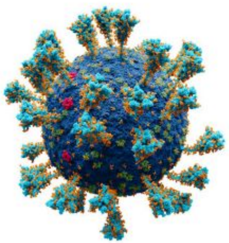
Answers for offline activities of AMIDD 2023

Jitao David Zhang, December 2023

Lecture 1: Introduction

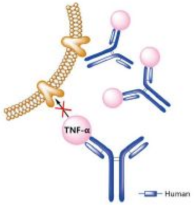
Offline Activities of Lecture 1, AMIDD 2023

1 Comirnaty
(COVID-19 Vaccine)



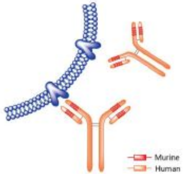
$37.806 Billion
Infectious Diseases

2 Humira
(Adalimumab)



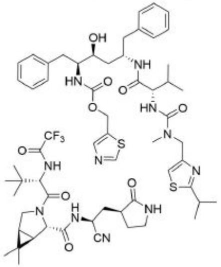
$21.237 Billion
Immunology

3 Keytruda
(Pembrolizumab)



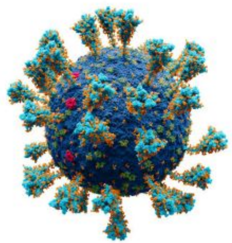
$20.937 Billion
Oncology

4 Paxlovid
(Ritonavir/Nirmatrelvir)



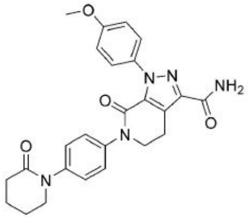
$18.933 Billion
Infectious Diseases

5 Spikevax
(CX-024414)



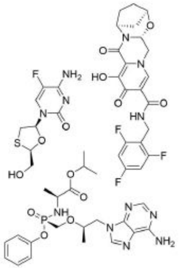
$18.435 Billion
Infectious Diseases

6 Eliquis
(Apixaban)



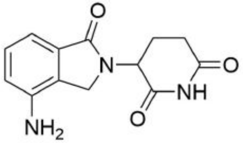
$11.789 Billion
Cardiology/Vascular Diseases

7 Biktarvy
(Bictegravir/Emtricitabine/Tenofovir Alafenamide)



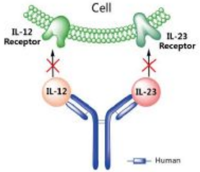
$10.390 Billion
Infectious Diseases

8 Revlimid
(Lenalidomide)



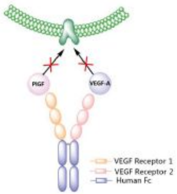
$9.978 Billion
Oncology

9 Stelara
(Ustekinumab)



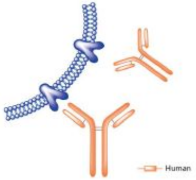
$9.723 Billion
Immunology

10 Eylea
(Aflibercept)



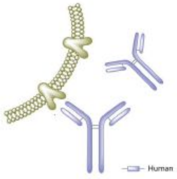
$9.639 Billion
Ophthalmology

11 Opdivo
(Nivolumab)



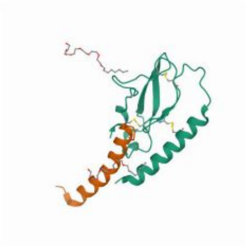
$9.299 Billion
Oncology

12 Dupixent
(Dupilumab)



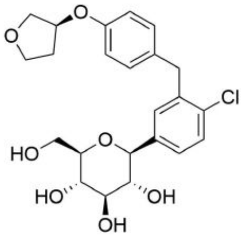
$9.095 Billion
Immunology

13 Ozempic
(Semaglutide)



$8.713 Billion
Diabetes

14 Jardiance
(Empagliflozin)



$8.388 Billion
Diabetes

Top 14 pharmaceuticals by sales in 2022

Poster compiled by the Jon Njardarson group at University of Arizona (https://njardarson.lab.arizona.edu). Citation: J. Chem. Ed. 2010, 87, 1348.

