Diao, M. Huang, B. Lin, C. Liu, Z. Xie, Y. Ma, P.M. Robson, M. Chung, Artificial intelligence–enabled rapid diagnosis of patients with covid-19, Nat. Med. 26 (8) (2020) 1224–1228, http://dx.doi.org/10.1038/s41591-020-0931-3.

[4] J.E. Arco, A. Ortiz, J. Ramirez, F.J. Martínez-Murcia, Y.-D. Zhang, J.M. Górriz, Uncertainty-driven ensembles of deep architectures for multiclass classification. Application to covid-19 diagnosis in chest x-ray images, arXiv:2011.14894.

[5] F.J. Martinez-Murcia, A. Ortiz, J. Ramírez, J.M. Górriz, R. Cruz, Deep residual transfer learning for automatic diagnosis and grading of diabetic retinopathy, Neurocomputing.

[6] D. Grapov, J. Fahrmann, K. Wanichthanarak, S. Khoomrung, Rise of deep learning for genomic, proteomic, and metabolomic data integration in precision medicine, omics: J. Integr. Biol. 22 (10) (2018) 630–636.

[7] Z. Zhang, Y. Zhao, X. Liao, W. Shi, K. Li, Q. Zou, S. Peng, Deep learning in omics: A survey and guideline, Brief. Funct. Genom. 18 (1) (2019) 41–57.

[8] K. Chaudhary, O.B. Poirion, L. Lu, L.X. Garmire, Deep learning–based multi-omics integration robustly predicts survival in liver cancer, Clin. Cancer Res. 24 (6) (2018) 1248–1259.

[9] J. Martorell-Marugán, S. Tabik, Y. Benhammou, C. del Val, I. Zwir, F. Herrera, P. Carmona-Sáez, Deep Learning in Omics Data Analysis and Precision Medicine, Exon Publications, 2019, pp. 37–53.

[10] A. Farhangfar, R. Greiner, C. Szepesvári, Learning to segment from a few well-selected training images, in: L. Bottou, M. Littman (Eds.), Proceedings of the 26th Annual International Conference on Machine Learning, ICML 2009, 2009, pp. 305–312, http://dx.doi.org/10.1145/1553374.1553413.

[11] N. Papernot, P. McDaniel, I. Goodfellow, S. Jha, Z.B. Celik, A. Swami, Practical black-box attacks against machine learning, in: Proceedings of the 2017 ACM on Asia COnference on Computer and Communications Security, 2017, pp. 506–519.

[12] S. Jain, B.C. Wallace, Attention is not explanation, arXiv preprint arXiv:1902.10186.

[13] S. Wiegreffe, Y. Pinter, Attention is not not explanation, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP, in: Association for Computational Linguistics, Hong Kong, China, 2019, pp. 11–20, http://dx.doi.org/10.18653/v1/D19-1002, URL https://www.aclweb.org/anthology/D19-1002.

[14] L. Steels, R. López de Mantaras, The barcelona declaration for the proper development and usage of artificial intelligence in europe, AI Commun. 31 (6) (2018) 485–494, http://dx.doi.org/10.3233/AIC-180607.

[15] J. Crowley, A. Oulasvirta, J. Shawe-Taylor, M. Chetouani, B. O'Sullivan, A. Paiva, A. Nowak, C. Jonker, D. Pedreschi, F. Giannotti, F. van Harmelen, J. Hajic, J. van den Hoven, R. Chatila, Y. Rogers, Toward ai systems that augment and empower humans by understanding us, our society and the world around us, Report of 761758 EU Project HumaneAI (available online) (2019) 1–32.

[16] D. Schneeberger, K. Stoeger, A. Holzinger, The european legal framework for medical ai, in: International Cross-Domain Conference for Machine Learning and Knowledge Extraction, Fourth IFIP TC 5, TC 12, WG 8.4, WG 8.9, WG 12.9 International Cross-Domain Conference, CD-MAKE 2020, Proceedings, Springer, Cham, 2020, pp. 209–226, http://dx.doi.org/10.1007/978-3-030-57321-8-12.

[17] H. Mueller, M.T. Mayrhofer, E.-B.V. Veen, A. Holzinger, The ten commandments of ethical medical ai, IEEE Comput. 54 (7) (2021) 119–123, http://dx.doi.org/10.1109/MC.2021.3074263.

[18] Z. Hussain, W. Slany, A. Holzinger, Investigating agile user-centered design in practice: A grounded theory perspective, in: Springer Lecture Notes in Computer Science LNCS, vol. 5889, Springer, Berlin Heidelberg, 2009, pp. 279–289, http://dx.doi.org/10.1007/978-3-642-10308-7_19.

[19] A. Holzinger, B. Malle, A. Saranti, B. Pfeifer, Towards multi-modal causability with graph neural networks enabling information fusion for explainable ai, Inf. Fusion 71 (7) (2021) 28–37, http://dx.doi.org/10.1016/j.inffus.2021.01.008.

[20] M. Zitnik, B. Zupan, Data fusion by matrix factorization, IEEE Trans. Pattern Anal. Mach. Intell. 37 (1) (2015) 41–53, http://dx.doi.org/10.1109/TPAMI.2014.2343973.

[21] M. Zitnik, F. Nguyen, B. Wang, J. Leskovec, A. Goldenberg, M.M. Hoffman, Machine learning for integrating data in biology and medicine: Principles, practice, and opportunities, Inf. Fusion 50 (10) (2019) 71–91, http://dx.doi.org/10.1016/j.inffus.2018.09.012.

[22] B. Cox, M. Kotlyar, A.I. Evangelou, V. Ignatchenko, A. Ignatchenko, K. Whiteley, I. Jurisica, S.L. Adamson, J. Rossant, T. Kislinger, Comparative systems biology of human and mouse as a tool to guide the modeling of human placental pathology, Mol. Syst. Biol. 5 (2009) 279.

[23] A.J. Enright, I. Iliopoulos, N.C. Kyrpides, C.A. Ouzounis, Protein interaction maps for complete genomes based on gene fusion events, Nature 402 (6757) (1999) 86–90.

[24] IMEx Consortium, http://www.imexconsortium.org/, (Online; Accessed 17 December 2020).

[25] N. Del-Toro, M. Duesbury, M. Koch, L. Perfetto, A. Shrivastava, D. Ochoa, O. Wagih, J. Pinero, M. Kotlyar, C. Pastrello, P. Beltrao, L. Furlong, I. Jurisica, H. Hermjakob, S. Orchard, P. Porras, Capturing variation impact on molecular interactions in the IMEx consortium mutations data set, Nature Methods 9 (4) (2012) 345–350.

[26] Gene Ontology Consortium, Creating the gene ontology resource: Design and implementation, Genome Res. 11 (8) (2001) 1425–1433.

[27] D.L. Wheeler, T. Barrett, D.A. Benson, S.H. Bryant, K. Canese, V. Chetvernin, D.M. Church, M. Dicuccio, R. Edgar, S. Federhen, L.Y. Geer, Y. Kapustin, O. Khovayko, D. Landsman, D.J. Lipman, T.L. Madden, D.R. Maglott, J. Ostell, V. Miller, K.D. Pruitt, G.D. Schuler, E. Sequeira, S.T. Sherry, K. Sirotkin, A. Souvorov, G. Starchenko, R.L. Tatusov, T.A. Tatusova, L. Wagner, E. Yaschenko, Database resources of the national center for biotechnology information, Nucl. Acids Res. 36 (Database issue) (2008) D13–21, http://dx.doi.org/10.1093/nar/gkl1031.

[28] V.A. McKusick, Mendelian inheritance in man and its online version, OMIM, Am. J. Hum. Genet. 80 (4) (2007) 588–604.

[29] L. Matthews, G. Gopinath, M. Gillespie, M. Caudy, D. Croft, B. de Bono, P. Garapati, J. Hemish, H. Hermjakob, B. Jassal, A. Kanapin, S. Lewis, S. Mahajan, B. May, E. Schmidt, I. Vastrik, G. Wu, E. Birney, L. Stein, P. D'Eustachio, Reactome knowledgebase of human biological pathways and processes, Nucl. Acids Res. 37 (Database issue) (2009) D619–22.

[30] UniProt Consortium, The universal protein resource (UniProt) 2009, Nucl. Acids Res. 37 (2009) D169–74, (Database issue).

