UNIVERSITÄT
LEIPZIG

Machine learning for integrating data in biology and medicine: Principles, practice and opportunities

Leipzig, 31.01.2020

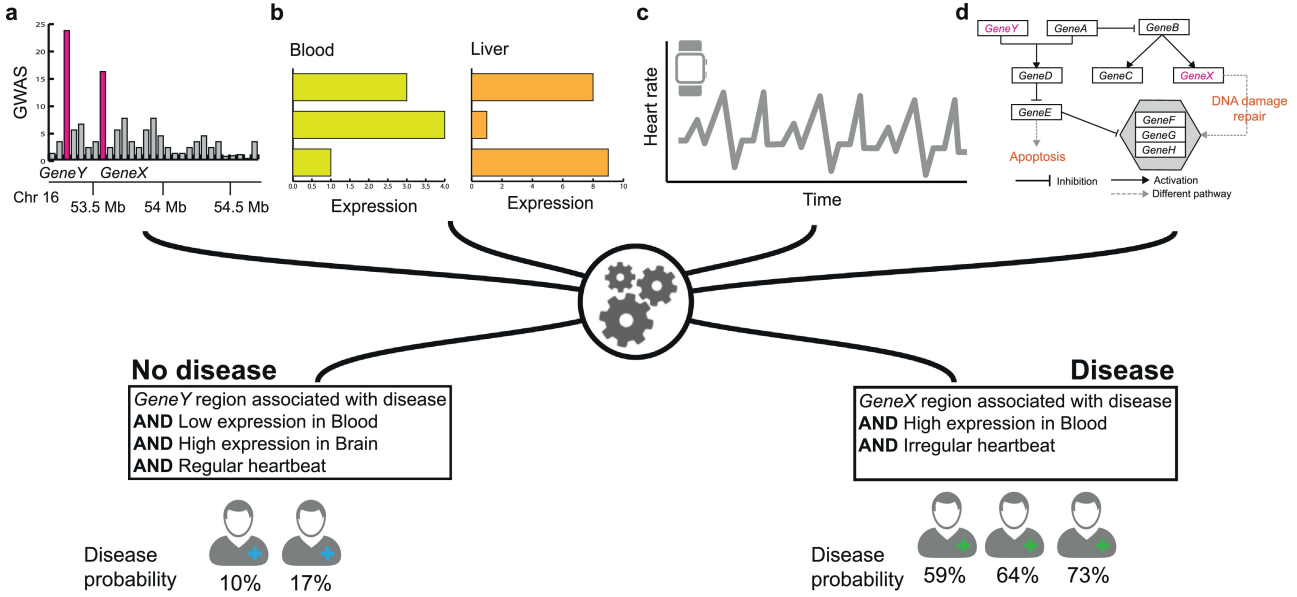
Vanessa Jehle

INHALT

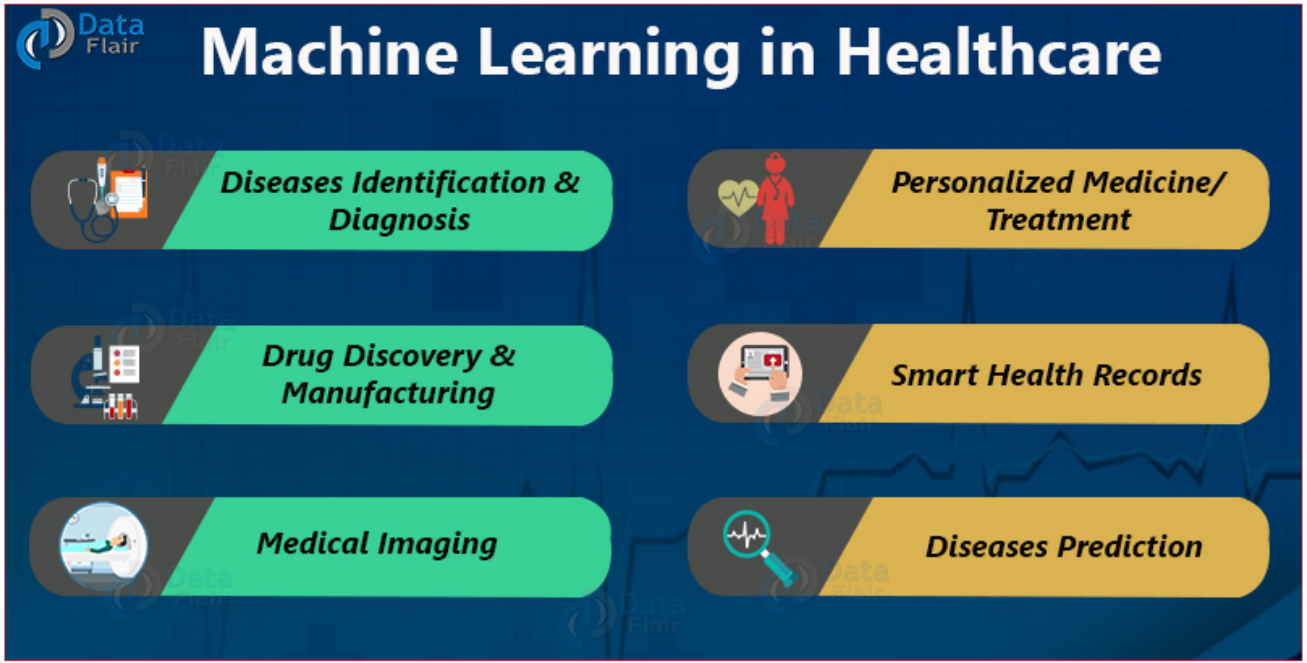
1. EINFÜHRUNG (1)

