Mathematical and Computational Biology In Drug Discovery (2025)

Dr. Jitao David Zhang

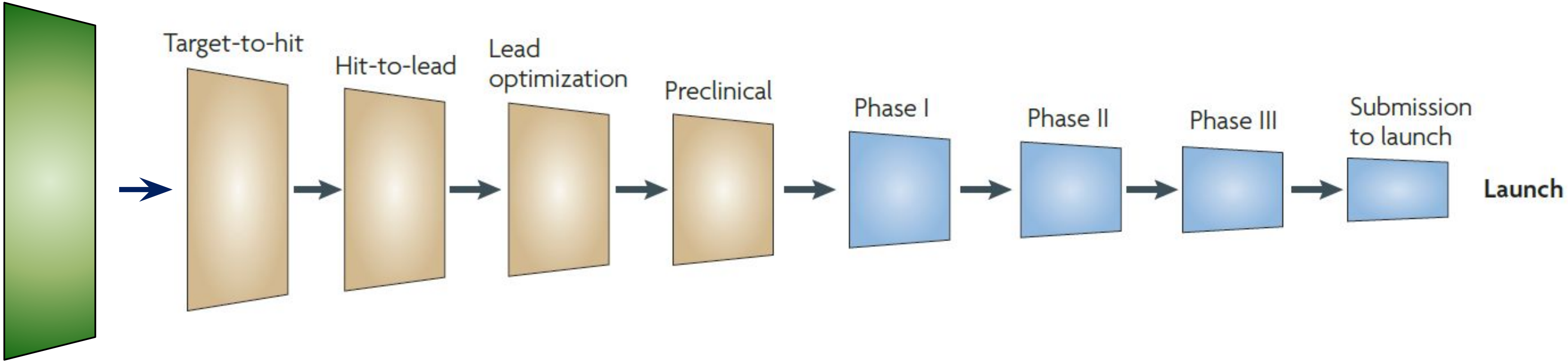
[1] Pharma Research and Early Development, Roche Innovation Center Basel, F. Hoffmann-La Roche
[2] Department of Mathematics and Computer Science, University of Basel

Administrivia

- Please fill the pre-course survey.
- Grades are given by participation (50%) and offline activities (50%).
- I hope that the course is more a seminar than a lecture: share your questions and let’s discuss!
- **Any more questions?**

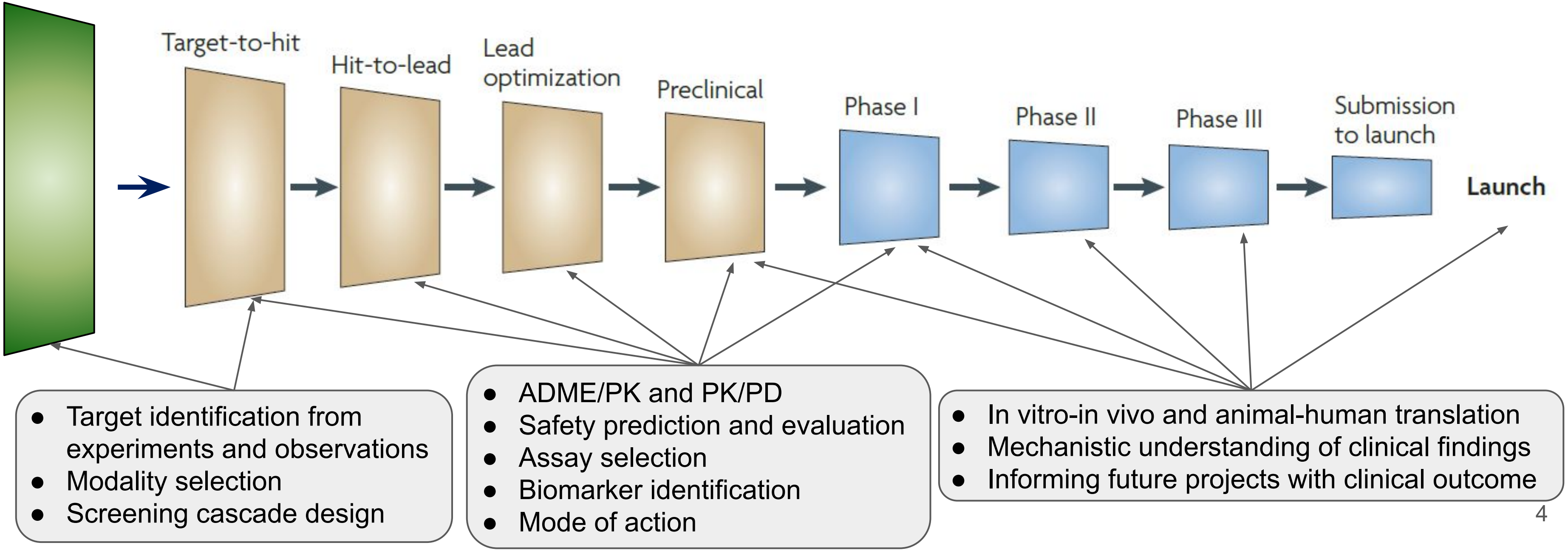
A linear view of drug discovery

Target identification & assessment



Mathematical and computational biology contributes at all R&D stages

Target identification & assessment



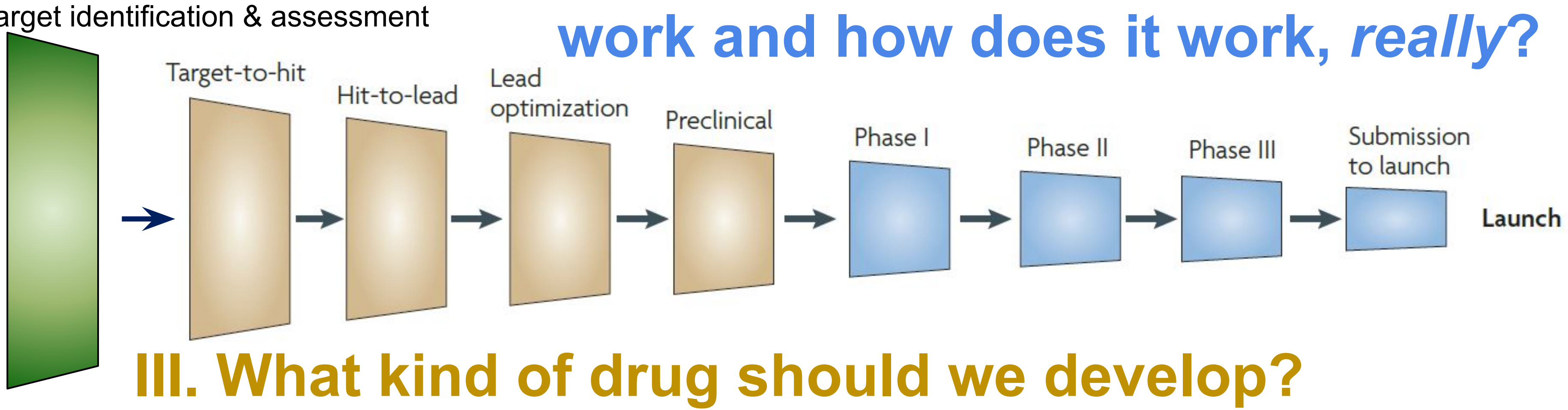
- Target identification from experiments and observations
- Modality selection
- Screening cascade design

- ADME/PK and PK/PD
- Safety prediction and evaluation
- Assay selection
- Biomarker identification
- Mode of action

- In vitro-in vivo and animal-human translation
- Mechanistic understanding of clinical findings
- Informing future projects with clinical outcome

Questions that we will address in this course

V: For which patients will the drug work and how does it work, *really*?