[31] M. Uhlen, P. Oksvold, L. Fagerberg, E. Lundberg, K. Jonasson, M. Forsberg, M. Zwahlen, C. Kampf, K. Wester, S. Hober, H. Wernerus, L. Bjarling, F. Ponten, Towards a knowledge-based human protein atlas, Nat. Biotechnol. 28 (12) (2010) 1248–1250, URL http://www.ncbi.nlm.nih.gov/pubmed/21139605.

[32] Network The Cancer Genome Atlas Research, J.N. Weinstein, E.A. Collisson, G.B. Mills, K.R.M. Shaw, B.A. Ozenberger, K. Ellrott, I. Shmulevich, C. Sander, J.M. Stuart, The cancer genome atlas pan-cancer analysis project, Nat. Genet. 45 (10) (2013) 1113–1120, http://dx.doi.org/10.1038/ng.2764, commentary.

[33] S.A. Forbes, D. Beare, P. Gunasekaran, K. Leung, N. Bindal, H. Boutselakis, M. Ding, S. Bamford, C. Cole, S. Ward, C.Y. Kok, M. Jia, T. De, J.W. Teague, M.R. Stratton, U. McDermott, P.J. Campbell, Cosmic: Exploring the world's knowledge of somatic mutations in human cancer, Nucleic Acids Res. 43 (D1) (2015) D805–D811, http://dx.doi.org/10.1093/nar/gku1075, arXiv:http://nar.oxfordjournals.org/content/43/D1/D805.full.pdf+html, URL http://nar.oxfordjournals.org/content/43/D1/D805.abstract.

[34] The Gene Ontology Consortium, Gene ontology consortium: Going forward, Nucleic Acids Res. 43 (D1) (2015) D1049–D1056, http://dx.doi.org/10.1093/nar/gku1179, arXiv:http://nar.oxfordjournals.org/content/43/D1/D1049.full.pdf+html, URL http://nar.oxfordjournals.org/content/43/D1/D1049.abstract.

[35] The UniProt Consortium, Uniprot: A hub for protein information, Nucleic Acids Res. 43 (D1) (2015) D204–D212, http://dx.doi.org/10.1093/nar/gku989, arXiv:http://nar.oxfordjournals.org/content/43/D1/D204.full.pdf+html, URL http://nar.oxfordjournals.org/content/43/D1/D204.abstract.

[36] S. Rahmati, M. Abovsky, C. Pastrello, M. Kotlyar, R. Lu, C.A. Cumbaa, P. Rahman, V. Chandran, I. Jurisica, pathdip 4: An extended pathway annotations and enrichment analysis resource for human, model organisms and domesticated species, Nucleic Acids Res. 48 (D1) (2020) D479–D488.

[37] M. Kotlyar, C. Pastrello, Z. Malik, I. Jurisica, Iid 2018 update: Context-specific physical protein-protein interactions in human, model organisms and domesticated species, Nucleic Acids Res. 47 (D1) (2019) D581–D589.

[38] T. Tokar, A. Hauschild, C. Pastrello, A. Rossos, I. Jurisica, mirdip v4.0 – integrative microrna targets prediction and tissue-specificity annotation, Nucl. Acids Res. 46 (D1) (2018) D360–D370.

[39] P. Braun, M. Tasan, M. Dreze, M. Barrios-Rodiles, I. Lemmens, H. Yu, J.M. Sahalie, R.R. Murray, L. Roncari, A.S. de Smet, K. Venkatesan, J.F. Rual, J. Vandenhaute, M.E. Cusick, T. Pawson, D.E. Hill, J. Tavernier, J.L. Wrana, F.P. Roth, M. Vidal, An experimentally derived confidence score for binary protein-protein interactions, Nat. Methods 6 (1) (2009) 91–107.

[40] M. Kotlyar, C. Pastrello, F. Pivetta, A. Lo Sardo, C. Cumbaa, H. Li, T. Naranian, Y. Niu, Z. Ding, F. Vafaee, F. Broackes-Carter, J. Petschnigg, G.B. Mills, A. Jurisicova, I. Stagljar, R. Maestro, I. Jurisica, In silico prediction of physical protein interactions and characterization of interactome orphans, Nature Methods 12 (1) (2015) 79–84.

[41] B. Azarkhalili, A. Saberi, H. Chitsaz, A. Sharifi-Zarchi, Deepathology: Deep multi-task learning for inferring molecular pathology from cancer transcriptome, Sci. Rep. 9 (1) (2019) 1–14.

[42] K.S.S. Enfield, E.A. Marshall, C. Anderson, K.W. Ng, S. Rahmati, Z. Xu, M. Fuller, K. Milne, D. Lu, R. Shi, D.A. Rowbotham, D.D. Becker-Santos, F.D. Johnson, J.C. English, C.E. MacAulay, S. Lam, W.W. Lockwood, R. Chari, A. Karsan, I. Jurisica, W.L. Lam, Epithelial tumor suppressor elf3 is a lineage-specific amplified oncogene in lung adenocarcinoma, Nature Commun. 10 (1) (2019) 5438.

[43] N. Pržulj, D.A. Wigle, I. Jurisica, Functional topology in a network of protein interactions, Bioinformatics 20 (2004) 340–348.

[44] A. Klimovskaia, D. Lopez-Paz, L. Bottou, M. Nickel, Poincaré maps for analyzing complex hierarchies in single-cell data, Nature Commun. 11 (1) (2020) 1–9.

[45] K. Fortney, M. Kotlyar, I. Jurisica, Inferring the functions of longevity genes with modular subnetwork biomarkers of caenorhabditis elegans aging, Genome Biol. 11 (2) (2010) R13.

[46] J. Wiens, S. Saria, M. Sendak, M. Ghassemi, V.X. Liu, F. Doshi-Velez, K. Jung, K. Heller, D. Kale, M. Saeed, P.N. Ossorio, S. Thadaney-Israni, A. Goldenberg, Do no harm: A roadmap for responsible machine learning for health care, Nat. Med. 25 (9) (2019) 1337–1340, http://dx.doi.org/10.1038/s41591-019-0548-6.

[47] R. Jörnsten, T. Abenius, T. Kling, L. Schmidt, E. Johansson, T.E.M. Nordling, B. Nordlander, C. Sander, P. Gennemark, K. Funa, B. Nilsson, L. Lindahl, S. Nelander, Network modeling of the transcriptional effects of copy number aberrations in glioblastoma, Mol. Syst. Biol. 7 (2011) 486, http://dx.doi.org/10.1038/msb.2011.17, URL http://www.ncbi.nlm.nih.gov/pubmed/21525872, http://www.pubmedcentral.nih.gov/articlerender.fcgi?.

[48] Y. Yang, L. Han, Y. Yuan, J. Li, N. Hei, H. Liang, C. Sotiriou, L. Bullinger, D. Spentzos, H. Zhao, A.S. Adler, H.Y. Chang, M.C. Abba, E. Lacunza, M. Butti, C.M. Aldaz, M. Buyse, A. Chakravarti, M.J. van de Vijver, C. Langer, D.G. Oscier, L.I. Furlong, A.L. Barabasi, Z.N. Oltvai, J.J. Cai, E. Borenstein, D.A. Petrov, W. Zhao, P. Langfelder, S. Horvath, B. Zhang, S. Horvath, K.I. Goh, H. Jeong, S.P. Mason, A.L. Barabasi, Z.N. Oltvai, H. Liang, W.H. Li, H. Yu, P.M. Kim, E. Sprecher, V. Trifonov, M. Gerstein, H. Yu, D. Greenbaum, H.X. Lu, X. Zhu, M. Gerstein, J. Sun, Z. Zhao, J.D. Han, J.M. Stuart, E. Segal, D. Koller, S.K. Kim, D.P. Bartel, K. Chen, N. Rajewsky, B.P. Lewis, C.B. Burge, D.P. Bartel, M. Ashburner, H. Yu, M. Gerstein, J. Lu, G.A. Calin, C.M. Croce, S.M. Hadad, J.T. Hwang, P. Langfelder, B. Zhang, S. Horvath, A. Grimson, J.P. Brunet, P. Tamayo, T.R. Golub, J.P. Mesirov, A. Alexa, J. Rahnenfuhrer, T. Lengauer, Gene co-expression network analysis reveals common system-level properties of prognostic genes across cancer types, Nature Commun. 5 (2014) 262–272, http://dx.doi.org/10.1038/ncomms4231.