Questions: (1) How many are small molecules, proteins, and oligonucleotides each? (2) Are there other modalities? (3) What patterns do you observe? (4) Do you have explanations for these patterns?

Drug	Modality	Drug	Modality
Comirnaty (COVID-19 vaccination)	Oligonucleotides/ Other modalities	Revlimid (lenalidomide, mutlple myeloma)	Small molecules
Humira (Adalimumab,anti-TNFalpha)	Proteins	Stelara (Ustekinumab, anti-IL12/IL23)	Proteins
Keytruda (Pembrolizumab, anti-PD1)	Proteins	Eylea (Aflibercept, anti-VEGF)	Proteins
Paxlovid (Ritonavir/Nirmatrelvir)	Small molecules	Opdivo (Nivolumab, anti-PD-1)	Proteins
Spikevax (COVID-19vaccination)	Oligonucleotides/ Other modalities	Dupixent (Dupilumab, anti-IL-4Ralpha)	Proteins
Eliquis (Apixaban,factor Xa inhibitor)	Small molecules	Ozempic (semaglutide,GLP1 analog)	Proteins (peptides)/Other modalities
Biktarvy (cocktail for HIV)	Small molecules	Jardiance (Empagliflozin, SGLT-2inhibitor)	Small molecules

What patterns do you observe based on the distribution of modalities, as well as on the indications (given by the color bars below the compounds)? How would you explain the observations of the patterns?

1. Many antibodies can be found under top-selling drugs. Antibody drugs are in general more expensive due to many factors (complexity of production, storage/delivery, indications, etc.). Due to the coronavirus pandemic, oligonucleotides are also much sold in 2022.
2. Life-threatening diseases (e.g. oncology and infectious disease), diseases with severe impact on life quality (e.g. autoimmune disease and ophthalmology), and chronic diseases (e.g. cardiovascular metabolism) tend to generate high sales.

Questions

Can be said that nowadays antibodies drug can be considered "old", because of the progress in gene therapy, like the new micro dystrophin of Sapreta Therapeutics against neuromuscular disease (DMD)? I would agree with the observation that antibodies are an established modality. Much work is still ongoing to improve antibodies, for instance bi-specifics, antibody-drug conjugates, etc.

I didn't select "other modalities" for any of the drugs listed because it's not so clear what does "other modalities" mean here. Some pharmaceuticals could be classified into multiple modalities. E.g., vaccines could be classified ad both "oligonucleotides" and "other modalities", since they can be encapsulated in nanoparticles, and often contain other components such as adjuvants. Likewise, combination therapies involve drugs with different modalities to achieve a greater therapeutic effect. Ultimately, the classification of pharmaceuticals can be somewhat fluid. As new technologies emerge, existing categories may need to be revised to accommodate new modalities. Definitely correct observation: the definition of modality is often not black-white, many drugs can be put in more than one categories.

It would be cool if we could discuss a bit more on the technical terms needed for those off-lecture activities. I had no idea (and I am still not sure) what oligonucleotides are or what exactly is meant by "small molecules". Besides that the lecture is very cool so far. Thank you! I will try to improve this in the future.

What happens if a molecule doesn't have the right form? Does it change so it fits or is it being discarded by the body? It depends. If a wrong modality is chosen, the probability is high that the drug candidate will fail in discovery or clinical development, usually due to lack of efficacy or excessive toxicity.

What modality is a vaccine? It depends. Often vaccines are killed viral particles, deactivated or recombinant viral proteins, sometimes mRNAs.

Lecture 2 on drug targets and mechanistic modelling

Follow up of questions on the video on Herceptin by Susan Desmond-Hellmann

Link to the video

Questions for the video

1. What is the **indication** of *Herceptin*? (Her2 positive breast cancer) What is its generic (USAN, or United States Adopted Name) name? (Trastuzumab)
2. What is the **gene target** of Herceptin? (Her2, ERBB2)
3. Which class best describes the target: Enzyme, Ion channel, Receptor and Kinase, or Structural protein? (Receptor and kinase)
4. In which year was the **target** of Herceptin described? When was Herceptin **approved**? (1987; 1998 in metastatic cancer and 2005 in the adjuvant setting)
5. What was the **improvement** of Herceptin compared with earlier antibodies? (humanized)
6. Why does a **biomarker** matter besides developing drugs? (diagnostic, higher chance of success due to patient stratification)
7. In the clinical trial of *Herceptin* for **metastatic breast cancer**, how much improvement in the **median survival** did Herceptin achieve? And how much improvement is in the **adjuvant setting** (Herceptin applied directly after operation)? (5.1 months improvement in median survival for metastatic breast cancer. Time to remission doubled in the adjuvant setting)