- komplexe biologische Systeme verstehen: Herausforderung für viele Forscher
- steigende Menge an verschiedenartigen Daten
- eine Datenart: viele wichtige Muster übersehen, worst case: falsch-positive Schlussfolgerung durch irrreführende Informationen
- Daten kombinieren: gesamtheitliches Bild des zu untersuchenden Phänomens

1. EINFÜHRUNG (2)



1. EINFÜRUNG (3)

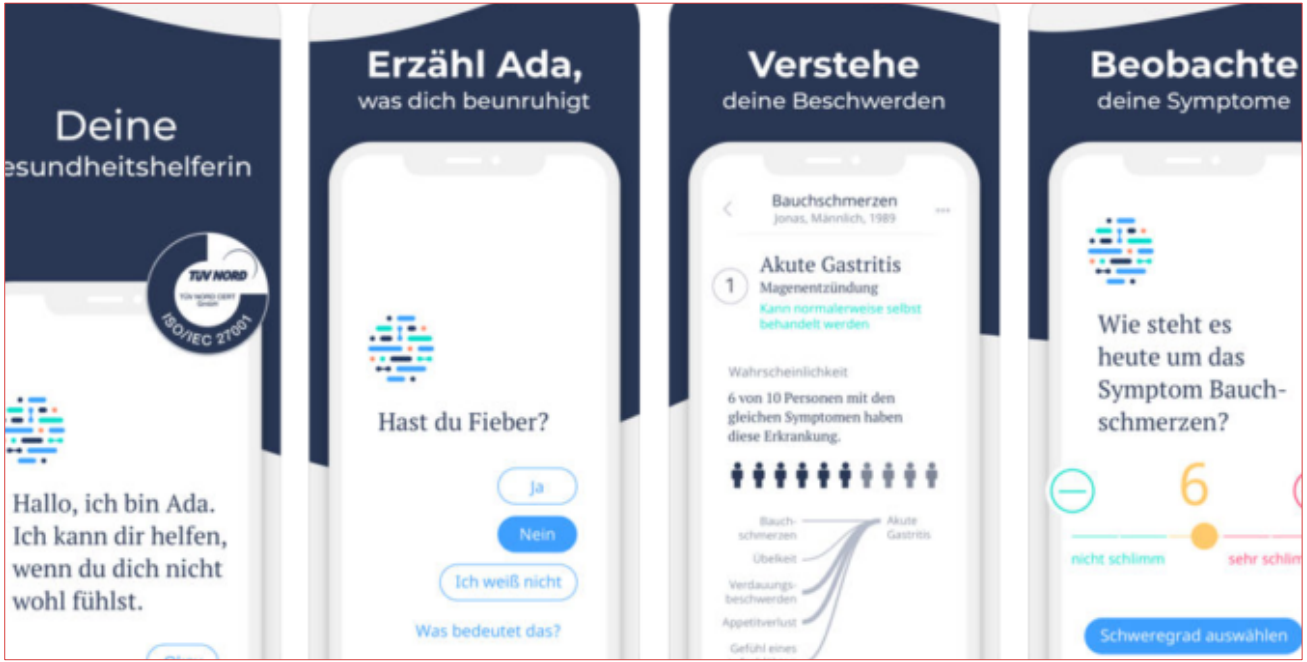


Wachsendes Angebot an Machine Learning Anwendungen im Medizinbereich

1. EINFÜHRUNG (4)

- Viele Anwendungen nutzen Datenintegration, um verlässliche Aussagen treffen zu können
- Beispiel: ADA
- „Ada stellt dir einfache Fragen und vergleicht deine Antworten mit Tausenden von ähnlichen Fällen, um die wahrscheinlichsten Ursachen für deine Symptome zu ermitteln“ (Website ada.com)
- Basiert auf umfassender Datenbank, Erweiterung durch Auswertung anonymisierter Patientendaten