[49] A.L. Barabasi, N. Gulbahce, J. Loscalzo, Network medicine: A network-based approach to human disease, Nat. Rev. Genet. 12 (1) (2011) 56–68.

[50] S.W.H. Wong, C. Pastrello, M. Kotlyar, C. Faloutsos, I. Jurisica, Modeling tumor progression via the comparison of stage-specific graphs, Methods 132 (2018) 34–41.

[51] A. Monette, A. Morou, N.A. Al-Banna, L. Rousseau, J.-B. Lattouf, S. Rahmati, T. Tokar, J.-P. Routy, J.-F. Cailhier, D.E. Kaufmann, I. Jurisica, R. Lapointe, Failed immune responses across multiple pathologies share pan-tumor and circulating lymphocytic targets, J. Clin. Invest. 129 (6) (2019) 2463–2479, http://dx.doi.org/10.1172/JCI125301, URL https://www.jci.org/articles/view/125301.

[52] D.M. Camacho, K.M. Collins, R.K. Powers, J.C. Costello, J.J. Collins, Next-generation machine learning for biological networks, Cell 173 (7) (2018) 1581–1592.

[53] A. Holzinger, Trends in interactive knowledge discovery for personalized medicine: Cognitive science meets machine learning, IEEE Intell. Inform. Bull. 15 (1) (2014) 6–14.

[54] M.E.J. Newman, A.L. Barabási, D.J. Watts, The Structure and Dynamics of Networks, Princeton Studies in Complexity, Princeton University Press, 2006.

[55] F. Emmert-Streib, M. Dehmer, Analysis of Microarray Data: A Network-Based Approach, Wiley-VCH, weinheim, Germany, 2008.

[56] E. Ben-Naim, H. Frauenfelder, Z. Toroczkai, Complex networks, in: Lecture Notes in Physics, Springer, 2004.

[57] M. Dehmer, F. Emmert-Streib, Y. Shi, Quantitative graph theory: A new branch of graph theory and network science, Inform. Sci. 418 (2017) 575–580, http://dx.doi.org/10.1016/j.ins.2017.08.009.

[58] M. Dehmer, F. Emmert-Streib, Quantitative Graph Theory. Theory and Applications, CRC Press, 2014.

[59] B. Wang, A.M. Mezlini, F. Demir, M. Fiume, Z. Tu, M. Brudno, B. Haibe-Kains, A. Goldenberg, Similarity network fusion for aggregating data types on a genomic scale, Nature Methods 11 (3) (2014) 333–340, http://dx.doi.org/10.1038/nMeth.2810.

[60] H. Bunke, Recent developments in graph matching, in: 15-th International Conference on Pattern Recognition, vol. 2, 2000, pp. 117–124.

[61] F. Emmert-Streib, M. Dehmer, Y. Shi, Fifty years of graph matching, network alignment and network comparison, Inform. Sci. (2016) 346–347, 180–197.

[62] M. Dehmer, Z. Chen, Y. Shi, Y. Zhang, S. Tripathi, M. Ghorbani, A. Mowshowitz, F. Emmert-Streib, On efficient network similarity measures, Appl. Math. Comput. 362 (2019) 124521.

[63] J. Devillers, A.T. Balaban, Topological Indices and Related Descriptors in QSAR and QSPR, Gordon and Breach Science Publishers, Amsterdam, The Netherlands, 1999.

[64] D. Bonchev, Information Theoretic Indices for Characterization of Chemical Structures, Research Studies Press, Chichester, 1983.

[65] R.E. Ulanowicz, Circumscribed complexity in ecological networks, in: J. Devillers, A.T. Balaban (Eds.), Advances in Network Complexity, Wiley-Blackwell, 2013, pp. 249–258, m. Dehmer and A. Mowshowitz and F. Emmert-Streib..

[66] M.M. Bronstein, J. Bruna, Y. LeCun, A. Szlam, P. Vandergheynst, Geometric deep learning: Going beyond euclidean data, IEEE Signal Process. Mag. 34 (4) (2017) 18–42, http://dx.doi.org/10.1109/MSP.2017.2693418.

[67] I. Laponogov, G. Gonzalez, M. Shepherd, A. Qureshi, D. Veselkov, G. Charkoftaki, V. Vasiliou, J. Youssef, R. Mirnezami, M. Bronstein, Network machine learning maps phytochemically rich hyperfoods to fight covid-19, Hum. Genom. 15 (1) (2021) 1–11, http://dx.doi.org/10.1186/s40246-020-00297-x.

[68] N. Díaz-Rodríguez, A. Lamas, J. Sanchez, G. Franchi, I. Donadello, S. Tabik, D. Filliat, P. Cruz, R. Montes, F. Herrera, Explainable neural-symbolic learning (x-nesyl) methodology to fuse deep learning representations with expert knowledge graphs: The monumai cultural heritage use case (2021). arXiv:2104.11914.

[69] Z.C. Lipton, The mythos of model interpretability, Queue 16 (3) (2018) 31–57, http://dx.doi.org/10.1145/3236386.3241340.

[70] J. DeYoung, S. Jain, N.F. Rajani, E. Lehman, C. Xiong, R. Socher, B.C. Wallace, ERASER: A benchmark to evaluate rationalized NLP models, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, in: Association for Computational Linguistics, Online, 2020, pp. 4443–4458, http://dx.doi.org/10.18653/v1/2020.acl-main.408, URL https://www.aclweb.org/anthology/2020.acl-main.408.

[71] P. Atanasova, J.G. Simonsen, C. Lioma, I. Augenstein, A diagnostic study of explainability techniques for text classification, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP, in: Association for Computational Linguistics, Online, 2020, pp. 3256–3274, http://dx.doi.org/10.18653/v1/2020.emnlp-main.263, URL https://www.aclweb.org/anthology/2020.emnlp-main.263.

[72] N. Díaz-Rodríguez, A. Härmä, R. Helaoui, I. Huitzil, F. Bobillo, U. Straccia, Couch potato or gym addict? Semantic lifestyle profiling with wearables and fuzzy knowledge graphs, in: Automatic Knowledge Base Construction (AKBC) Workshop at NIPS, 2017, pp. 1–10.

[73] N.D. Rodríguez, M.P. Cuéllar, J. Lilius, M.D. Calvo-Flores, A fuzzy ontology for semantic modelling and recognition of human behaviour, Knowl.-Based Syst. 66 (2014) 46–60.

[74] C. Biemann, M. Riedl, Text: Now in 2D! a framework for lexical expansion with contextual similarity, J. Lang. Model. 1 (1) (2013) 55–95.

[75] N. Díaz-Rodríguez, S. Gronroos, F. Wickstrom, J. Lilius, H. Eertink, A. Braun, P. Dillen, J. Crowley, J. Alexandersson, An ontology for wearables data interoperability and ambient assisted living application development, in: Recent Developments and the New Direction in Soft-Computing Foundations and Applications, Springer, 2017, pp. 1–5, http://dx.doi.org/10.1007/978-3-319-75408-6_43.

[76] H. Jeong, S.P. Mason, A.L. Barabasi, Z.N. Oltvai, Lethality and centrality in protein networks, Nature 411 (6833) (2001) 41–42.

[77] M.W. Hahn, A.D. Kern, Comparative genomics of centrality and essentiality in three eukaryotic protein-interaction networks, Mol. Biol. Evol. 22 (4) (2005) 803–806.

[78] S. K. Maslov S., Nspecificity and stability in topology of protein networks, Science 296 (2002) 910–913.

[79] A.C. Gavin, M. Bosche, R. Krause, P. Grandi, M. Marzioch, A. Bauer, J. Schultz, J.M. Rick, A.M. Michon, C.M. Cruciat, M. Remor, C. Hofert, M. Schelder, M. Brajenovic, H. Ruffner, A. Merino, K. Klein, M. Hudak, D. Dickson, T. Rudi, V. Gnau, A. Bauch, S. Bastuck, B. Huhse, C. Leutwein, M.A. Heurtier, R.R. Copley, A. Edelmann, E. Querfurth, V. Rybin, G. Drewes, M. Raida, T. Bouwmeester, P. Bork, B. Seraphin, B. Kuster, G. Neubauer, G. Superti-Furga, Functional organization of the yeast proteome by systematic analysis of protein complexes, Nature 415 (6868) (2002) 141–147.