III. What kind of drug should we develop?

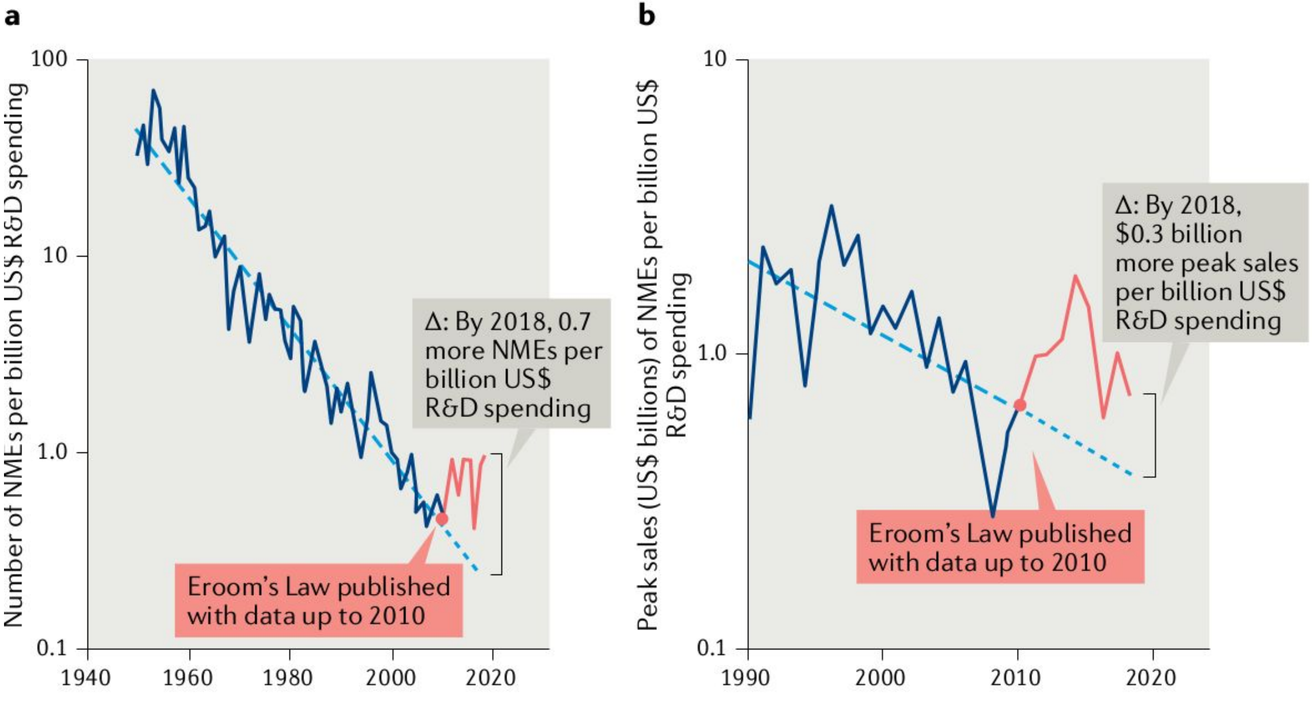
IV. What efficacy and safety profiles can we expect?

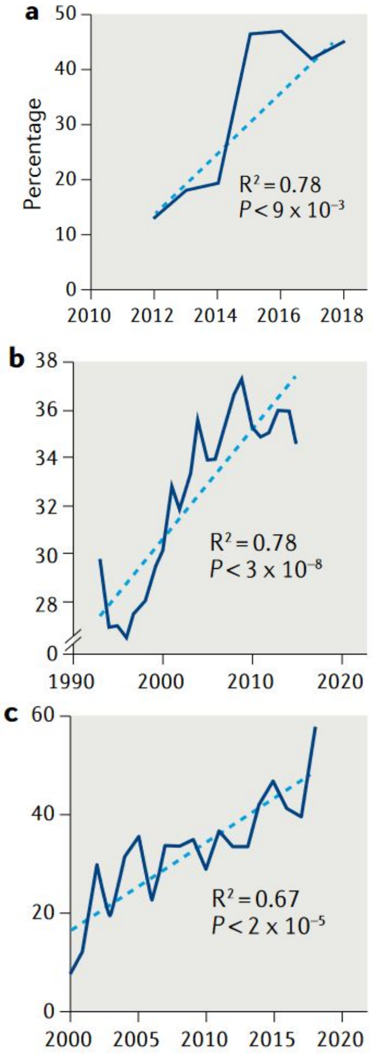
I. What makes a good drug target?

II. What can we do if there are no good targets?

Part 1: A few numbers

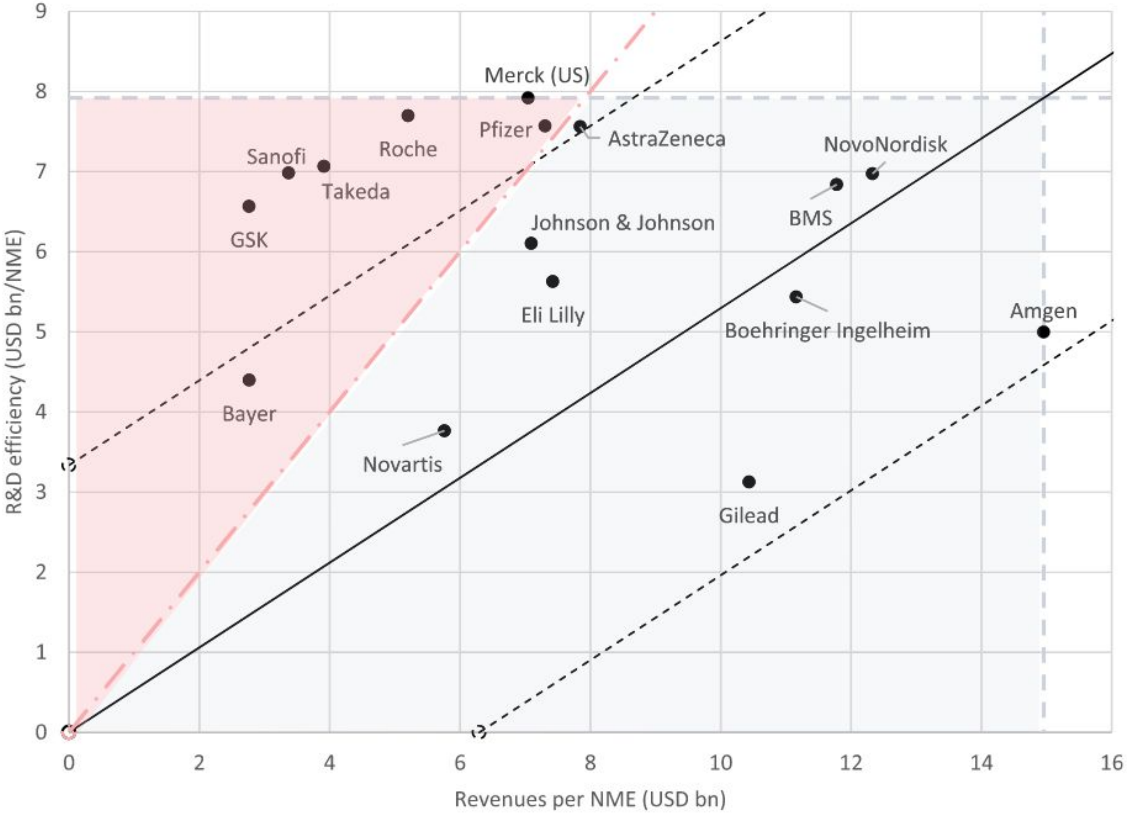
The Eroom’s Law



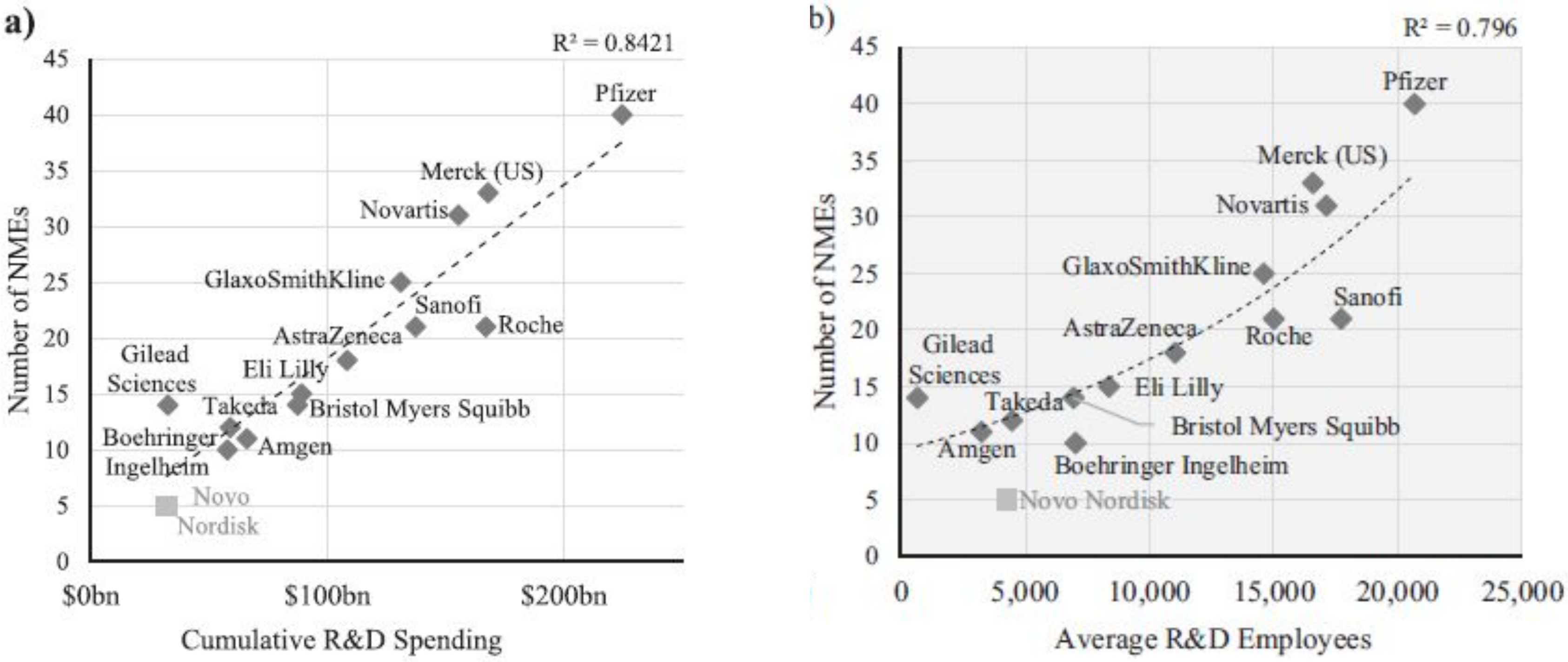


Left: R&D cost by year. Right: correlation with genetic evidence (a), narrow indication (b), and rare diseases (c).