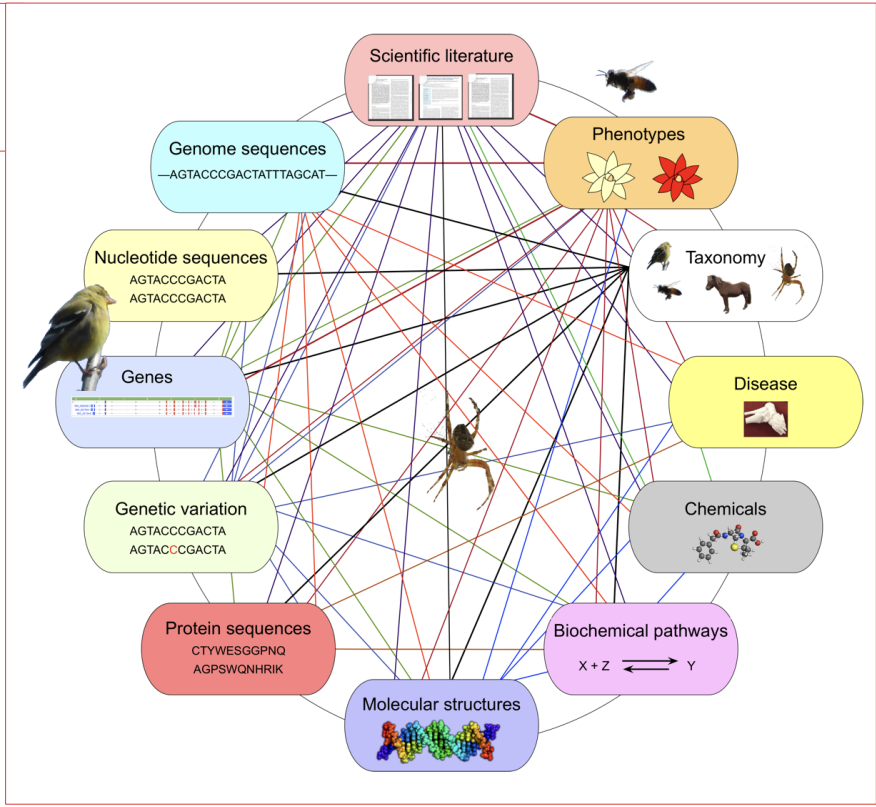
ADA



ADA Werbefoto

2. GRUNDLAGEN: DATEN

- Verschiedene Datensätze aus unterschiedlichen Quellen



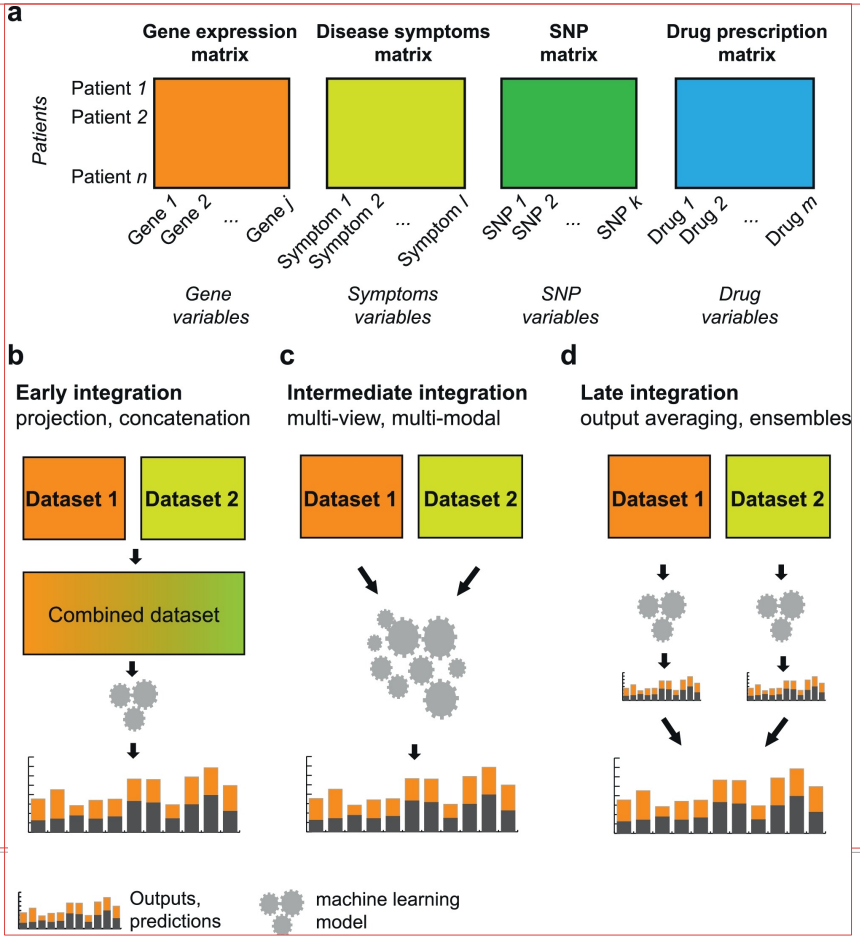
Biologische Datensätze

2. GRUNDLAGEN: HERAUSFORDERUNG DER DATENINTEGRATION

- biomedizinische Datensätze: hochdimensional, unvollständig, verzerrt, heterogen, verrauscht, dynamisch
- hochdimensional, aber spärlich: wenig qualitativ hochwertige Beispiele (z. B. GWAS)
- unvollständig und verzerrt durch Einschränkungen der Messtechnolgie sowie natürliche und physikalische Einschränkungen, voreingenommene Forscher
- dynamisch: Krebszellen, Bakterien, Viren entwickeln sich rasant um Arzneimittelresistenz zu erlangen