[80] S. Wuchty, Topology and weights in a protein domain interaction network–a novel way to predict protein interactions, BMC Genom. 7 (2006) 1471–2164, (Electronic) Journal Article.

[81] R. Sharan, I. Ulitsky, R. Shamir, Network-based prediction of protein function, Mol. Syst. Biol. 3 (2007) 88.

[82] E. Ravasz, A.L. Somera, D.A. Mongru, Z.N. Oltvai, A.L. Barabasi, Hierarchical organization of modularity in metabolic networks, Science 297 (5586) (2002) 1551–1556, 22192386 1095-9203 Journal Article.

[83] H. Yu, M. Gerstein, Genomic analysis of the hierarchical structure of regulatory networks, Proc. Natl. Acad. Sci. U S A0027-8424 (Print) Journal article.

[84] J.D. Han, N. Bertin, T. Hao, D.S. Goldberg, G.F. Berriz, L.V. Zhang, D. Dupuy, A.J. Walhout, M.E. Cusick, F.P. Roth, M. Vidal, Evidence for dynamically organized modularity in the yeast protein-protein interaction network, Nature 430 (6995) (2004) 88–93.

[85] E. Segal, M. Shapira, A. Regev, D. Pe'er, D. Botstein, D. Koller, N. Friedman, Module networks: Identifying regulatory modules and their condition-specific regulators from gene expression data, Nat. Genet. 34 (2) (2003) 166–176.

[86] A. Gavin, P. Aloy, P. Grandi, R. Krause, M. Boesche, M. Marzioch, C. Rau, L. Jensen, S. Bastuck, B. Dumpelfeld, A. Edelmann, M. Heurtier, V. Hoffman, C. Hoefert, K. Klein, M. Hudak, A. Michon, M. Schelder, M. Schirle, M. Remor, T. Rudi, S. Hooper, A. Bauer, T. Bouwmeester, G. Casari, G. Drewes, G. Neubauer, J. Rick, B. Kuster, P. Bork, R. Russell, G. Superti-Furga, Proteome survey reveals modularity of the yeast cell machinery, Nature 440 (2006) 631–636.

[87] S.W. Wong, N. Cercone, I. Jurisica, Comparative network analysis via differential graphlet communities, Proteomics 15 (2–3) (2015) 608–617.

[88] S. Shen-Orr, R. Milo, S. Mangan, U. Alon, Network motifs in the transcriptional regulation network of escherichia coli, Nat. Gen. 31 (2002) 64–68.

[89] J.J. Rice, A. Kershenbaum, G. Stolovitzky, Lasting impressions: Motifs in protein-protein maps may provide footprints of evolutionary events, Proc. Natl. Acad. Sci. U S A 102 (9) (2005) 3173–3174, rice;
J. Jeremy Kershenbaum, Aaron Stolovitzky, Gustavo comment united states proceedings of the national academy of sciences of the united states of america, Proc. Natl. Acad. Sci. U S A 102 (9) (2005) 3173–3174, Epub 2005 Feb 22.

[90] M. Singh, C. Venugopal, T. Tokar, N. McFarlane, M.K. Subapanditha, M. Qazi, D. Bakhshinyan, P. Vora, N.K. Murty, I. Jurisica, S.K. Singh, Therapeutic targeting of the premetastatic stage in human lung-to-brain metastasis, Cancer Res. 78 (17) (2018) 5124–5134.

[91] A. Monette, D. Bergeron, A. Ben Amor, L. Meunier, C. Caron, A.M. Mes-Masson, N. Kchir, K. Hamzaoui, I. Jurisica, R. Lapointe, Immune-enrichment of non-small cell lung cancer baseline biopsies for multiplex profiling define prognostic immune checkpoint combinations for patient stratification, J. Immunother. Cancer 7 (1) (2019) 86.

[92] M. Kotlyar, K. Fortney, I. Jurisica, Network-based characterization of drug-regulated genes, drug targets, and toxicity, Methods 57 (4) (2012) 499–507.

[93] M. Xiong, Big Data in Omics and Imaging: Integrated Analysis and Causal Inference, CRC Press, 2018.

[94] M. Barrios-Rodiles, K.R. Brown, B. Ozdamar, R. Bose, Z. Liu, R.S. Donovan, F. Shinjo, Y. Liu, J. Dembowy, I.W. Taylor, V. Luga, N. Przulj, M. Robinson, H. Suzuki, Y. Hayashizaki, I. Jurisica, J.L. Wrana, High-throughput mapping of a dynamic signaling network in mammalian cells, Science 307 (5715) (2005) 1621–1625.

[95] W. Hu, Z. Feng, L. Ma, J. Wagner, J.J. Rice, G. Stolovitzky, A.J. Levine, A single nucleotide polymorphism in the mdm2 gene disrupts the oscillation of p53 and mdm2 levels in cells, Cancer Res. 67 (6) (2007) 2757–2765.

[96] R. Barshir, O. Basha, A. Eluk, I.Y. Smoly, A. Lan, E. Yeger-Lotem, The tissuenet database of human tissue protein-protein interactions, Nucl. Acids Res. 41 (Database issue) (2013) D841–4, http://dx.doi.org/10.1093/nar/gks1198, URL http://www.ncbi.nlm.nih.gov/pubmed/23193266.

[97] R. Jansen, D. Greenbaum, M. Gerstein, Relating whole-genome expression data with protein-protein interactions, Genome Res. 12 (1) (2002) 37–46.

[98] K.R. Brown, Igor Jurisica, Unequal evolutionary conservation of human protein interactions in interologous networks, Genome Biology 8 (5) (2007) R95, http://dx.doi.org/10.1186/gb-2007-8-5-r95.

[99] D.R. Rhodes, S.A. Tomlins, S. Varambally, V. Mahavisno, T. Barrette, S. Kalyana-Sundaram, D. Ghosh, A. Pandey, A.M. Chinnaiyan, Probabilistic model of the human protein-protein interaction network, Nat. Biotechnol. 23 (8) (2005) 951–959.

[100] T. Kato, Y. Murata, K. Miura, K. Asai, P.B. Horton, T. Koji, W. Fujibuchi, Network-based de-noising improves prediction from microarray data, BMC Bioinformatics (7) (2006) S4.

[101] P.F. Jonsson, P.A. Bates, Global topological features of cancer proteins in the human interactome, Bioinformatics 22 (18) (2006) 2291–2297.

[102] S. Wachi, K. Yoneda, R. Wu, Interactome-transcriptome analysis reveals the high centrality of genes differentially expressed in lung cancer tissues, Bioinformatics 21 (23) (2005) 4205–4208.

[103] P.P. Reis, T. Tokar, R.S. Goswami, Y. Xuan, M. Sukhai, A.L. Seneda, E. Móz, Luis, B. Perez-Ordonez, C. Simpson, D. Goldstein, R. Gilbert, P. Gullane, J. Irish, I. Jurisica, S. Kamel-Reid, A 4-gene signature from histologically normal surgical margins predicts local recurrence in patients with oral carcinoma: Clinical validation, Sci. Rep. 10 (1) (2020) 1–8, http://dx.doi.org/10.1038/s41598-020-58688-y.

[104] T. Tokar, C. Pastrello, V.R. Ramnarine, C.-Q. Zhu, K.J. Craddock, L.A. Pikor, E.A. Vucic, S. Vary, F.A. Shepherd, M.-S. Tsao, W.L. Lam, I. Jurisica, Differentially expressed micrornas in lung adenocarcinoma invert effects of copy number aberrations of prognostic genes, Oncotarget 9 (10) (2018) 9137–9155, http://dx.doi.org/10.18632/oncotarget.24070, URL https://www.oncotarget.com/article/24070/.

[105] V. Mandilaras, S. Garg, M. Cabanero, Q. Tan, C. Pastrello, J. Burnier, K. Karakasis, L. Wang, N.C. Dhani, M.O. Butler, P.L. Bedard, L.L. Siu, B. Clarke, P.A. Shaw, T. Stockley, I. Jurisica, A.M. Oza, S. Lheureux, Tp53 mutations in high grade serous ovarian cancer and impact on clinical outcomes: A comparison of next generation sequencing and bioinformatics analyses, International journal of gynecological cancer : Official journal of the International Gynecological Cancer Society.