Questions

Is HER2 only connected to breast cancer or also other types of cancer? Good question. While ERBB2/Her2 overexpression is mostly known for certain breast cancers, HER2 overexpression is also the cause of some stochmar and gastric cancers, for instance gastro-oesophageal adenocarcinoma. See relevant information provided by the European Society for Medical Oncology (ESMO).

Lecture 3 on statistical modelling and machine learning

Questions about machine learning

Which of the following statements are true? More than one option is possible.

- Features can be categorical,ordinal, or numerical. True
- Time in PK/PD modelling can be treated with Convolutional Neural Networks. False, though other neural networks may be used.
- Replacing a missing value with the feature mean is a safe operation against overfitting. False, it can even cause overfitting under certain circumstances.

What can we do with unbalanced dataset? Undersampling the majority class, oversampling the minority class, using balanced metrics (e.g. F1 score), or use generative models to enhance minority classes.

What is the main difference between supervised learning and unsupervised learning? Either we use the label of the data to guide the learning process (supervised learning), or labels are not available or we do not use the labels (unsupervised learning)

What clustering methods do you know? k-means clustering, density-based clustering, hierarchical clustering, etc.

What does overfitting mean? What measures can we take to avoid overfitting? The model has a good performance when applied to training/test data, however the performance is worse when it is applied to unseen data. In another word, overfitting means poor generalization ability of the model. To avoid overfitting, we may limit the complexity of the model, or enhance data used for training.

How can we measure model complexity and goodness of fit? Complexity is defined by the model used (e.g. linear model versus tree-model), the number of parameters, as well as the relationships between the parameters. Akaike information criterion and Bayesian information criterion can be used to judge the complexity.

Questions about Hill’s causality criteria

Take a two-sample t-test as an example: which of the following statistics indicate the strength of the difference?

- Total sample size No, it does not indicate the strength of the difference: a large sample size does not mean automatically a bigger difference
- The t-statistic Yes
- The p-value Yes

What does biological gradient mean in the context of Hill's causality criteria? The dose-response curve.

Hill claimed that 'No formal tests of significance can answer these questions.' What are the strengths and limitations of tests of significance when our goal is to establish causality?

1. A statistical test may or may test the causality directly.
2. Other factors mentioned by Hill, for instance biology, consistency, specificity, are not always embedded in the statistical test.
3. No test is a rigorous proof of causality. Experimentations, or at least mechanistic understanding, are needed to establish causality

Questions from students

What are examples for formal tests? For instance *t*-tests, Fisher’s exact test, etc.

In what situations might a statistically significant result not imply a causal relationship? Consider the toy example: compare the sales of ice creams on days with wildfire, versus days without, we expect to see a difference. However, neither ice creams or wildfire causa each other directly. Instead, a common factor (temperature) causes both. Other causal structures may lead to similar observations.

Very interesting papers, specially Hil's one. It makes me very happy to learn that you enjoyed Hill’s talk. In my opinion the paper should be read by more people.

I really enjoyed the reading of Hill. His caution against blindly using the "P"(value) as the criterion to judge on a "significant" difference in outcome between treatment groups without going into the study design, performing cross-validation or using biologist's domain knowledge is as valid today as it was back in the days. I cannot agree with you more!

Lecture 4 on causal inference

Questions

Use an example to explain a pipe structure to someone without statistical background/ Use an example to explain a fork structure to someone without statistical background. Most of you found interesting examples. Congratulations!