→ für umfassendes Verständnis entscheidend, vielfältige Informationsquellen zu vereinen

2.1 METHODEN DER DATENINTEGRATION (1)



2.1 METHODEN DER DATENINTEGRATION (2)

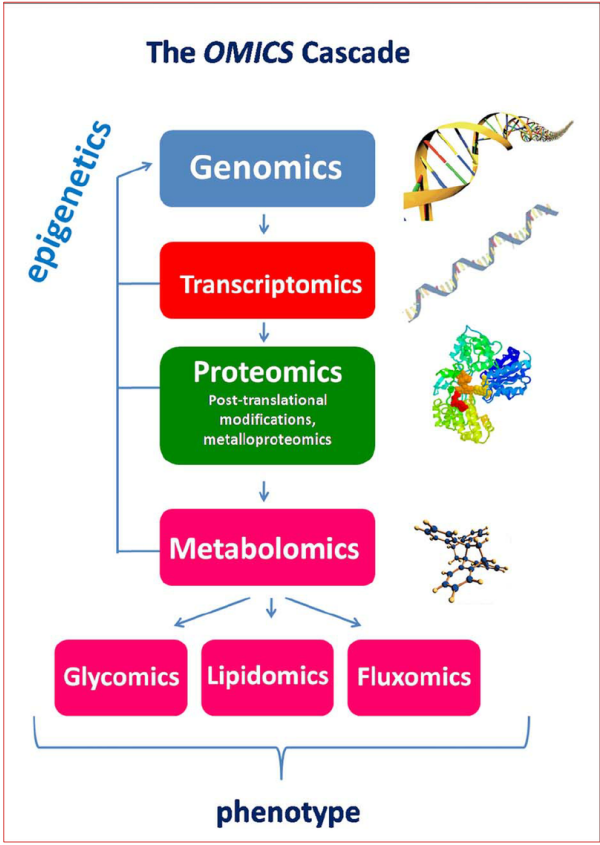
- early integration: Kombination von Daten vor der Analyse
  - Transformation der Datensätze zu einer merkmalsbasierten Tabelle oder grafischen Darstellung -> Eingabe für Machine Learning -> Ausgabe
  - Vorteil: Abhängigkeiten zwischen den Daten werden erhalten
  - Nachteil: Individuelle Eigenschaften der einzelnen Datensätze gehen verloren (Sturktur, Informationsgrad)
- late integration: Kombination der Ergebnisse mehrerer unabhängiger Analysen
  - Erste Modelle durch Analyse jedes einzelnen Datensatzes, zweites Modell mit Vorhersagen der ersten Modelle als Eingabe -> Ausgabe

2.1 METHODEN DER DATENINTEGRATION (3)

  - Vorteil: Struktur der einzelnen Datensätze wird erhalten
  - Nachteil: Beziehung zwischen den Datensätzen geht verloren, z. B. Korrelation oder Interaktion
  - Gewichtung der ersten Modelle schwierig
- early und late integration: Es kommen grundsätzlich alle Machine Learning Verfahren in Frage
- intermediate integration: Kombination von Daten als ein integraler Bestandteil des Analyse-Prozesses
  - Modell lernt gemeinsame Darstellung vieler Datensätze
  - Algorithmen, die auf Vielzahl von Datensätzen eingehen
  - Vorteil: Eigenschaften der Datensätze und Korrelationen bleiben erhalten
  - Nachteil: Oft neuer Algorithmus nötig

3. HAUPTTEIL

- Heterogene Daten zu jeder dieser Ebenen
- Fokus auf Patientenpopulation, insbesondere Pharmakologie
- Potential der Datenintegration, entscheidende Rolle bezüglich Gesundheit und Krankheit beim Menschen



3.1 RECHNERGESTÜTZE PHARMAKOLOGIE

Nutzt Daten um:

- Vorherzusagen, wie sich Medikamente auf den menschlichen Körper auswirken
- Entscheidungsfindung bei der Entwicklung von Arzneimitteln zu unterstützen
- Verbesserung der klinischen Praxis
- Verhinderung von unerwünschten Nebenwirkungen