[106] A.D. King, N. Przulj, I. Jurisica, Protein complex prediction via cost-based clustering, Bioinformatics 20 (17) (2004) 3013–3020.

[107] J. I. Przulj N., Corneil D.G., Efficient estimation of graphlet frequency distributions in protein-protein interaction networks, Bioinformatics 22 (2006) 974–980.

[108] S.W.H. Wong, C. Pastrello, M. Kotlyar, C. Faloutsos, I. Jurisica, Sdregion: Fast spotting of changing communities in biological networks, Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.

[109] D. Berrar, W. Dubitzky, Deep learning in bioinformatics and biomedicine, Brief. Bioinform. 22 (2) (2021) 1513–1514.

[110] T. Ideker, O. Ozier, B. Schwikowski, A.F. Siegel, Discovering regulatory and signalling circuits in molecular interaction networks, Bioinformatics 18 (18 Suppl) (2002) S233–40.

[111] H.Y. Chuang, E. Lee, Y.T. Liu, D. Lee, T. Ideker, Network-based classification of breast cancer metastasis, Mol. Syst. Biol. 3 (2007) 140.

[112] S. Nacu, R. Critchley-Thorne, P. Lee, S. Holmes, Gene expression network analysis and applications to immunology, Bioinformatics 23 (7) (2007) 850–858.

[113] Y.C. Hwang, C.C. Lin, J.Y. Chang, H. Mori, H.F. Juan, H.C. Huang, Predicting essential genes based on network and sequence analysis, Mol. Biosyst. 5 (12) (2009) 1672–1678.

[114] J. I. Geraci J., Liu G., Algorithms for systematic identification of small subgraphs (2012).

[115] S.-J. Schramm, S.S. Li, V. Jayaswal, D.C.Y. Fung, A.E. Campain, C.N.I. Pang, R.A. Scolyer, Y.H. Yang, G.J. Mann, M.R. Wilkins, Disturbed protein-protein interaction networks in metastatic melanoma are associated with worse prognosis and increased functional mutation burden, Pigment Cell Melanoma Res. 26 (5) (2013) 708–722, http://dx.doi.org/10.1111/pcmr.12126, URL http://www.ncbi.nlm.nih.gov/pubmed/23738911.

[116] N. Ramanan, S. Natarajan, Causal learning from predictive modeling for observational data, Front. Big Data 3 (2020) 34.
[117] R. Guo, L. Cheng, J. Li, P.R. Hahn, H. Liu, A survey of learning causality with data: Problems and methods, ACM Comput. Surv. 53 (4) (2020) 1–37.
[118] B. Schölkopf, Causality for machine learning, arXiv:1911.10500.
[119] H. Lu, A.L. Yuille, M. Liljeholm, P.W. Cheng, K.J. Holyoak, Bayesian Generic priors for causal learning, Psychol. Rev. 115 (4) (2008) 955–984, http://dx.doi.org/10.1037/a0013256.
[120] J. Peters, D. Janzing, B. Schölkopf, Elements of Causal Inference: Foundations and Learning Algorithms, Cambridge University Press, Cambridge (MA), 2017.
[121] D. Lopez-Paz, R. Nishihara, S. Chintala, B. Schölkopf, L. Bottou, Discovering causal signals in images, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 6979–6987.
[122] D.C. Castro, I. Walker, B. Glocker, Causality matters in medical imaging, Nature Commun. 11 (1) (2020) 1–10.
[123] O. Goudet, D. Kalainathan, P. Caillou, I. Guyon, D. Lopez-Paz, M. Sebag, Causal generative neural networks, arXiv preprint arXiv:1711.08936.
[124] M. Rojas-Carulla, M. Baroni, D. Lopez-Paz, Causal discovery using proxy variables, arXiv:1702.07306.
[125] B. Schölkopf, Causality for machine learning, arXiv preprint arXiv:1911.10500.
[126] R. Moraffah, M. Karami, R. Guo, A. Raglin, H. Liu, Causal interpretability for machine learning-problems, methods and evaluation, ACM SIGKDD Explor. Newsl. 22 (1) (2020) 18–33.
[127] J. Pearl, Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference, Morgan Kaufmann, San Francisco, 1988.
[128] J. Pearl, Causality: Models, Reasoning, and Inference, 2nd ed., Cambridge University Press, Cambridge, 2009.
[129] A.-H. Karimi, J. von Kügelgen, B. Schölkopf, I. Valera, Algorithmic recourse under imperfect causal knowledge: A probabilistic approach, arXiv:2006.06831.
[130] S. Wachter, B. Mittelstadt, C. Russell, Counterfactual explanations without opening the black box: Automated decisions and the GDPR, Harv. J. Law Technol. 31 (2018) 841–887.
[131] M. Prosperi, Y. Guo, M. Sperrin, J.S. Koopman, J.S. Min, X. He, S. Rich, M. Wang, I.E. Buchan, J. Bian, Causal inference and counterfactual prediction in machine learning for actionable healthcare, Nat. Mach. Intell. 2 (7) (2020) 369–375.
[132] M. Sharma, S. Mindermann, J. Brauner, G. Leech, A. Stephenson, T. Gavenčiak, J. Kulveit, Y.W. Teh, L. Chindelevitch, Y. Gal, How robust are the estimated effects of nonpharmaceutical interventions against covid-19? Advances in Neural Information Processing Systems 33.
[133] A. Jesson, S. Mindermann, U. Shalit, Y. Gal, Identifying causal-effect inference failure with uncertainty-aware models, Advances in Neural Information Processing Systems 33.
[134] T. Nair, D. Precup, D.L. Arnold, T. Arbel, Exploring uncertainty measures in deep networks for multiple sclerosis lesion detection and segmentation, Med. Image Anal. 59 (2020) 101557, http://dx.doi.org/10.1016/j.media.2019.101557.
[135] T. Miller, Contrastive explanation: A structural-model approach, arXiv preprint arXiv:1811.03163.
[136] N. Elzein, The demand for contrastive explanations, Philos. Stud. 176 (5) (2019) 1325–1339.
[137] B. Krarup, S. Krivic, F. Lindner, D. Long, Towards contrastive explanations for comparing the ethics of plans, arXiv preprint arXiv:2006.12632.
[138] A. Dhurandhar, P.-Y. Chen, R. Luss, C.-C. Tu, P. Ting, K. Shanmugam, P. Das, Explanations based on the missing: Towards contrastive explanations with pertinent negatives, Adv. Neural Inf. Process. Syst. 31 (2018) 592–603.
[139] W. Liang, J. Zou, Z. Yu, Alice: Active learning with contrastive natural language explanations, arXiv preprint arXiv:2009.10259.
[140] Y. Liu, Z. Li, Q. Ge, N. Lin, M. Xiong, Deep feature selection and causal analysis of alzheimer's disease, Front. Neurosci. 13 (2019) 1198.
[141] T.R. Besold, A. d. Garcez, S. Bader, H. Bowman, P. Domingos, P. Hitzler, K.-U. Kühnberger, L.C. Lamb, D. Lowd, P.M.V. Lima, et al., Neural-symbolic learning and reasoning: A survey and interpretation, arXiv preprint arXiv:1711.03902.
[142] J.-R. King, F. Charton, D. Lopez-Paz, M. Oquab, Back-to-back regression: Disentangling the influence of correlated factors from multivariate observations, NeuroImage 220 (2020) 117028.
[143] A. Barredo-Arrieta, J. Del Ser, Plausible counterfactuals: Auditing deep learning classifiers with realistic adversarial examples, in: 2020 International Joint Conference on Neural Networks, IJCNN, IEEE, 2020, pp. 1–7.
[144] S.L. Hyland, M. Faltys, M. Hueser, X. Lyu, T. Gumbsch, C. Esteban, C. Bock, M. Horn, M. Moor, B. Rieck, M. Zimmermann, D. Bodenham, K.M. Borgwardt, G. Raetsch, T.M. Merz, Early prediction of circulatory failure in the intensive care unit using machine learning, Nat. Med. 26 (3) (2020) 364–373, http://dx.doi.org/10.1038/s41591-020-0789-4.
[145] A. Holzinger, C. Biemann, C.S. Pattichis, D.B. Kell, What do we need to build explainable ai systems for the medical domain? arXiv:1712.09923.
[146] R. Caruana, Y. Lou, J. Gehrke, P. Koch, M. Sturm, N. Elhadad, Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission, in: 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '15, ACM, 2015, pp. 1721–1730, http://dx.doi.org/10.1145/2783258.2788613.
[147] A. Holzinger, Interactive machine learning for health informatics: When do we need the human-in-the-loop? Brain Inform. 3 (2) (2016) 119–131, http://dx.doi.org/10.1007/s40708-016-0042-6.
[148] F. Buccafurri, T. Eiter, G. Gottlob, N. Leone, Enhancing model checking in verification by ai techniques, Artificial Intelligence 112 (1–2) (1999) 57–104, http://dx.doi.org/10.1016/S0004-3702(99)00039-9.
[149] A. Holzinger, G. Langs, H. Denk, K. Zatloukal, H. Müller, Causability and explainability of artificial intelligence in medicine, Wiley Interdiscip. Rev. Data Min. Knowl. Discov. 9 (4) (2019) 1–13, http://dx.doi.org/10.1002/widm.1312.
[150] A. Holzinger, Usability engineering methods for software developers, Commun. ACM 48 (1) (2005) 71–74, http://dx.doi.org/10.1145/1039539.1039541.
[151] A. Holzinger, A. Carrington, H. Müller, Measuring the quality of explanations: The system causability scale (scs). comparing human and machine explanations, KI - Künstliche intelligenz (german journal of artificial intelligence), in: Kristian Kersting (Ed.), Special Issue on Interactive Machine Learning, Vol. 34 (2), TU Darmstadt, 2020, pp. 193–198, http://dx.doi.org/10.1007/s13218-020-00636-z.
[152] S. Grigorescu, B. Trasnea, T. Cocias, G. Macesanu, A survey of deep learning techniques for autonomous driving, J. Field Robotics 37 (3) (2020) 362–386.
[153] K. Muhammad, A. Ullah, J. Lloret, J. Del Ser, V.H.C. de Albuquerque, Deep learning for safe autonomous driving: Current challenges and future directions, IEEE Transactions on Intelligent Transportation Systems.
[154] D.S. Ting, Y. Liu, P. Burlina, X. Xu, N.M. Bressler, T.Y. Wong, Ai for medical imaging goes deep, Nat. Med. 24 (5) (2018) 539–540.
[155] K. Muhammad, S. Khan, J. Del Ser, V.H.C. de Albuquerque, Deep learning for multigrade brain tumor classification in smart healthcare systems: A prospective survey, IEEE Transactions on Neural Networks and Learning Systems.
[156] A. Diez-Olivan, J. Del Ser, D. Galar, B. Sierra, Data fusion and machine learning for industrial prognosis: Trends and perspectives towards industry 4.0, Inf. Fusion 50 (2019) 92–111.
[157] K.P. Tellakula, R. Saravana Kumar, S. Deb, A survey of ai imaging techniques for covid-19 diagnosis and prognosis, Appl. Comput. Sci. 17 (2) (2021) 40–55.
[158] Y. Xin, L. Kong, Z. Liu, Y. Chen, Y. Li, H. Zhu, M. Gao, H. Hou, C. Wang, Machine learning and deep learning methods for cybersecurity, IEEE Access 6 (2018) 35365–35381.
[159] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Vol. 1 (Long and Short Papers), in: Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186, http://dx.doi.org/10.18653/v1/N19-1423, URL https://www.aclweb.org/anthology/N19-1423.
[160] N. Rethmeier, V.K. Saxena, I. Augenstein, in: R.P. Adams, V. Gogate (Eds.), TX-Ray: Quantifying and Explaining Model-Knowledge Transfer in (Un-)Supervised NLP, UAI, AUAI Press, 2020, p. 197, URL http://dblp.uni-trier.de/db/conf/uai/uai2020.html#RethmeierSA20.
[161] H. Daumé III, Frustratingly easy domain adaptation, in: Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, in: Association for Computational Linguistics, Prague, Czech Republic, 2007, pp. 256–263, URL https://www.aclweb.org/anthology/P07-1033.
[162] J. Blitzer, R. McDonald, F. Pereira, Domain adaptation with structural correspondence learning, in: Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Sydney, Australia, 2006, pp. 120–128, URL https://www.aclweb.org/anthology/W06-1615.
[163] D. Wright, I. Augenstein, Transformer based multi-source domain adaptation, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP, Association for Computational Linguistics, Online, 2020, pp. 7963–7974, http://dx.doi.org/10.18653/v1/2020.emnlp-main.639, URL https://www.aclweb.org/anthology/2020.emnlp-main.639.
[164] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial networks, Commun. ACM 63 (11) (2020) 139–144.
[165] Y. Ganin, V. Lempitsky, Unsupervised Domain Adaptation by Backpropagation, in: International Conference on Machine Learning, 2015, pp. 1180–1189.
[166] P. Atanasova, D. Wright, I. Augenstein, Generating label cohesive and well-formed adversarial claims, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP, Association for Computational Linguistics, Online, 2020, pp. 3168–3177, http://dx.doi.org/10.18653/v1/2020.emnlp-main.256, URL https://www.aclweb.org/anthology/2020.emnlp-main.256.
[167] C. Lin, S. Bethard, D. Dligach, F. Sadeque, G. Savova, T.A. Miller, Does BERT need domain adaptation for clinical negation detection? J. Am. Med. Inform. Assoc. 27 (4) (2020) 584–591.
[168] T. Lesort, V. Lomonaco, A. Stoian, D. Maltoni, D. Filliat, N. Díaz-Rodríguez, Continual learning for robotics: Definition, framework, learning strategies, opportunities and challenges, Inf. Fusion 58 (2020) 52–68.
[169] J. Bjerva, W. Kouw, I. Augenstein, Back to the future – sequential alignment of text representations, in: Proceedings of the 34th AAAI Conference on Artificial Intelligence..