Why is collider structure particularly challenging for causal inference? (I copied one student's response, because I found it is so well written) Because uncorrelated variables can suddenly become correlated when we condition on the collider. The example from the lecture for example would be papers published in a scientific journal, where a negative correlation between trustworthiness of a study and how "newsworthy" it is appears.

What do we mean with refutation in causal inference? Analysis that we do to test the strength, validity, and confidence of the estimated causality.

Please find resources (articles, videos, etc.) to learn more about Mendelian Randomization. Use your own words explain it (ChatGPT detection is turned on, for your information). It means that we can use genetic information of individuals to avoid issues of confounding, since genetic information can be thought outcome of a natural randomized trial.

Lecture 5 & 6

Questions for the abstract of Bollag et al., 2010

1. What is the **indication** of PLX4032? BRAF-mutant melanoma
2. What is the **gene target** of PLX4032? The mutant form of BRAF, primarily V600E. The authors also suggested an activity against the V600K mutation.
3. The malignancy depends on which **biological pathway**? The RAF/MEK/ERK pathway.
4. What is the **Mechanism of Action** of PLX4032? PLX4032 inhibits the kinase activity of mutant BRAF, which inhibits ERK phosphorylation and blocks the RAF/MEK/ERK pathway in BRAF mutant cells.
5. What **went wrong** in the first **Phase I clinical trial**? And how was it **solved**? Patients did not respond, i.e. doctors observed no tumour regressions. The drug developer changed the formulation from crystalline to amorphous. The new formulation allowed higher drug exposures, which lead to high response rate.
6. What was the **dosing regimen** in the final **Phase I** clinical trial, and what is the response rate? Oral dose, 960 mg twice every day (bid, latin *bis in die*); Response rate: 81%.

Questions for Bollag et al., 2010 (I)

1. We learned that many drugs target one of the four protein types: GPCRs, ion channels, kinases, and nuclear receptors. Which type does the target of PLX4032 belong to? Kinases
2. How was the efficacy of PLX4032 tested? Cell lines, xenografts (mouse), and finally patients. Beagle dogs, cynomolgus monkeys, and rats were used to test safety, not efficacy.
3. Why was PLX4032 chosen for further development, but not PLX4720? Better scaling of pharmacokinetic (PK) properties.
4. How was the exposure of PLX4032 in the blood quantified? Which mathematical operation was used? Area under of curve of plasma concentrations - integration.
5. How was the final dosing regimen (960-mg BID) determined? It was the maximum tolerated dose, toxicities detected in higher-dose groups.

Questions for Bollag et al., 2010 (II)

6. How did patients with the V600K mutation in BRAF respond? They responded better, with 71% and 100% reduction in tumour dimensions.

7. What measures were taken to demonstrate the effect of BRAF inhibition in patient biopsies? Phosphorylated-ERK and Ki67 levels were measured. Values from both measurements were found to be decreased in the later measurements, showing the reduction in ERK pathway activity.These are biomarkers, i.e. measurements that correlates drug treatment with efficacy.

8. What side effects of PLX4032 were reported? Besides fatigue, rash, and joint pain, 31% patients treated at the maximum tolerated dose (MTD) developed skin lesions described as cutaneous squamous cell carcinomas, keratoacanthoma type.

9. What measures were taken against side effects and safety concerns of PLX4032? **Limiting the dose**, resection of the lesion, dermatological monitoring.

10. Where do you think mathematics and informatics is used in the discovery and development of PLX4032? *Almost every step*, especially X-ray data analysis, data summary and modelling.

Lecture 7 on Biological sequence analysis (no offline activity - feedback only)

Lecture 8: structure-based and ligand-based drug discovery

Questions about sequence analysis

Q1: Compare *p(*ACGTGGT|M*)* and *p(*ACCTGGT|M*)* The two sequences differ only at one position, but the problem is trickier than it seems! The contributions of first alphabet *A* and the last three alphabets *GGT* are cancelled out, leaving us to compare *p(CGT)* and *p(CCT)*, which equals p(G|C)p(T|G)/p(C|C)/p(T|C)=0.078*0.208/.298/.302, which gives 0.180. About half of us got the answer correctly, which is great. If you got a different answer, please check whether now you understand it better.