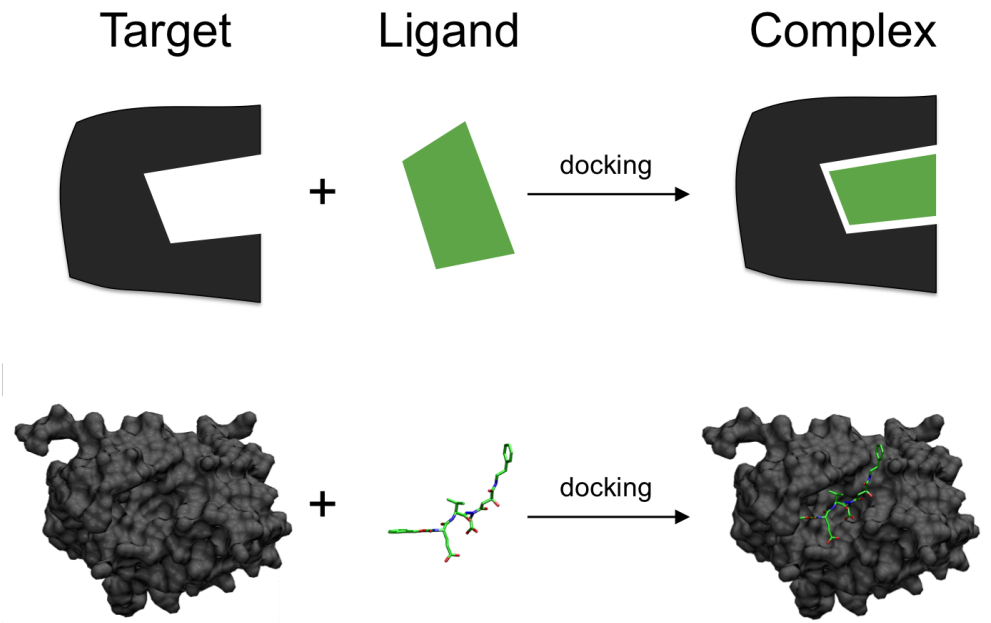
DRUG-TARGET INTERACTION PREDICTION (1)

- Medikamente haben eine Auswirkung auf den menschlichen Körper, indem sie sich an Zielstrukturen (targets) binden und deren Funktion aktivieren oder hemmen
- Zielstrukturen: oft Moleküle, die für Entstehen oder Fortbestehen einer Krankheit wichtig sind
- wichtig, wesentliche Eigenschaften von Arzneimitteln zu verstehen (Nebenwirkungen, Wirkungsweise, medizinische Indikation)
- traditioneller Ansatz Molekulares Docking

DRUG-TARGET INTERACTION PREDICTION (2)

- Molekulares Docking: Modellierung der Bindung von Arzneimitteln an medizinisch relevante Zielproteine, Bestimmung der Wahrscheinlichkeit der Interaktion mittels 3D-Modellierung und Computersimulation
- Ligandendocking als alternativen Ansatz
- bestimmt ein abstraktes Modell wichtiger chemischer Eigenschaften und modelliert Wirkstoffkandidaten flexibel, um Bindungsvorgang einschätzen zu können
- schlechte Ergebnisse, wenn Zielprotein nur kleine Anzahl bindender Liganden hat und das abstrakte Modell eine schlechte Qualität aufweist

LIGANDENDOCKING



Ligand (grün) dockt an Zielprotein (schwarz) an -> Protein-Ligand-Komplex

DRUG-TARGET INTERACTION PREDICTION (3)

- Machine Learning basierend auf dem „guilt-by-association principle“: ähnliche Arzneimittel neigen dazu an ähnliche Zielstrukturen zu binden und umgekehrt
- Nutzung von chemischen Strukturen der Arzneimittel und DNA Sequenzen als Eingabe
- Ähnlichkeitsmaße (drug-drug, target-target) und Informationen zu Interaktion (drug-target)
- viele Methoden beziehen zusätzliche Informationen in das Modell ein, z. B. Nebenwirkungen

→ Vorhersage der Interaktion genauer

- z. B. Darstellung als heterogenes Netzwerk, Vorhersage mit random walks

DRUG-DRUG INTERACTION (1)

- Komplexe Krankheiten oder Begleiterkrankungen: mehrere Arzneimittel
- Höheres Risiko von Nebenwirkungen aufgrund Arzneimittelwechselwirkungen
- Risiko einer übermäßigen Reaktion auf Medikamente
- Wechselwirkungen sind schwierig zu bestimmen: Kombination von Arzneimitteln können sich unterschiedlich klinisch manifestieren, nur bei bestimmten Untergruppen von Patienten zutreffend sein
- unmöglich alle Arzneimittelpaare zu prüfen

DRUG-DRUG INTERACTION (2)

- n Arzneimittel: n(n-1)/2 paarweise Kombinationen, 100 Arzneimittel → 4950 Kombinationen
- rechnergestützte Methoden um Kombinationen zu bestimmen, die Wechselwirkungen haben könnten
- Wechselwirkungen: Synergist und Antagonist
- Synergist: Wirkstoffe, die sich gegenseitig verstärken (additiv und überadditiv)
- Antagonist: Wirkstoff, der anderen Wirkstoff in seiner Wirkungsweise hemmt
- Gemessen durch Dosis-Wirkungs-Kurve oder Zellviabilität