[170] D. Doran, S. Schulz, T.R. Besold, What does explainable ai really mean? a new conceptualization of perspectives, arXiv preprint arXiv:1710.00794.

[171] A. Bennetot, J.-L. Laurent, R. Chatila, N. Díaz-Rodríguez, Towards explainable neural-symbolic visual reasoning, in: NeSy Workshop IJCAI, 2019, pp. 1–10.

[172] V. Vapnik, A. Chervonenkis, On the uniform convergence of relative frequencies of events to their probabilities, Theory Probab. Appl. 16 (2) (1971) 264–280, http://dx.doi.org/10.1137/1116025.

[173] V. Vapnik, A. Chervonenkis, The necessary and sufficient conditions for consistency in the empirical risk minimization method, Pattern Recognit. Image Anal. 1 (3) (1991) 283–305.

[174] T. Poggio, R. Rifkin, S. Mukherjee, P. Niyogi, General conditions for predictivity in learning theory, Nature 428 (6981) (2004) 419–422, http://dx.doi.org/10.1038/nature02341.

[175] A. Majkowska, S. Mittal, D.F. Steiner, J.J. Reicher, S.M. McKinney, G.E. Duggan, K. Eswaran, P.-H. Cameron Chen, Y. Liu, S.R. Kalidindi, et al., Chest radiograph interpretation with deep learning models: Assessment with radiologist-adjudicated reference standards and population-adjusted evaluation, Radiology 294 (2) (2020) 421–431.

[176] S. McKinney, M. Sieniek, V. Godbole, N. Antropova, H. Ashrafian, T. Back, M. Chesus, G. Corrado, A. Darzi, M. Etemadi, F. Garcia-Vicente, F. Gilbert, M. Halling-Brown, H. D, S. Jansen, A. Karthikesalingam, C. Kelly, D. King, J. Ledsam, D. Melnick, H. Mostofi, L. Peng, J. Reicher, s. B. Romera-Parede, R. Sidebottom, M. Suleyman, D. Tse, K. Young, J. De Fauw, S. Shetty, International evaluation of an ai system for breast cancer screening, Nature 577 (7788) (2020) 89–94.

[177] H. Xu, S. Mannor, Robustness and generalization, Mach. Learn. 86 (3) (2012) 391–423, http://dx.doi.org/10.1007/s10994-011-5268-1.