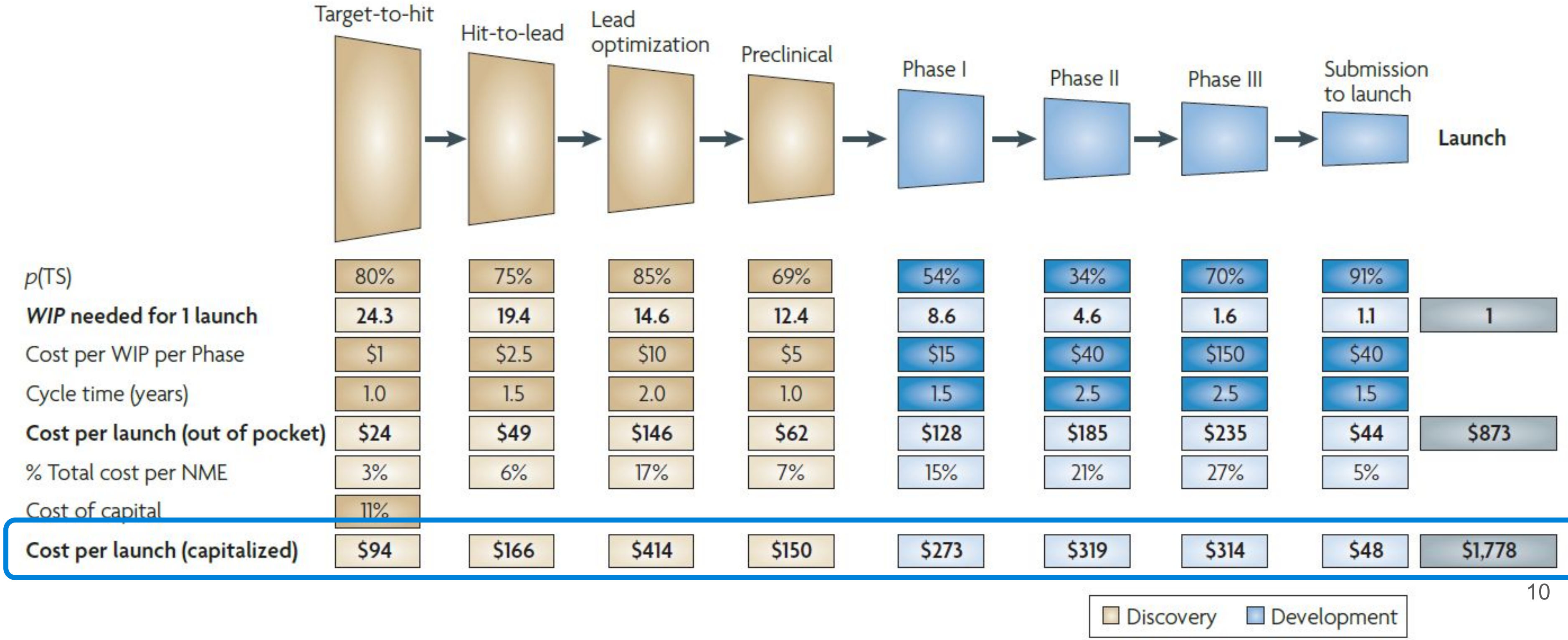
R&D productivity of leading pharma companies (2001–2020)



Relationship between new molecule entities (NMEs) and spending as well as employees



Discussion: what’s your take?



	Target-to-hit	Hit-to-lead	Lead optimization	Preclinical	Phase I	Phase II	Phase III	Submission to launch	Launch
p(TS)	80%	75%	85%	69%	54%	34%	70%	91%	
***WIP* needed for 1 launch**	**24.3**	**19.4**	**14.6**	**12.4**	**8.6**	**4.6**	**1.6**	**1.1**	**1**
Cost per WIP per Phase	$1	$2.5	$10	$5	$15	$40	$150	$40	
Cycle time (years)	1.0	1.5	2.0	1.0	1.5	2.5	2.5	1.5	
Cost per launch (out of pocket)	**$24**	**$49**	**$146**	**$62**	**$128**	**$185**	**$235**	**$44**	**$873**
% Total cost per NME	3%	6%	17%	7%	15%	21%	27%	5%	
Cost of capital	11%								
Cost per launch (capitalized)	**$94**	**$166**	**$414**	**$150**	**$273**	**$319**	**$314**	**$48**	**$1,778**

🟫 Discovery 🟦 Development

Learnings from numbers

1. Cost of target assessment and identification is not explicit.
2. Clinical studies are expensive, but picking a wrong target is twice as expensive.
3. It is probably wise to *infer* efficacy and safety profiles of drugs as accurately as possible.

Clinical activities of large pharma companies

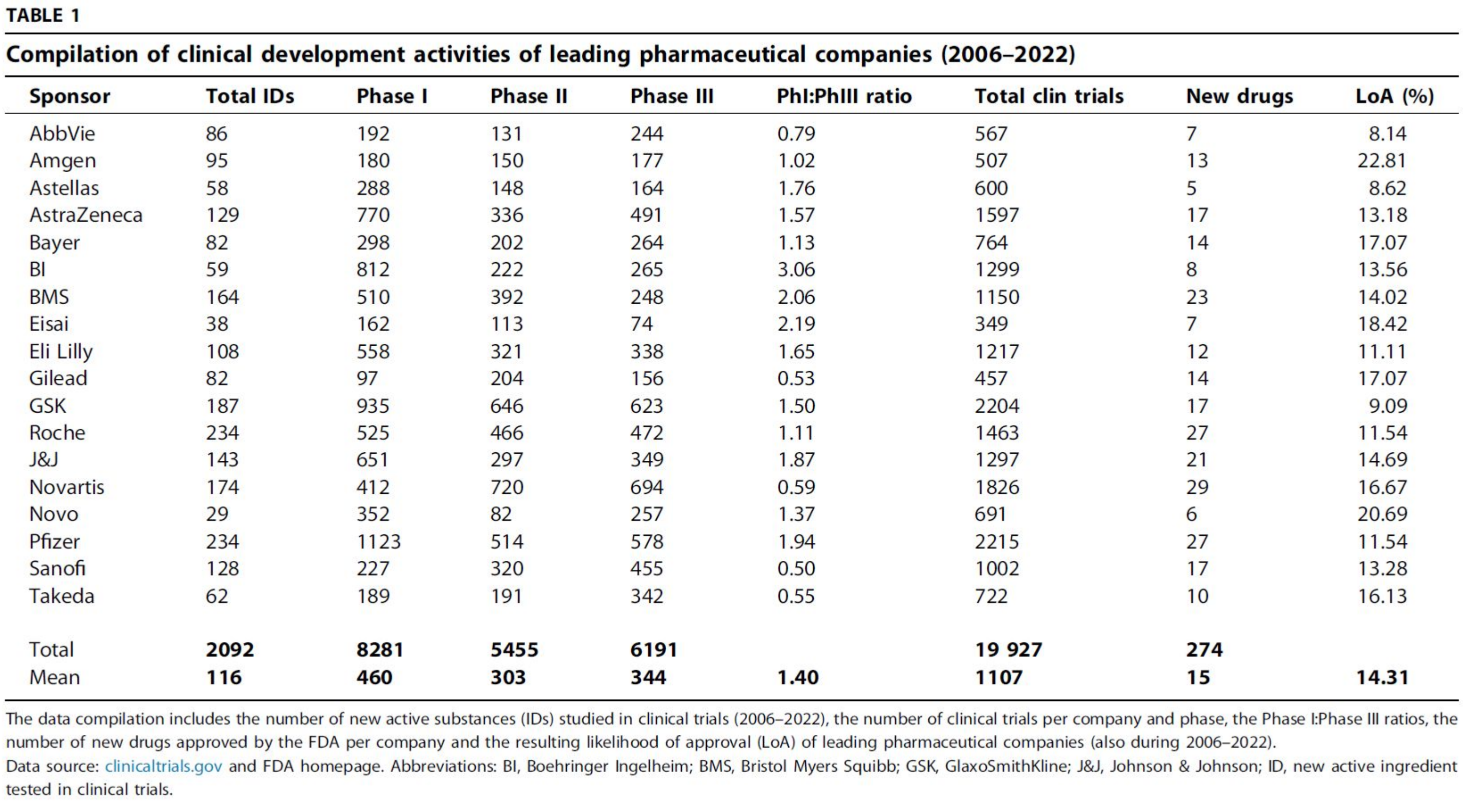


TABLE 1

Compilation of clinical development activities of leading pharmaceutical companies (2006–2022)