DRUG-DRUG INTERACTION (3)

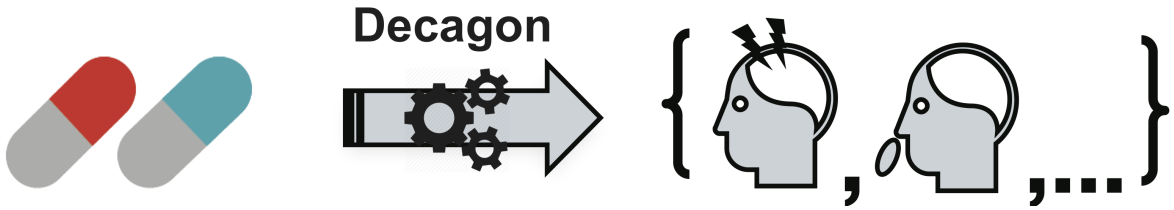
- Rechnergestütze Methoden nutzen Dosis-Wirkungs-Kurve oder Zellviabilität, um Arzneimittel mit möglichen Wechselwirkungen zu bestimmen
- Ansätze: klassifikations- oder ähnlichkeitsbasiert
- Klassifikationsbasiert: Vorhersage der Wechselwirkung als binäres Klassifikationsproblem, bekannte Wirkstoffe mit Wechselwirkungen als positive Beispiele, alle anderen Medikamentenpaare als negative Beispiele

DRUG-DRUG INTERACTION (4)

- Ähnlichkeitsbasiert: ähnliche Wirkstoffe haben ähnliche Interaktionsmuster
- Kombination verschiedener Ähnlichkeitsmaße zweier Wirkstoffe (z. B. chemische Substruktur, Nebenwirkungen, off-target side effects)
- Zusammenfassen der Ähnlichkeitsmaße mittels Clustering oder label propagation um Wechselwirkungen vorherzusagen
- Problem: Oft kann vorhergesagt werden, ob eine Wechselwirkung auftritt, aber nicht, wie sich diese im Patient darstellt

DRUG-DRUG INTERACTION (5)

- Neue Verfahren konzentrieren sich nicht nur auf die Vorhersage, ob Wechselwirkungen auftreten, sondern bestimmen, wie sich ein bestimmtes Wirkstoffpaar innerhalb einer Patientenpopulation klinisch manifestiert
- Nutzen molekulare Daten sowie Daten über Wirkstoff und Patienten um Nebenwirkungen vorherzusagen
- Beispiel: **Decagon**

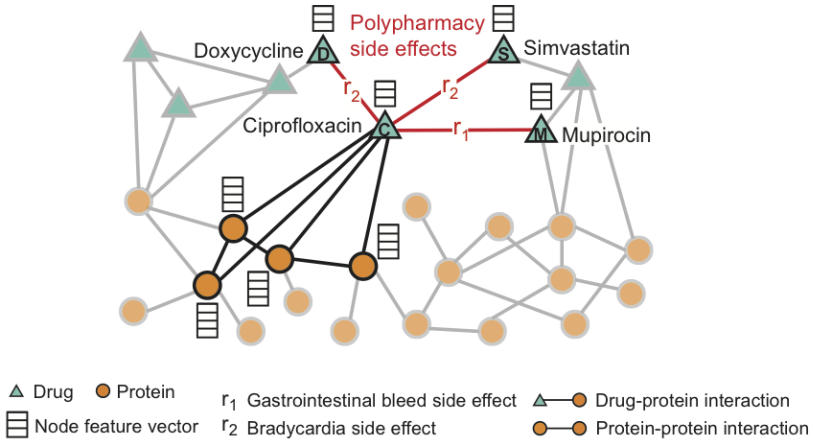


DECAGON (1)

- graph convolutional neural network
- multimodaler Graph mit Protein-Protein Interaktionen, Wirkstoff-Protein Interaktionen und Wechselwirkungen zweier Wirkstoffe
- Jede Art von Nebenwirkung als unterschiedlicher Kantentyp im multimodalen Graph
- Nutzung des Graphs zur Entwicklung eines convolutional neural networks um die klinische Manifestation von Nebenwirkungen bestimmter Wirkstoffpaare vorherzusagen

DECAGON (2)

- Verwendung von molekularen und pharmakologischen Daten sowie Daten der Patientenpopulation
- erkennen und priorisieren von Wechselwirkungen für weitere Untersuchungen (z. B. klinische Studien)



DECAGON (3)

- Encoder: GCN arbeitet auf Graph und erzeugt Einbettungen (embeddings) für Knoten
- Decoder: Nutzt embeddings, um Wahrscheinlichkeit für fehlenden Knoten (unbekannte Wechselwirkung zweier Medikamente) vorherzusagen
- Training: End-to-end, Minimierung der Verlustfunktion, Weitergabe des Gradienten der Verlustfunktion über Encoder und Decoder um alle trainierbaren Parameter zu optimieren