[178] N. Díaz-Rodríguez, R. Binkyte-Sadauskiene, W. Bakkali, S. Bookseller, P. Tubaro, A. Bacevicius, R. Chatila, Questioning causality on sex, gender and COVID-19, and identifying bias in large-scale data-driven analyses: the Bias Priority Recommendations and Bias Catalog for Pandemics, arXiv preprint arXiv:2104.14492.

[179] J. Pearl, E. Bareinboim, Transportability of causal and statistical relations: A formal approach, in: 11th International IEEE Conference on Data Mining Workshops, IEEE, 2011, pp. 540–547, http://dx.doi.org/10.1109/ICDMW.2011.169.

[180] D. Tsipras, S. Santurkar, L. Engstrom, A. Turner, A. Madry, Robustness may be at odds with accuracy, arXiv:1805.12152.

[181] Y. Dong, F. Liao, T. Pang, H. Su, J. Zhu, X. Hu, J. Li, Boosting adversarial attacks with momentum, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, 2018, pp. 9185–9193, http://dx.doi.org/10.1109/CVPR.2018.00957.

[182] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, R. Fergus, Intriguing properties of neural networks, arXiv preprint arXiv:1312.6199.

[183] A. Margeloiu, N. Simidjievski, M. Jamnik, A. Weller, Improving interpretability in medical imaging diagnosis using adversarial training, arXiv preprint arXiv:2012.01166.

[184] D. Merkulov, I.V. Oseledets, Empirical study of extreme overfitting points of neural networks, J. Commun. Technol. Electron. 64 (12) (2019) 1527–1534.

[185] M. Everingham, L. Van Gool, C.K. Williams, J. Winn, A. Zisserman, The pascal visual object classes (voc) challenge, Int. J. Comput. Vis. 88 (2) (2010) 303–338.

[186] Y. Belinkov, A. Poliak, S. Shieber, B. Van Durme, A. Rush, On adversarial removal of hypothesis-only bias in natural language inference, in: Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics, *SEM 2019, Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 256–262, http://dx.doi.org/10.18653/v1/S19-1028, URL https://www.aclweb.org/anthology/S19-1028.

[187] S. Lapuschkin, S. Wäldchen, A. Binder, G. Montavon, W. Samek, K.-R. Müller, Unmasking clever hans predictors and assessing what machines really learn, Nature Commun. 10 (2019) 1096.

[188] A. Jain, M. Ravula, J. Ghosh, Biased models have biased explanations, arXiv preprint arXiv:2012.10986.

[189] N. Liu, M. Du, X. Hu, Adversarial machine learning: An interpretation perspective, arXiv preprint arXiv:2004.11488.

[190] A. Heuillet, F. Couthouis, N. Díaz-Rodríguez, Explainability in deep reinforcement learning, Knowl.-Based Syst. 214 (2021) 106685.

[191] T.W. Killian, H. Zhang, J. Subramanian, M. Fatemi, M. Ghassemi, An empirical study of representation learning for reinforcement learning in healthcare, arXiv preprint arXiv:2011.11235.

[192] M.T. Ribeiro, S. Singh, C. Guestrin, why should I trust you?: Explaining the predictions of any classifier, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 1135–1144.

[193] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, W. Samek, On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation, PLoS One 10 (7) (2015) e0130140.

[194] S. Gehrmann, H. Strobelt, R. Krueger, H. Pfister, A.M. Rush, Visual interaction with deep learning models through collaborative semantic inference, IEEE Trans. Vis. Comput. Graphics 26 (1) (2019) 884–894, http://dx.doi.org/10.1109/TVCG.2019.2934595.

[195] U. Schmid, B. Finzel, Mutual Explanations for Cooperative Decision Making in Medicine, KI-Künstliche Intelligenz, 2020, pp. 1–7.

[196] S. Bruckert, B. Finzel, U. Schmid, The next generation of medical decision support: A roadmap toward transparent expert companions, Front. Artif. Intell. 3 (2020) 75.

[197] S. Wallkotter, S. Tulli, G. Castellano, A. Paiva, M. Chetouani, Explainable agents through social cues: A review, arXiv preprint arXiv:2003.05251.

[198] E. Tjoa, C. Guan, A survey on explainable artificial intelligence (XAI): towards medical XAI, CoRR abs/1907.07374. arXiv:1907.07374.

[199] T. Lei, R. Barzilay, T. Jaakkola, Rationalizing neural predictions, in: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Austin, Texas, 2016, pp. 107–117, http://dx.doi.org/10.18653/v1/D16-1011, URL https://www.aclweb.org/anthology/D16-1011.

[200] P. Atanasova, J.G. Simonsen, C. Lioma, I. Augenstein, Generating fact checking explanations, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 7352–7364, http://dx.doi.org/10.18653/v1/2020.acl-main.656, URL https://www.aclweb.org/anthology/2020.acl-main.656.

[201] A. Singh, S. Sengupta, A.R. Mohammed, I. Faruq, V. Jayakumar, J. Zelek, V. Lakshminarayanan, et al., What is the optimal attribution method for explainable ophthalmic disease classification? in: International Workshop on Ophthalmic Medical Image Analysis, Springer, 2020, pp. 21–31.

[202] B. Kim, M. Wattenberg, J. Gilmer, C. Cai, J. Wexler, F. Viegas, R. Sayres, Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav), 2018, arXiv:1711.11279.

[203] D. Stammbach, E. Ash, e-fever: Explanations and summaries for automated fact checking, in: Proceedings of the 2020 Truth and Trust Online Conference, TTO 2020, Hacks Hackers, 2020, p. 32.

[204] S. Enarvi, M. Amoia, M. Del-Agua Teba, B. Delaney, F. Diehl, S. Hahn, K. Harris, L. McGrath, Y. Pan, J. Pinto, L. Rubini, M. Ruiz, G. Singh, F. Stemmer, W. Sun, P. Vozila, T. Lin, R. Ramamurthy, Generating medical reports from patient-doctor conversations using sequence-to-sequence models, in: Proceedings of the First Workshop on Natural Language Processing for Medical Conversations, Association for Computational Linguistics, Online, 2020, pp. 22–30, http://dx.doi.org/10.18653/v1/2020.nlpmc-1.4, URL https://www.aclweb.org/anthology/2020.nlpmc-1.4.

[205] J. Maynez, S. Narayan, B. Bohnet, R. McDonald, On faithfulness and factuality in abstractive summarization, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 1906–1919, http://dx.doi.org/10.18653/v1/2020.acl-main.173, URL https://www.aclweb.org/anthology/2020.acl-main.173.

[206] O.-M. Camburu, T. Rocktäschel, T. Lukasiewicz, P. Blunsom, e-SNLI: Natural language inference with natural language explanations, in: S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, R. Garnett (Eds.), Advances in Neural Information Processing Systems 31, Curran Associates, Inc., 2018, pp. 9539–9549.

[207] N. Kotonya, F. Toni, Explainable automated fact-checking: A survey, in: Proceedings of the 28th International Conference on Computational Linguistics, International Committee on Computational Linguistics, Barcelona, Spain (Online), 2020, pp. 5430–5443, URL https://www.aclweb.org/anthology/2020.coling-main.474.

[208] G. Andrew, R. Arora, J. Bilmes, K. Livescu, Deep canonical correlation analysis, in: Proceedings of the 30th International Conference on International Conference on Machine Learning, Vol. 28, ICML'13, JMLR.org, 2013, pp. III–1247–III–1255.

[209] L.G. Valiant, Projection learning, Mach. Learn. 37 (2) (1999) 115–130, http://dx.doi.org/10.1023/A:1007678005361.

[210] T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient estimation of word representations in vector space, in: Y. Bengio, Y. LeCun (Eds.), 1st International Conference on Learning Representations, ICLR 2013 Scottsdale, Arizona, USA, May (2013) 2-4, Workshop Track Proceedings, 2013, pp. 1–10, http://arxiv.org/abs/1301.3781.

[211] S. Jain, B.C. Wallace, Attention is not explanation, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Vol. 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, USA, 2019, pp. 3543–3556, http://dx.doi.org/10.18653/v1/N19-1357, URL https://www.aclweb.org/anthology/N19-1357.

[212] C. Meister, S. Lazov, D. Wright, I. Augenstein, R. Cotterell, Is sparse attention more interpretable? in: Proceedings of ACL-IJCNLP, Association for Computational Linguistics, 2021, pp. 1–10.