Sponsor	Total IDs	Phase I	Phase II	Phase III	PhI:PhIII ratio	Total clin trials	New drugs	LoA (%)
AbbVie	86	192	131	244	0.79	567	7	8.14
Amgen	95	180	150	177	1.02	507	13	22.81
Astellas	58	288	148	164	1.76	600	5	8.62
AstraZeneca	129	770	336	491	1.57	1597	17	13.18
Bayer	82	298	202	264	1.13	764	14	17.07
BI	59	812	222	265	3.06	1299	8	13.56
BMS	164	510	392	248	2.06	1150	23	14.02
Eisai	38	162	113	74	2.19	349	7	18.42
Eli Lilly	108	558	321	338	1.65	1217	12	11.11
Gilead	82	97	204	156	0.53	457	14	17.07
GSK	187	935	646	623	1.50	2204	17	9.09
Roche	234	525	466	472	1.11	1463	27	11.54
J&J	143	651	297	349	1.87	1297	21	14.69
Novartis	174	412	720	694	0.59	1826	29	16.67
Novo	29	352	82	257	1.37	691	6	20.69
Pfizer	234	1123	514	578	1.94	2215	27	11.54
Sanofi	128	227	320	455	0.50	1002	17	13.28
Takeda	62	189	191	342	0.55	722	10	16.13
Total	**2092**	**8281**	**5455**	**6191**		**19 927**	**274**	
Mean	**116**	**460**	**303**	**344**	**1.40**	**1107**	**15**	**14.31**

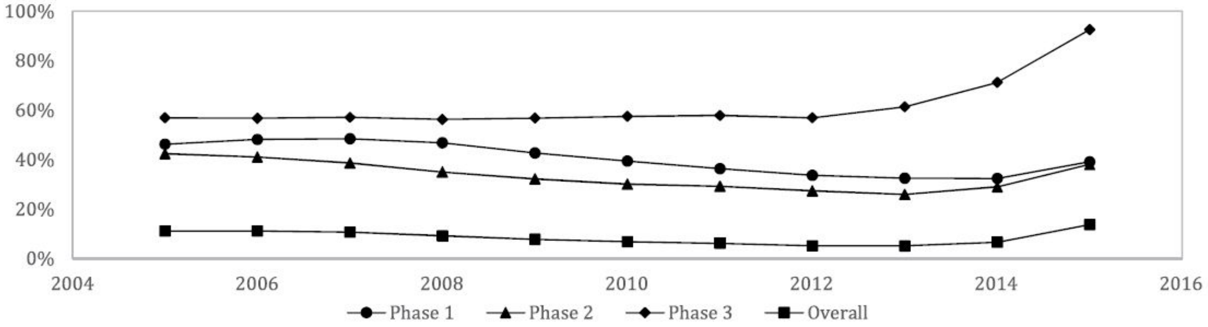
The data compilation includes the number of new active substances (IDs) studied in clinical trials (2006–2022), the number of clinical trials per company and phase, the Phase I:Phase III ratios, the number of new drugs approved by the FDA per company and the resulting likelihood of approval (LoA) of leading pharmaceutical companies (also during 2006–2022).

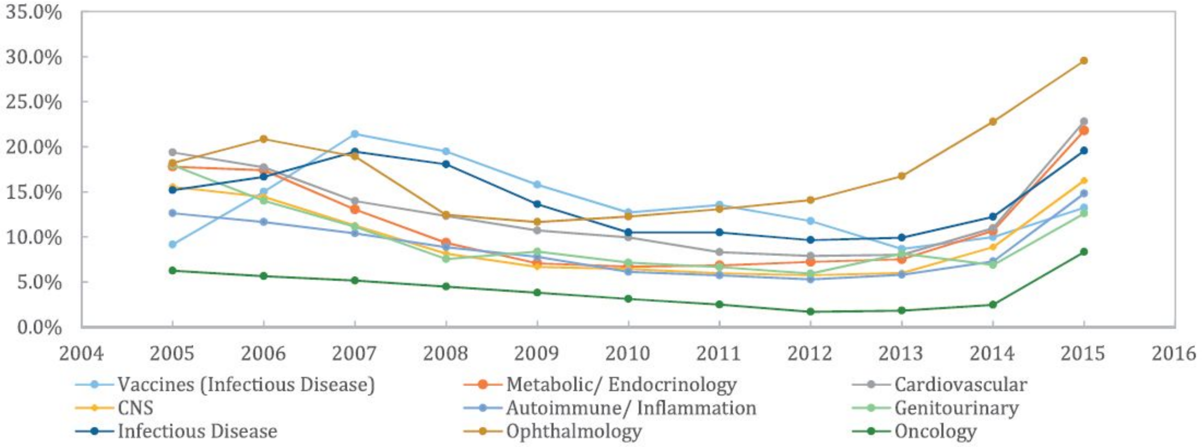
Data source: clinicaltrials.gov and FDA homepage. Abbreviations: BI, Boehringer Ingelheim; BMS, Bristol Myers Squibb; GSK, GlaxoSmithKline; J&J, Johnson & Johnson; ID, new active ingredient tested in clinical trials.

LoA: likelihood of approval, also known as the probability of technical and regulatory success (PTRS), covers from the first-in-human study to drug approval and marketing authorization by the FDA.

Schuhmacher, *et al.* 2025. “Benchmarking R&D Success Rates of Leading Pharmaceutical Companies: An Empirical Analysis of FDA Approvals (2006–2022).” Drug Discovery Today 30 (2): 104291. https://doi.org/10.1016/j.drudis.2025.104291.

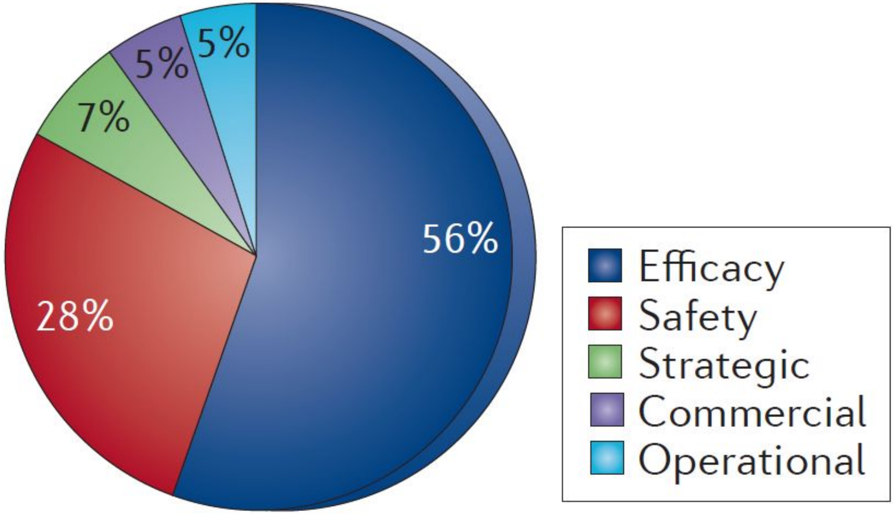
Clinical trial success rates by phase and indications

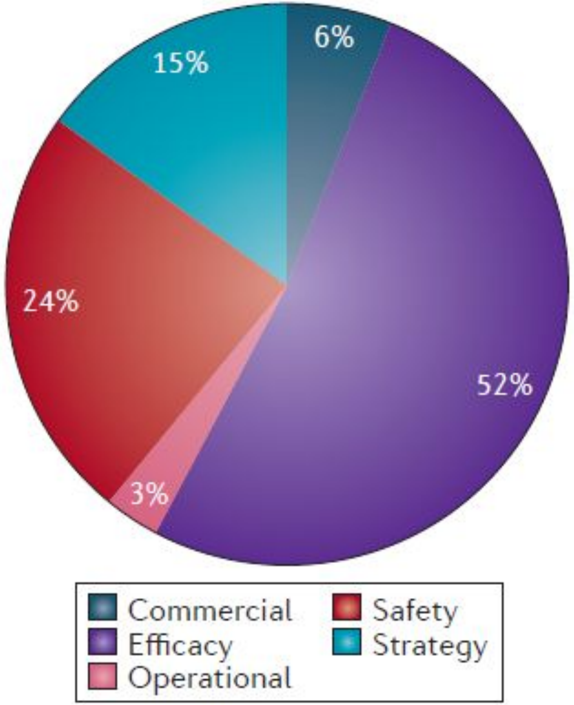




Wong, Chi Heem, Kien Wei Siah, and Andrew W Lo. 2019. "Estimation of Clinical Trial Success Rates and Related Parameters." Biostatistics 20 (2): 273–86. https://doi.org/10.1093/biostatistics/kxx069.