DECAGON (4) - ERGEBNISSE

- Ergebnisse: Nebenwirkungen mit starker molekularer Basis leichter vorhergesagt als umweltbedingte oder phänotypische, da der Graph hauptsächlich pharmakogenomische Daten enthält
- Vorhersage unbekannter Wechselwirkungen: Für 5 der wahrscheinlichsten 10 Wechselwirkungen Hinweise in medizinischer Literatur

- Pharmakogenomik: Untersuchung, wie Gene die Reaktion einer Person auf Medikamente beeinflussen

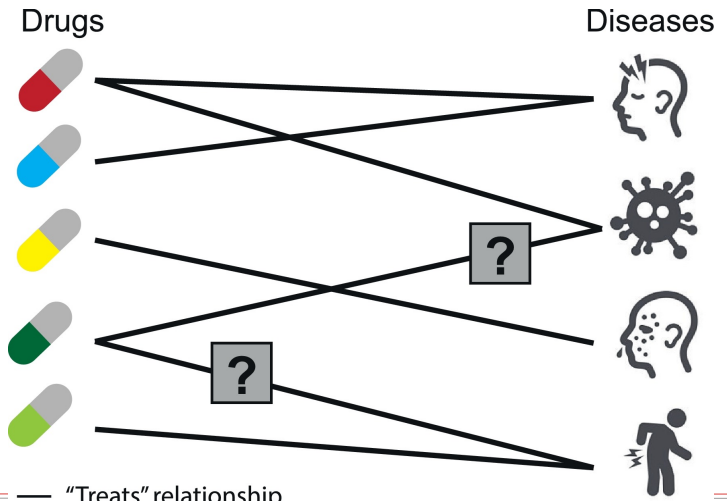
DECAGON (5) – ERGEBNISSE VORHERSAGE

New polypharmacy side effect predictions given by (drug *i*, side effect type *r*, drug *j*) triples that were assigned the highest probability scores by *Decagon*

k	Polypharmacy effect r	Drug i	Drug j	Evidence
1	Sarcoma	Pyrimethamine	Aliskiren	Stage *et al.* (2015)
4	Breast disorder	Tolcapone	Pyrimethamine	Bicker *et al.* (2017)
6	Renal tubular acidosis	Omeprazole	Amoxicillin	Russo *et al.* (2016)
8	Muscle inflammation	Atorvastatin	Amlodipine	Banakh *et al.* (2017)
9	Breast inflammation	Aliskiren	Tioconazole	Parving *et al.* (2012)

DRUG REPURPOSING (1)

- Neue Anwendungsgebiete für bekannte Medikamente
- Viele Medikamente haben multiple Zielstrukturen, Einsatz für mehr als einen Zweck möglich
- Verschiedene Erkrankungen teilen sich genetische Faktoren, molekulare pathways und Symptome

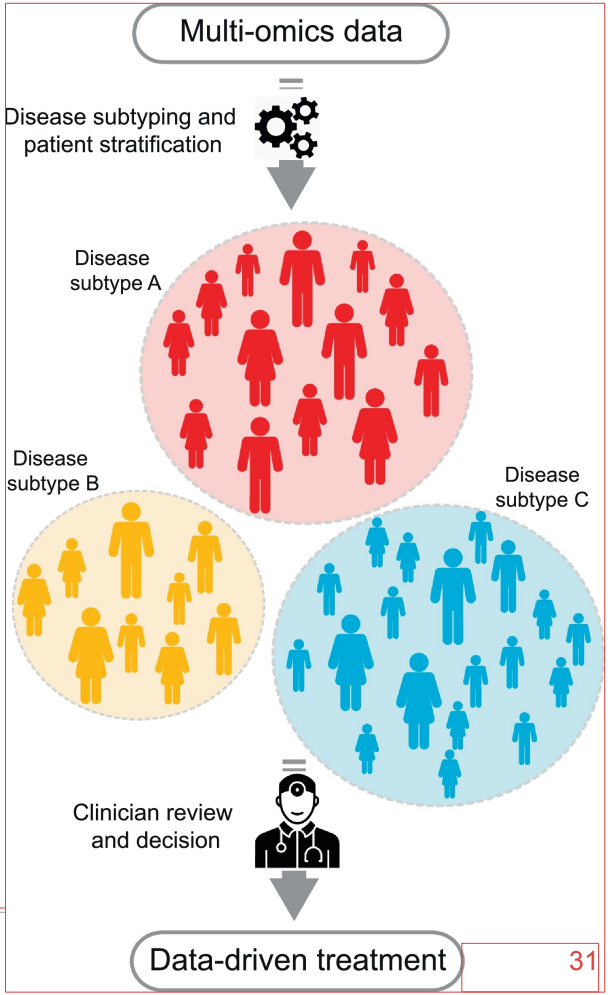


DRUG REPURPOSING (2)