Q2: We have got a RNA sequence by sequencing sputum from a patient (see below). Please follow the instructions to find out the original genome of the sequence, and ideally the gene encoding the sequences?

ATGTTTGTTTTTCTTGTTTTATTGCCACTAGTCTCTAGTCAGTGTGTTAATCTTACAACCAGAACTCAATTACCCCCTGCATACACTAA
TTCTTTCACACGTGGTGTTTATTACCCTGACAAAGTTTTCAGATCCTCAGT

It turns out to be a stretch of sequence from the ORF1ab gene in the coronavirus genome.

Questions about *Evaluation of the Biological Activity of Compounds: Techniques and Mechanism of Action Studies*

Q1. An important chemical and mathematical concept was not described in the book chapter: what does *the Law of Mass Action* mean? (An ODE model of reaction rate and reactant mass)

Q2: Which quantity measures binding affinity directly: dissociation constant (K_D) or the concentration of the test compound that produces 50 percent inhibition (IC_{50})? (K_D)

Q3: In Figure 2.3, what do x- and y-axis represent in panel (A) and panel (B), respectively? (concentrations in in x-axis; y-axis: counts per minute of radioactivity (A), percentage of binding of the labelled compound)

Q4: What is a sigmoidal curve? (A S-shaped, logistic or logit curve)

Q5: Do IC_{50} values indicate a particular mechanism of action (MoA)? (No)

Q6: In a certain enzymatic assay,, two compounds have the following pIC50 values: 7.2 (Compound A), 9.3 (Compound B). If all other conditions are held constant, what is the relationship between binding affinities of the two compounds with regard to the target? (B>A)

Q7: Why is DMSO often used in bioassays? (solvent, control)

Q8: Can you use your own language to describe what is the Hill function? (discussed in Lecture 5)

Q9: What statistical measure is used to measure the signal-noise ratio in screening? Can you use your own language explaining it? (how well can we separate positive controls from negative controls)

Q10: Why logarithm (usually base 10) transformation is often preferred to represent quantities such as IC_{50} and K_i? (presentation, as well as statistical mechanistics)

Questions from you:

1. On page 19: what is meant with "displacement of a labelled ligand"? (I do not know what 'displacement' means in that context)
2. I didn't quite understand the application of the Z value and when it usually is used

Lecture 9: From interactions to networks

Questions

Q1. What is the method commonly used to benchmark performance of different techniques of computer-aided drug design (CADD)? (Receiver Operating Characteristic curves)

Q2: What do we mean by molecular dynamics? (A computer simulation method to analyze the movements of atoms and molecules using Newtonian mechanistics)

Q3: What are the three basic methods to represent target and ligand structures *in silico*? (atomic, surface, and grid representations)

Q4: What sampling algorithms are there for protein-ligand docking? Can you explain one of them using your words? (systematic algorithms, molecular dynamics simulations, Monte Carlo search with Metropolis Criterion and genetic algorithms)

Q5: What are the key steps in structure-based virtual high-throughput screenings (SB-vHTS)? (preparing structures, posing, scoring)

Q6: What is the usual starting point of structure-based CADD campaign? (Experimentally determined protein structures, preferably in complex with ligands)

Q7: What do we mean by 'pharmacophore'? (model of the target binding site which summarizes steric and electronic features needed for optimal interaction of a ligand with a target, a "subgraph" of a molecule with interesting properties for drug design/protein binding)

Q8: In QSAR analysis, why it is important to select optimal descriptors/features? (to reduce noise, to increase generalized performance, and for hypothesis generation)

Q9: What do we mean by the acronyms *DMPK* and *ADMET*? (DMPK=drug metabolism and pharmacokinetics; ADMET= absorption, distribution, metabolism, excretion, and the potential for toxicity)

Q10: Why common CADD methods have difficulties handling protein-protein interaction and protein-DNA interactions? (large interaction size, lack of user-friendly tools, and comparably little training data)

Lecture 10 on omics and MoA studies (PCA example - the results were discussed in the lecture)

Lecture 11 on PK/PD modelling (reading only - no offline activity)