Failure analysis of Phase II and III trials: 2011-2012 (l.) & 2013-2015 (r.)





Arrowsmith, John, and Philip Miller. “Phase II and Phase III Attrition Rates 2011–2012.” Nature Reviews Drug Discovery 12, no. 8 (August 1, 2013): 569–569. https://doi.org/10.1038/nrd4090; Harrison, Richard K. “Phase II and Phase III Failures: 2013–2015.” Nature Reviews Drug Discovery 15 (November 4, 2016): 817–18. https://doi.org/10.1038/nrd.2016.184.

Essential preclinical work towards the Target Product Profile (TPP)

Classification	Examples
Potency and efficacy	● Mechanism and mode of action ● In vivo efficacy
Preclinical pharmacology	● Pharmacokinetic/pharmacodynamic (PK/PD) relationship ● Human pharmacokinetics and dose prediction ● Metabolism, clearance, and target tissue distribution ● Drug-drug interactions: CYP (cytochrome P450) induction, inhibition (inc. time-dependent) ● Combination therapy
Preclinical safety	● Regulatory guidance about requirements and experiment design ● *In vivo* studies: single/repeat dose toxicity, metabolites ● *In vitro* studies: genotoxicity, cytotoxicity, off-target activity, hERG (human ether-a-go-go-related gene), BSEP (balt salt export bump)
Chemistry, manufacturing, and controls (CMC)	● GMP (Good Manufacturing Practice) grade synthesis, number of steps, manufacturing timelines ● Drug substance and drug product stability ● Formulation supporting the selected route of administration ● Global access
Diagnostics and biomarkers	Measurements that may stratify patients, predict pharmacokinetics, efficacy, safety, and potentially lead to dose adjustment, for example metabolites, protein abundance, DNA mutation, etc.

Adapted from Demarest *et al.* (2025)

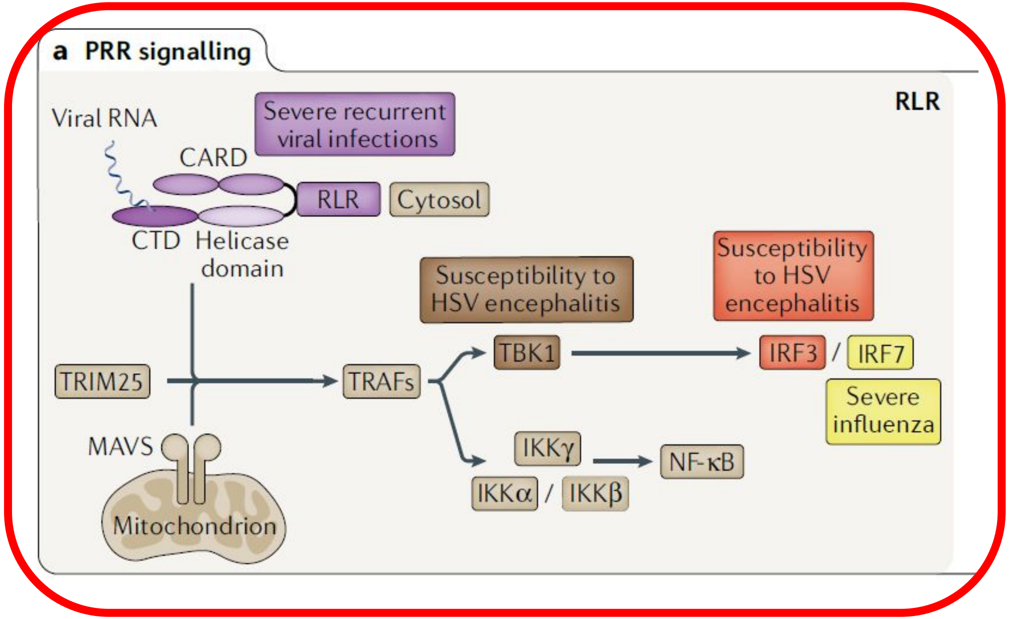
Part 2: A few propositions

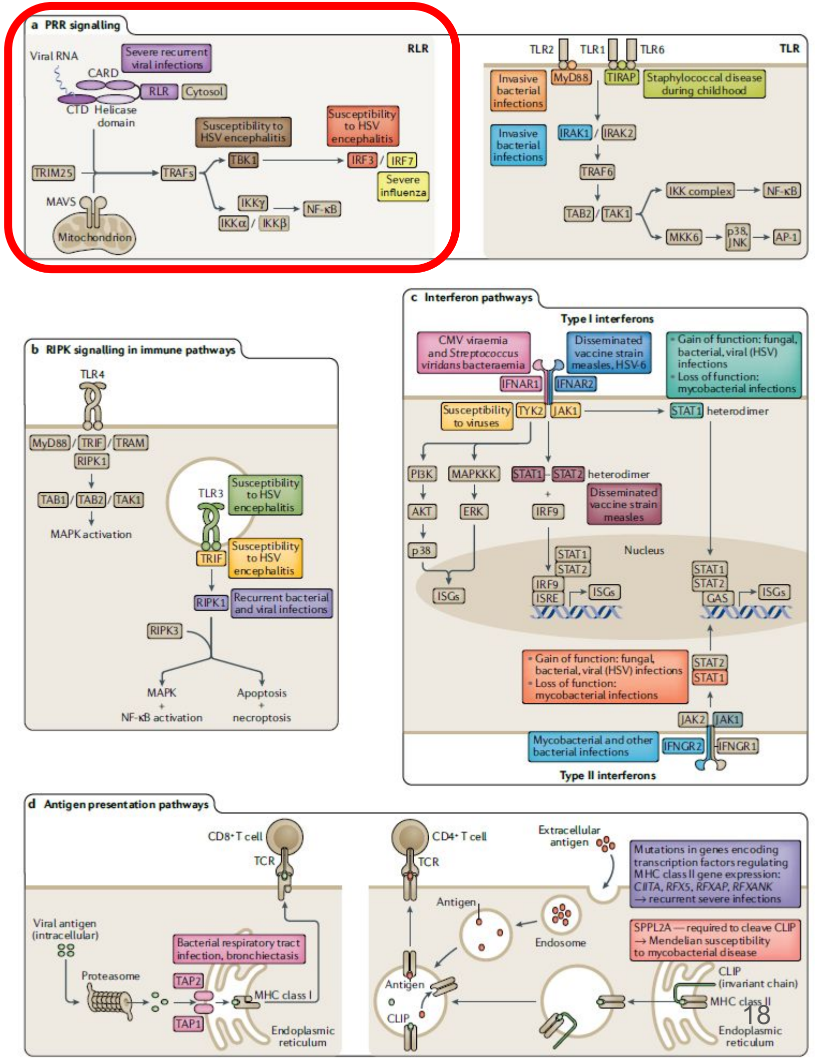
Propositions about the course

1. Human (disease) biology is a hierarchical complex adaptive system.
2. Drug discovery aims at identifying *new agents* that change the system's behaviour by interacting with, modifying, introducing, or removing biological entities with acceptable benefit and risk profiles.
3. We use mathematical and computational biology to study the system in order to predict and study the effect of modulation.

Complex Adaptive System

1. Parallel information channels
2. Conditional actions (if/then)
3. Modularity




a PRR signalling
RLR
Viral RNA
Severe recurrent viral infections
CARD
RLR
Cytosol
CTD
Helicase domain
Susceptibility to HSV encephalitis
Susceptibility to HSV encephalitis
TRIM25
TRAFs
TBK1
IRF3
IRF7
Severe influenza
MAVS
IKKγ
NF-κB
IKKα
IKKβ
Mitochondrion
TLR2
TLR1
TLR6
TLR
Invasive bacterial infections
MyD88
TIRAP
Staphylococcal disease during childhood
Invasive bacterial infections
IRAK1
IRAK2
TRAF6
IKK complex
NF-κB
TAB2
TAK1
MKK6
p38, JNK
AP-1
b RIPK signalling in immune pathways
TLR4
MyD88
TRIF
TRAM
RIPK1
TAB1
TAB2
TAK1
MAPK activation
TLR3
Susceptibility to HSV encephalitis
TRIF
Susceptibility to HSV encephalitis
RIPK1
Recurrent bacterial and viral infections
RIPK3
MAPK + NF-κB activation
Apoptosis + necroptosis
c Interferon pathways
Type I interferons
CMV viraemia and Streptococcus viridans bacteraemia
Disseminated vaccine strain measles, HSV-6
• Gain of function: fungal, bacterial, viral (HSV) infections
• Loss of function: mycobacterial infections
IFNAR1
IFNAR2
Susceptibility to viruses
TYK2
JAK1
STAT1 heterodimer
PI3K
MAPKKK
STAT1
STAT2
heterodimer
Disseminated vaccine strain measles
AKT
ERK
IRF9
p38
Nucleus
STAT1
STAT2
IRF9
ISRE
ISGs
ISGs
STAT1
STAT2
GAS
ISGs
• Gain of function: fungal, bacterial, viral (HSV) infections
• Loss of function: mycobacterial infections
STAT2
STAT1
JAK2
JAK1
Mycobacterial and other bacterial infections
IFNGR2
IFNGR1
Type II interferons
d Antigen presentation pathways
CD8⁺ T cell
TCR
Viral antigen (intracellular)
Bacterial respiratory tract infection, bronchiectasis
Proteasome
TAP2
TAP1
MHC class I
Endoplasmic reticulum
CD4⁺ T cell
TCR
Extracellular antigen
Mutations in genes encoding transcription factors regulating MHC class II gene expression: CIITA, RFX5, RFXAP, RFXANK → recurrent severe infections
Antigen
Endosome
SPPL2A — required to cleave CLIP → Mendelian susceptibility to mycobacterial disease
Antigen
CLIP
CLIP (invariant chain)
MHC class II
Endoplasmic reticulum