- 4 Ansätze:
- 1) Vorhersage neuer Anwendungsgebiete bekannter Arzneimittel durch das Wissen der Zusammenhänge zwischen Protein und Zielstrukturen
- 2) Analyse der Aktivierung der Genexpression nach unterschiedlichen medikamentösen Behandlungsplänen
- 3) Vorhersagen basierend auf Nebenwirkungen
- 4) Ähnlichkeitsmaße (Krankheit, Medikament), unter Zuhilfenahme unterschiedlicher biomedizinischer Erkenntnisse

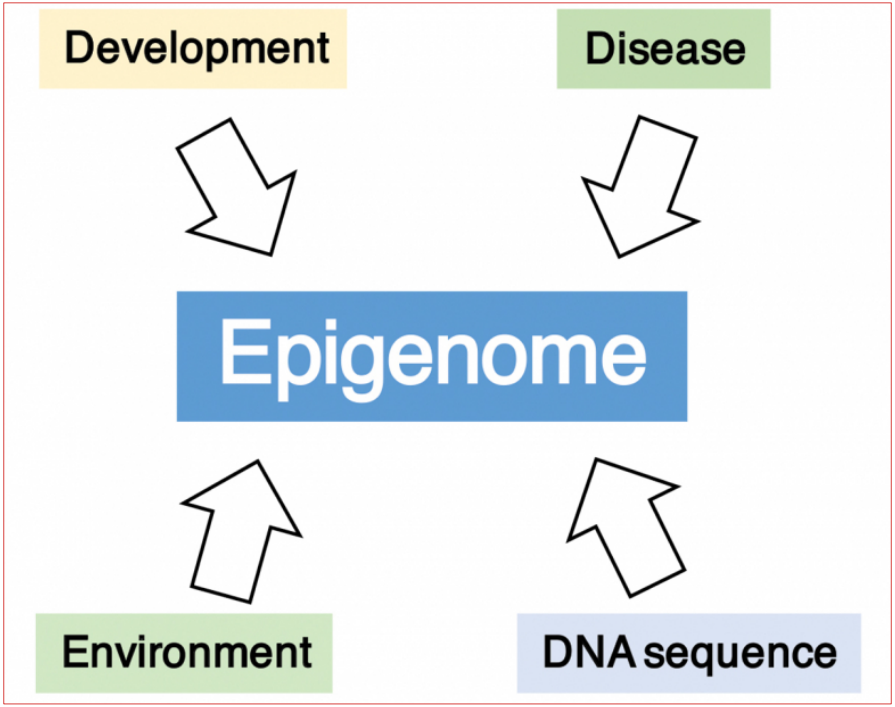
3.2 ANDERE ANWENDUNGEN

- Subtypisierung von Krankheiten



3.2 ANDERE ANWENDUNGEN

- Analyse von epigenetischen Daten
- Steuerung der Genexpression



4. ZUSAMMENFASSUNG

- Datenintegration ist in vielen Anwendungen in der Biologie und Medizin für gute Ergebnisse zwingend notwendig
- Viele verschiedene Datensätze, unterschiedliche Ansätze und Anwendungen in diversen Bereichen
- Pharmakologie bietet zahlreiche Anwendungsfälle und birgt großes Potential für bedeutsame Entwicklungen
- Forschung: starker Fokus auf Wechselwirkungen zwischen Arzneimitteln

UNIVERSITÄT
LEIPZIG

VIELEN DANK!

QUELLEN

- Marinka Zitnik, Monica Agrawal, Jure Leskovec, Modeling polypharmacy side effects with graph convolutional networks, *Bioinformatics*, Volume 34, Issue 13, 01 July 2018, Pages i457–i466
- https://f1000research.com/posters/8-1262
- Chen, X., Liu, M., & Yan, G. (2012). Drug-target interaction prediction by random walk on the heterogeneous network. *Molecular bioSystems, 8* 7, 1970-8
- Emmons, S. (2019). *Learning on Graphs: Supervised and Unsupervised Methods.*
- Aggarwal, C. and Reddy, C. (2015). *Healthcare Data Analytics*. CRC Press.
- Zitnik, M., Nguyen, F., Wang, B., Leskovec, J., Goldenberg, A. and Hoffman, M. (2019). Machine learning for integrating data in biology and medicine: Principles, practice, and opportunities. *Information Fusion*, 50, pp.71-91.

QUELLEN ABBILDUNGEN

- https://data-flair.training/blogs/machine-learning-in-healthcare/
- https://smw.ch/article/doi/smw.2017.14523
- https://en.wikipedia.org/wiki/File:Docking_representation_2.png
- https://digitalworldbiology.com/products/exploring-biological-databases
- https://www.researchgate.net/profile/Carlos_Ambrosio/publication/259465484/figure/fig1/AS:297133921062915@1447853655427/Schematic-representation-of-the-development-of-the-omic-fields-The-omics-fever-began.png
- https://www.gim-radar.de/wp-content/uploads/2018/11/ada_app-672x372.png
- https://data-flair.training/blogs/machine-learning-in-healthcare/