[213] J. Valdez, M. Kim, M. Rueschman, V. Socrates, S.S. Sahoo, Provcare semantic provenance knowledgebase: Evaluating scientific reproducibility of research studies, in: AMIA 2017, American Medical Informatics Association Annual Symposium, Washington, DC, USA, November (2017) 4-8, AMIA, 2017, pp. 1–10, URL http://knowledge.amia.org/65881-amia-1.3897810/t003-1.3901461/f003-1.3901462/2731669-1.3901539/2731984-1.3901536.

[214] H. Lakkaraju, E. Kamar, R. Caruana, J. Leskovec, Faithful and customizable explanations of black box models, in: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics and Society, 2019, pp. 131–138.

[215] M. Muzammal, R. Talat, A.H. Sodhro, S. Pirbhulal, A multi-sensor data fusion enabled ensemble approach for medical data from body sensor networks, Inf. Fusion 53 (2020) 155–164.

[216] F. Rundo, G.L. Banna, L. Prezzavento, F. Trenta, S. Conoci, S. Battiato, 3d non-local neural network: A non-invasive biomarker for immunotherapy treatment outcome prediction. case-study: Metastatic urothelial carcinoma, Journal of Imaging.

[217] F. Sattler, K.-R. Müller, W. Samek, Clustered federated learning: Model-agnostic distributed multi-task optimization under privacy constraints, IEEE Trans. Neural Netw. Learn. Syst. 31 (9) (2020) 3400–3413.

[218] L. Ruff, J.R. Kauffmann, R.A. Vandermeulen, G. Montavon, W. Samek, M. Kloft, T.G. Dietterich, K.-R. Müller, A unifying review of deep and shallow anomaly detection, Proc. IEEE 109 (5) (2021) 756–795.

[219] F. Cabitza, R. Rasoini, G.F. Gensini, Unintended consequences of machine learning in medicine, JAMA 318 (6) (2017) 517–518.

[220] A. Holzinger, B. Haibe-Kains, I. Jurisica, Why imaging data alone is not enough: Ai-based integration of imaging, omics, and clinical data, Eur. J. Nucl. Med. Mol. Imaging 46 (13) (2019) 2722–2730, http://dx.doi.org/10.1007/s00259-019-04382-9.

[221] R. Bhattacharyya, M.J. Ha, Q. Liu, R. Akbani, H. Liang, V. Baladandayuthapani, Personalized network modeling of the pan-cancer patient and cell line interactome, JCO Clin. Cancer Inform. 4 (2020) 399–411.

[222] M. Abadi, A. Chu, I. Goodfellow, H.B. McMahan, I. Mironov, K. Talwar, L. Zhang, Deep learning with differential privacy, in: Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, 2016, pp. 308–318.

[223] M. Friedrich, A. Köhn, G. Wiedemann, C. Biemann, Adversarial learning of privacy-preserving text representations for de-identification of medical records, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Florence, Italy, 2019, pp. 5829–5839, http://dx.doi.org/10.18653/v1/P19-1584, URL https://www.aclweb.org/anthology/P19-1584.

[224] J. Wang, W. Bao, L. Sun, X. Zhu, B. Cao, S.Y. Philip, Private model compression via knowledge distillation, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, 2019, pp. 1190–1197.

[225] Z.B. Celik, D. Lopez-Paz, P. McDaniel, Patient-driven privacy control through generalized distillation, in: 2017 IEEE Symposium on Privacy-Aware Computing, PAC, IEEE, 2017, pp. 1–12.

[226] M.R. Mehra, S.S. Desai, S. Kuy, T.D. Henry, A.N. Patel, Cardiovascular disease, drug therapy, and mortality in covid-19, N. Engl. J. Med. 382 (25) (2020) e102, http://dx.doi.org/10.1056/NEJMoa2007621.

[227] J.J. Mulvey, C.M. Magro, L.X. Ma, G.J. Nuovo, R.N. Baergen, Analysis of complement deposition and viral rna in placentas of covid-19 patients, Ann. Diagn. Pathol. 46 (2020) 151530, http://dx.doi.org/10.1016/j.anndiagpath.2020.151530.

[228] H. Zeng, W. Chen, R. Zheng, S. Zhang, J.S. Ji, X. Zou, C. Xia, K. Sun, Z. Yang, H. Li, Changing cancer survival in china during 2003–15: A pooled analysis of 17 population-based cancer registries, Lancet Glob. Health 6 (5) (2018) e555–e567, http://dx.doi.org/10.1016/S2214-109X(18)30127-X.

[229] F. Fang, R. Steen, A. Casadevall, Misconduct accounts for the majority of retracted scientific publications, Proc. Nat. Acad. Sci. 42 (109) (2012) 17028–17033.

[230] S. Orchard, S. Kerrien, S. Abbani, B. Aranda, J. Bhate, S. Bidwell, A. Bridge, L. Briganti, F. Brinkman, G. Cesareni, A. Chatr-aryamontri, E. Chautard, C. Chen, M. Dumousseau, J. Goll, R. Hancock, L. Hannick, I. Jurisica, J. Khadake, D. Lynn, U. Mahadevan, L. Perfetto, A. Raghunath, S. Ricard-Blum, B. Roechert, L. Salwinski, V. Stümpflen, M. Tyers, P. Uetz, I. Xenarios, H. Hermjakob, Protein interaction data curation: the international molecular exchange (IMEx) consortium, Nature Commun. 10 (1) (2019) 345–350.

[231] L. Perfetto, C. Pastrello, N. Del-Toro, M. Duesbury, M. Iannuccelli, M. Kotlyar, L. Licata, B. Meldal, K. Panneerselvam, S. Panni, N. Rahimzadeh, S. Ricard-Blum, L. Salwinski, A. Shrivastava, G. Cesareni, M. Pellegrini, S. Orchard, I. Jurisica, H. Hermjakob, P. Porras, The IMEx coronavirus interactome: An evolving map of coronaviridae-host molecular interactions, database, 2020, http://dx.doi.org/10.1093/database/baaa096, published online 2020 Nov 18.

[232] P. Porras, E. Barrera, A. Bridge, N. del Toro, G. Cesareni, M. Duesbury, H. Hermjakob, M. Iannuccelli, I. Jurisica, M. Kotlyar, L. Licata, R. Lovering, D. Lynn, B. Meldal, B. Nanduri, K. Paneerselvam, S. Panni, C. Pastrello, M. pellegrini, L. Perfetto, N. Rahimzadeh, P. Ratan, S. Ricard-Blum, L. Salwinski, G. Shirodkar, S. Anjali, S. Orchard, Towards a unified open access dataset of molecular interactions, Nat. Commun. 1 (11) (2020) 6144–6156, http://dx.doi.org/10.1038/s41467-020-19942-z, published online 2020 Dec 1.

[233] M. Hund, D. Boehm, W. Sturm, M. Sedlmair, T. Schreck, T. Ullrich, D.A. Keim, L. Majnaric, A. Holzinger, Visual analytics for concept exploration in subspaces of patient groups: Making sense of complex datasets with the doctor-in-the-loop, Brain Inform. 3 (4) (2016) 233–247, http://dx.doi.org/10.1007/s40708-016-0043-5.

[234] D. Keim, G. Andrienko, J.-D. Fekete, C. Gorg, J. Kohlhammer, G. Melançon, Visual analytics: Definition, process, and challenges, in: Lecture notes in computer science, vol. 4950, 2008, pp. 154–176, http://dx.doi.org/10.1007/978-3-540-70956-5_7.

[235] D. Streeb, M. El-Assady, D.A. Keim, M. Chen, Why visualize? Arguments for visual support in decision making, IEEE Comput. Graph. Appl. 41 (2) (2021) 17–22, http://dx.doi.org/10.1109/MCG.2021.3055971.

[236] A. holzinger, H. Mueller, Toward human-ai interfaces to support explainability and causability in medical ai, IEEE IEEE COMPUTER, 54, (10). http://dx.doi.org/10.1109/MC.2021.3092610.

[237] M. Lopez-Cobo, G. De Prato, G. Alaveras, R. Righi, S. Samoili, J. Hradec, L. Ziemba, K. Pogorzelska, M. Cardona, Academic offer and demand for advanced profiles in the eu. in: Artificial Intelligence, High Performance Computing and Cybersecurity, Publications Office of the European Union, Luxembourg, Brussels, 2019, http://dx.doi.org/10.2760/016541.