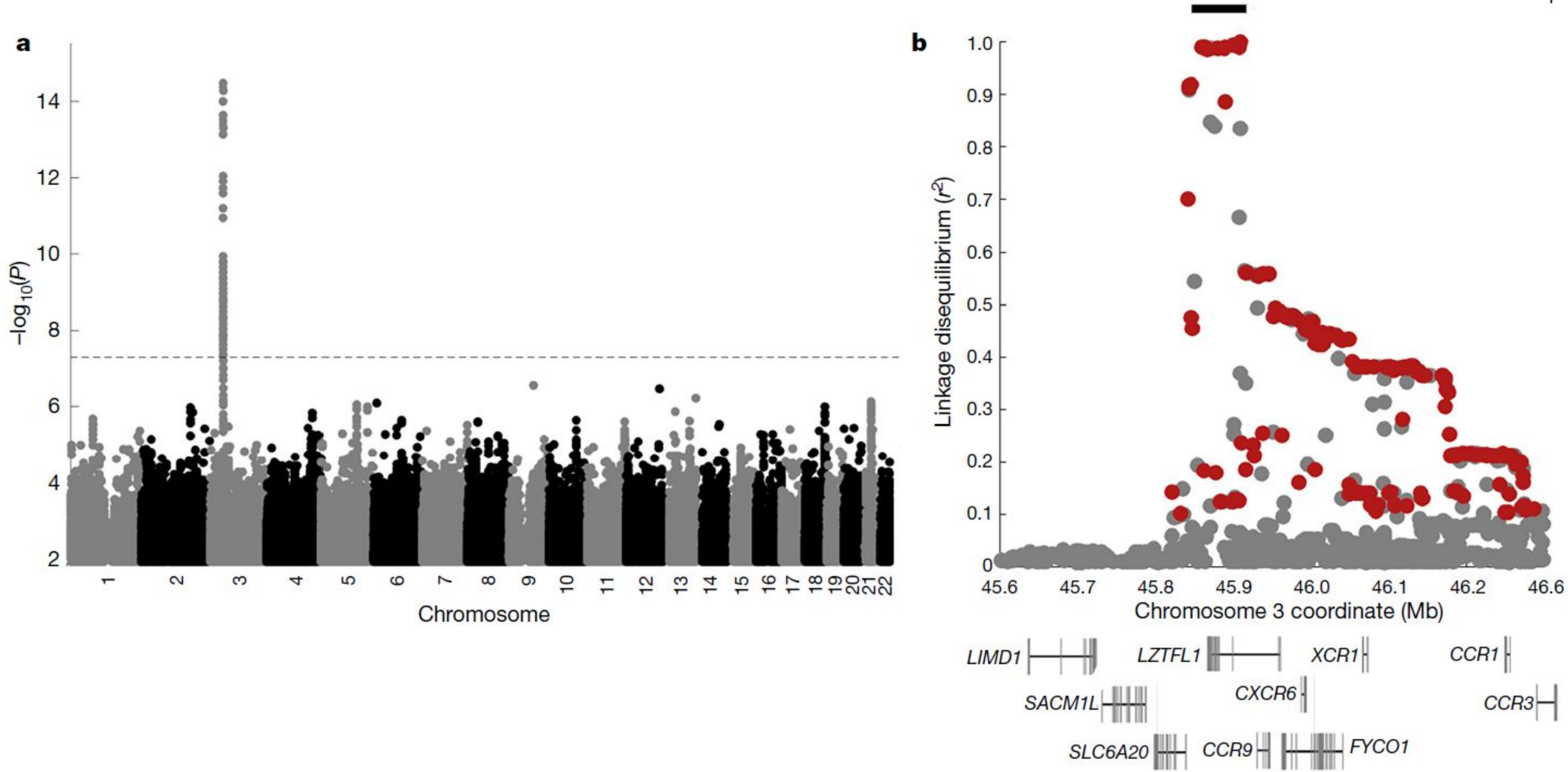
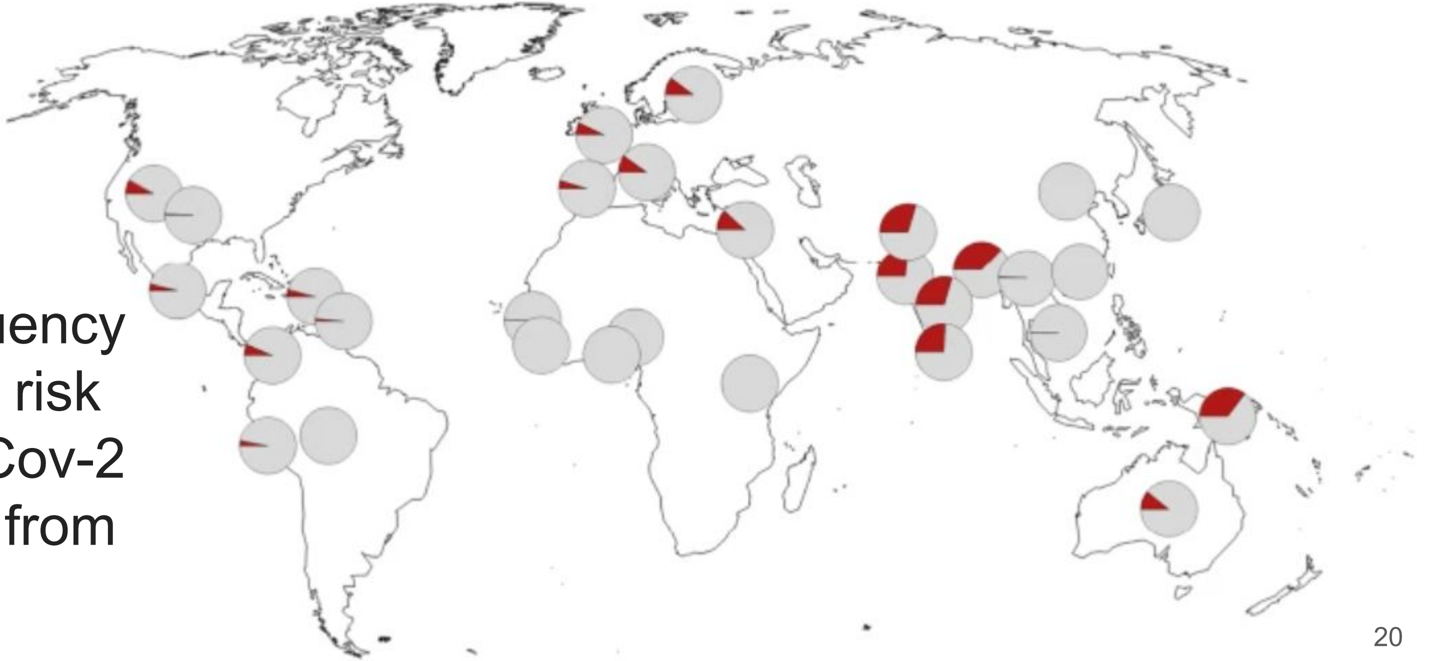


Fig. 1 | Genetic variants associated with severe COVID-19. a, Manhattan plot of a genome-wide association study of 3,199 hospitalized patients with COVID-19 and 897,488 population controls. The dashed line indicates genome-wide significance ($P = 5 \times 10^{-8}$). Data were modified from the COVID-19 Host Genetics Initiative[2] (https://www.covid19hg.org/). **b**, Linkage disequilibrium between the index risk variant (rs35044562) and genetic variants in the 1000 Genomes Project. Red circles indicate genetic variants for which the alleles are correlated to the risk variant ($r^2 > 0.1$) and the risk alleles match the Vindija 33.19 Neanderthal genome. The core Neanderthal haplotype ($r^2 > 0.98$) is indicated by a black bar. Some individuals carry longer Neanderthal-like haplotypes. The location of the genes in the region are indicated below using standard gene symbols. The *x* axis shows hg19 coordinates.

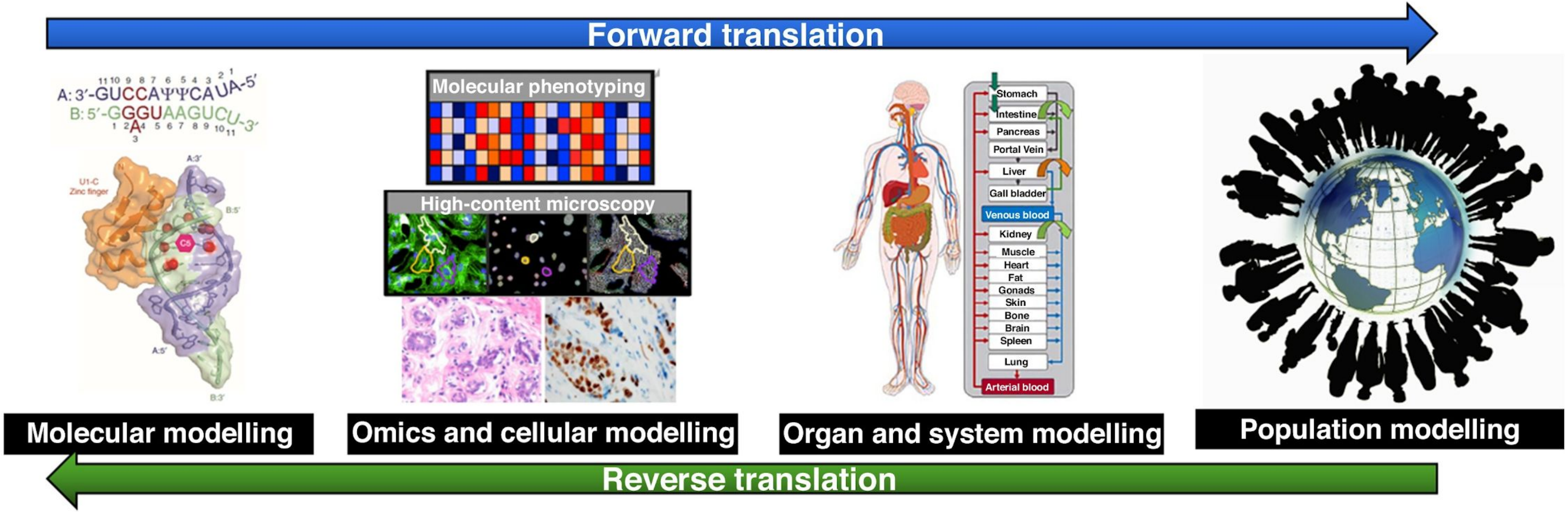
Complex Adaptive System

4. Adaptation and evolution



Minor allele frequency at rs35044562, a risk allele for SARS-Cov-2 that we inherited from Neanderthals.

A multiscale-modelling view of drug discovery



Drug Discovery Today

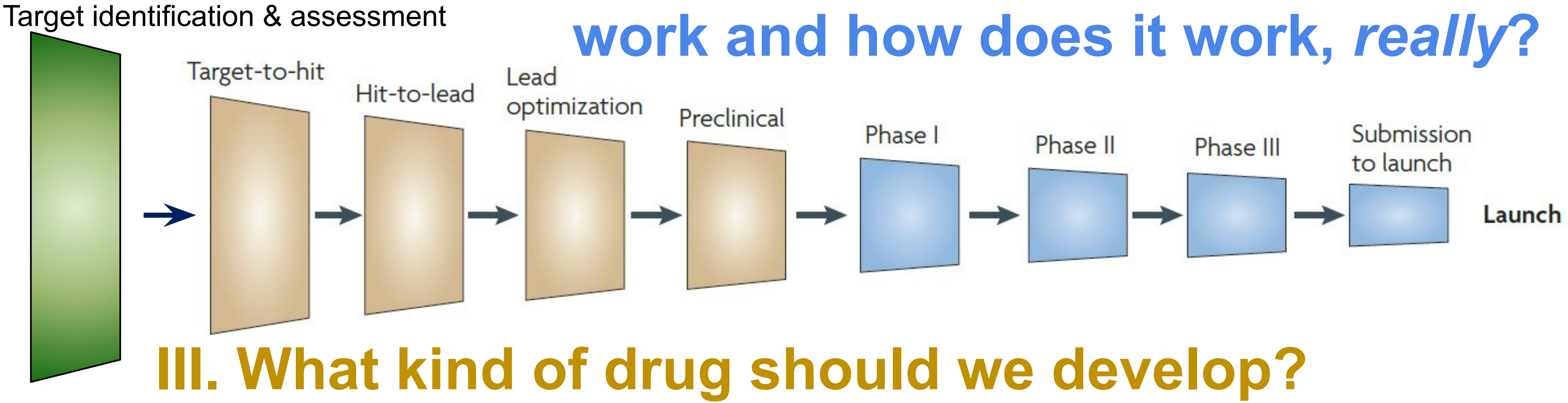
Drug Discovery	Biology	Math./Comp.
Target identification, assessment, and phenotypic screening	● Genomics ● Genetics ● Gene expression ● Chemical biology	● Statistical modelling ● Machine learning ● Mechanistic modelling
Drug modality and preclinical modelling	● RNA, antisense oligonucleotides, and antibodies ● Gene expression ● Network analysis	● Monte-Carlo methods ● Generative models ● Clustering
Biomarker, clinical modelling and reverse translation	● Population genetics ● Gene expression ● Pharmacokinetics and pharmacodynamics	● Causal analysis ● Machine learning ● Agent-based modelling

Common modelling approaches

- Statistical modelling and machine learning
- Causal inference
- Mechanistic modelling
    - ODEs (compartment models)
    - Agent-based models (particle models)
    - Networks (graphical and boolean models)

Questions that we will address in this course

V: For which patients will the drug work and how does it work, *really*?



III. What kind of drug should we develop?

IV. What efficacy and safety profiles can we expect?

I. What makes a good drug target?

II. What can we do if there are no good targets?

Take-home messages

- Drug discovery is a time-consuming and risky process with potentially high return of investment.
- Drug discovery aims at identifying agents interacting with or modifying biological molecules in order to modulate human disease biology, which is a hierarchical complex adaptive system.
- Mathematical and computational biology studies interactions and dynamics of the biological system, and it contributes to drug discovery at multiple stages.

Offline activities

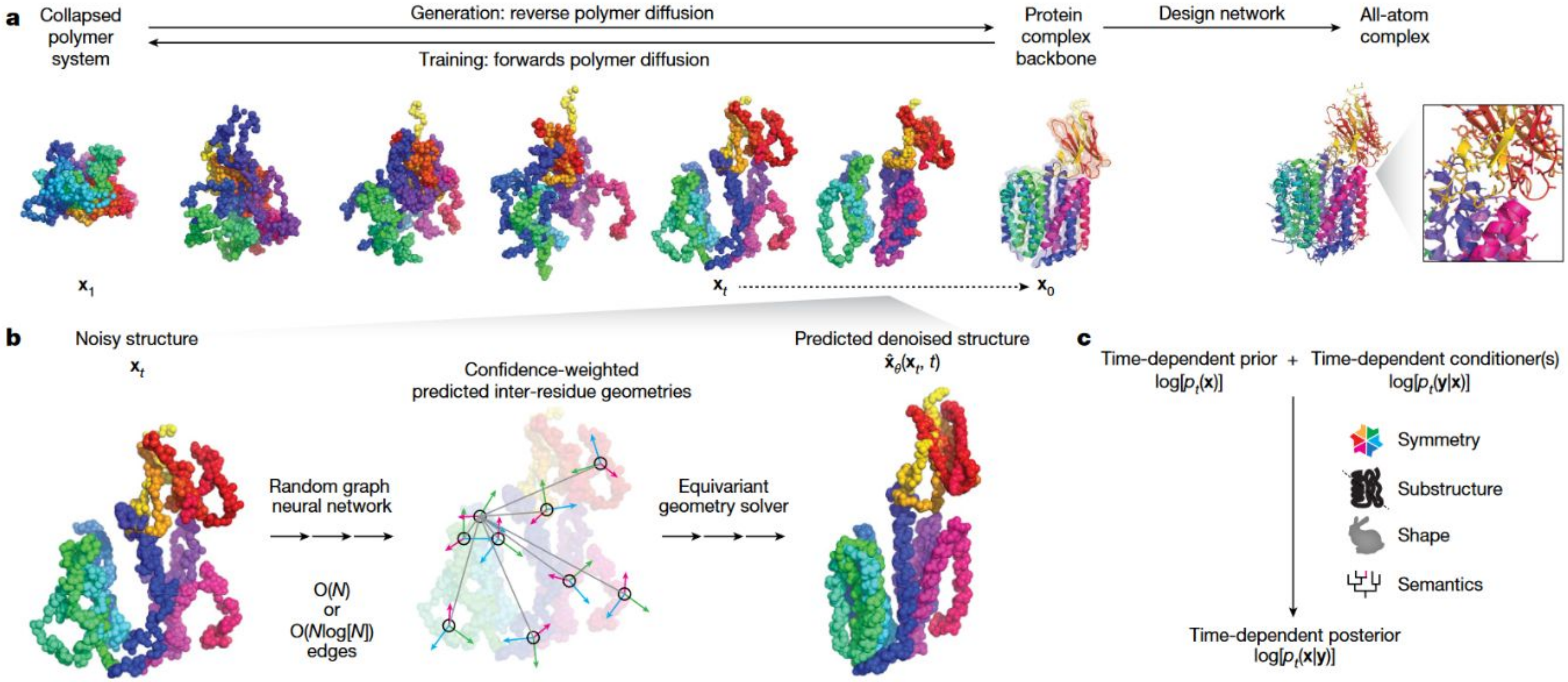
1. Fill the pre-course survey.
2. Read ‘How GLP-1 went from being a hard-to-handle hormone to a blockbuster success’ by Asher Mullar and Lotte Bjerre Knudsen (Nature Reviews Drug Discovery, 2024). What surprises you most?

References

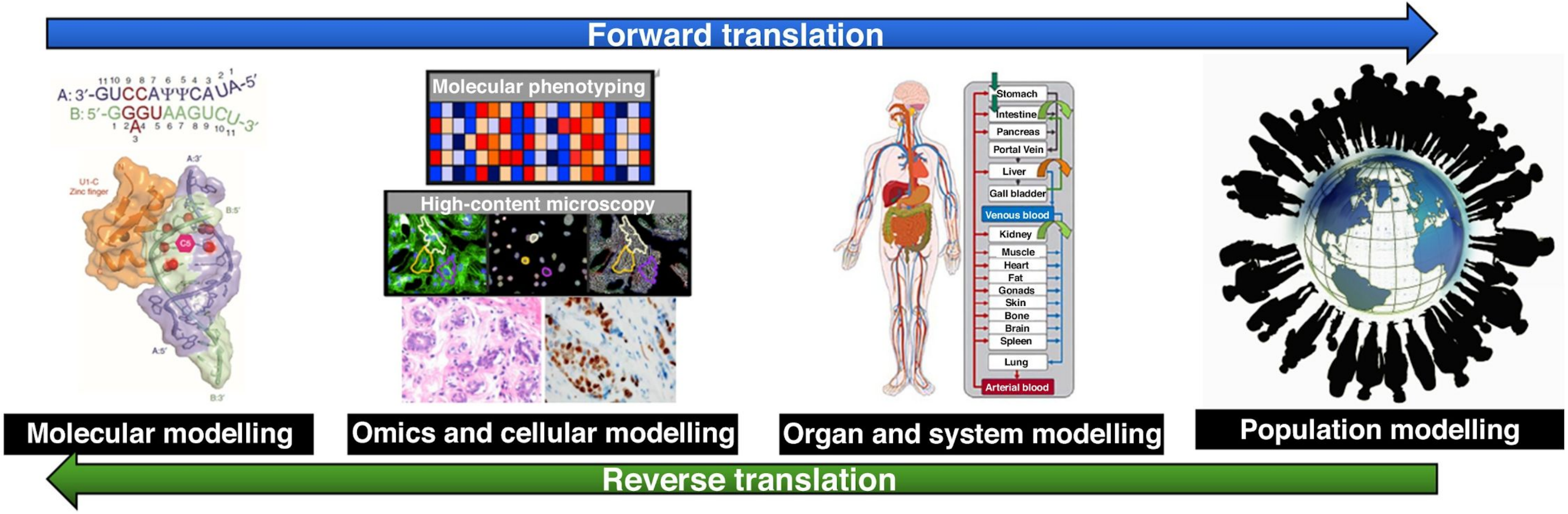
1. Paul *et al.* “How to Improve R&D Productivity: The Pharmaceutical Industry’s Grand Challenge.” Nature Reviews Drug Discovery, 2010.
2. Ringel, Michael S., Jack W. Scannell, Mathias Baedeker, and Ulrik Schulze. 2020. “Breaking Eroom’s Law.” Nature Reviews Drug Discovery 19 (12): 833–34. https://doi.org/10.1038/d41573-020-00059-3.
3. Morgan, Paul, Dean G. Brown, Simon Lennard, Mark J. Anderton, J. Carl Barrett, Ulf Eriksson, Mark Fidock, et al. 2018. “Impact of a Five-Dimensional Framework on R&amp;D Productivity at AstraZeneca.” Nature Reviews Drug Discovery 17 (3): 167–81. https://doi.org/10.1038/nrd.2017.244.
4. Harrison, Richard K. 2016. “Phase II and Phase III Failures: 2013–2015.” Nature Reviews Drug Discovery 15 (November): 817–18. https://doi.org/10.1038/nrd.2016.184.
5. Schuhmacher, Alexander, Lucas Wilisch, Michael Kuss, Andreas Kandelbauer, Markus Hinder, and Oliver Gassmann. “R&D Efficiency of Leading Pharmaceutical Companies – A 20-Year Analysis.” Drug Discovery Today 26, no. 8 (August 1, 2021): 1784–89. https://doi.org/10.1016/j.drudis.2021.05.005.
6. Schuhmacher, Alexander, Markus Hinder, Alexander von Stegmann und Stein, Dominik Hartl, and Oliver Gassmann. “Analysis of Pharma R&D Productivity – a New Perspective Needed.” Drug Discovery Today 28, no. 10 (October 1, 2023): 103726. https://doi.org/10.1016/j.drudis.2023.103726.
7. Zhang, Jitao David, Lisa Sach-Peltason, Christian Kramer, Ken Wang, and Martin Ebeling. 2020. “Multiscale Modelling of Drug Mechanism and Safety.” Drug Discovery Today 25 (3): 519–34. https://doi.org/10.1016/j.drudis.2019.12.009.
8. Holland, John H. 2006. “Studying Complex Adaptive Systems.” Journal of Systems Science and Complexity 19 (1): 1–8. https://doi.org/10.1007/s11424-006-0001-z.
9. Kwok, Andrew J., Alex Mentzer, and Julian C. Knight. 2021. “Host Genetics and Infectious Disease: New Tools, Insights and Translational Opportunities.” Nature Reviews Genetics 22 (3): 137–53. https://doi.org/10.1038/s41576-020-00297-6.
10. Zeberg, Hugo, and Svante Pääbo. 2020. “The Major Genetic Risk Factor for Severe COVID-19 Is Inherited from Neanderthals.” Nature 587 (7835): 610–12. https://doi.org/10.1038/s41586-020-2818-3.
11. Sturm, Gregor. 2020. “Hallmarks of Good Scientific Software”. https://grst.github.io/bioinformatics/2020/07/16/hallmarks-scientific-software.html
12. Ingraham, J. B. et al. Illuminating protein space with a programmable generative model. Nature 623, 1070–1078 (2023).
13. Demarest, James F., Ruxandra Draghia-Akli, Tomas Cihlar, Kenneth Bradley, John A. T. Young, Richa Chandra, Sujata Vaidyanathan, et al. “Antiviral Target Compound Profile for Pandemic Preparedness.” Nature Reviews Drug Discovery 24, no. 2 (February 2025): 151–52. https://doi.org/10.1038/s41573-024-01102-3.

Backup and License

Chroma: a generative model for proteins and protein complexes learning from evolution


a
Collapsed polymer system
Generation: reverse polymer diffusion
Training: forwards polymer diffusion
Protein complex backbone
Design network
All-atom complex
$\mathbf{x}_1$
$\mathbf{x}_t$
$\mathbf{x}_0$
b
Noisy structure
$\mathbf{x}_t$
Random graph neural network
O(N) or O(Nlog[N]) edges
Confidence-weighted predicted inter-residue geometries
Equivariant geometry solver
Predicted denoised structure
$\hat{\mathbf{x}}_\theta(\mathbf{x}_t, t)$
c
Time-dependent prior
$\log[p_t(\mathbf{x})]$
+
Time-dependent conditioner(s)
$\log[p_t(\mathbf{y}|\mathbf{x})]$
Symmetry
Substructure
Shape
Semantics
Time-dependent posterior
$\log[p_t(\mathbf{x}|\mathbf{y})]$

A multiscale-modelling view of drug discovery



Drug Discovery Today

Skipped in the class

Complementary views of biological systems

- Metabolism
- Energy
- Information machine
- Evolution
- Computing machine
- Network
- ...

Skipped in the class

An example of complementary views

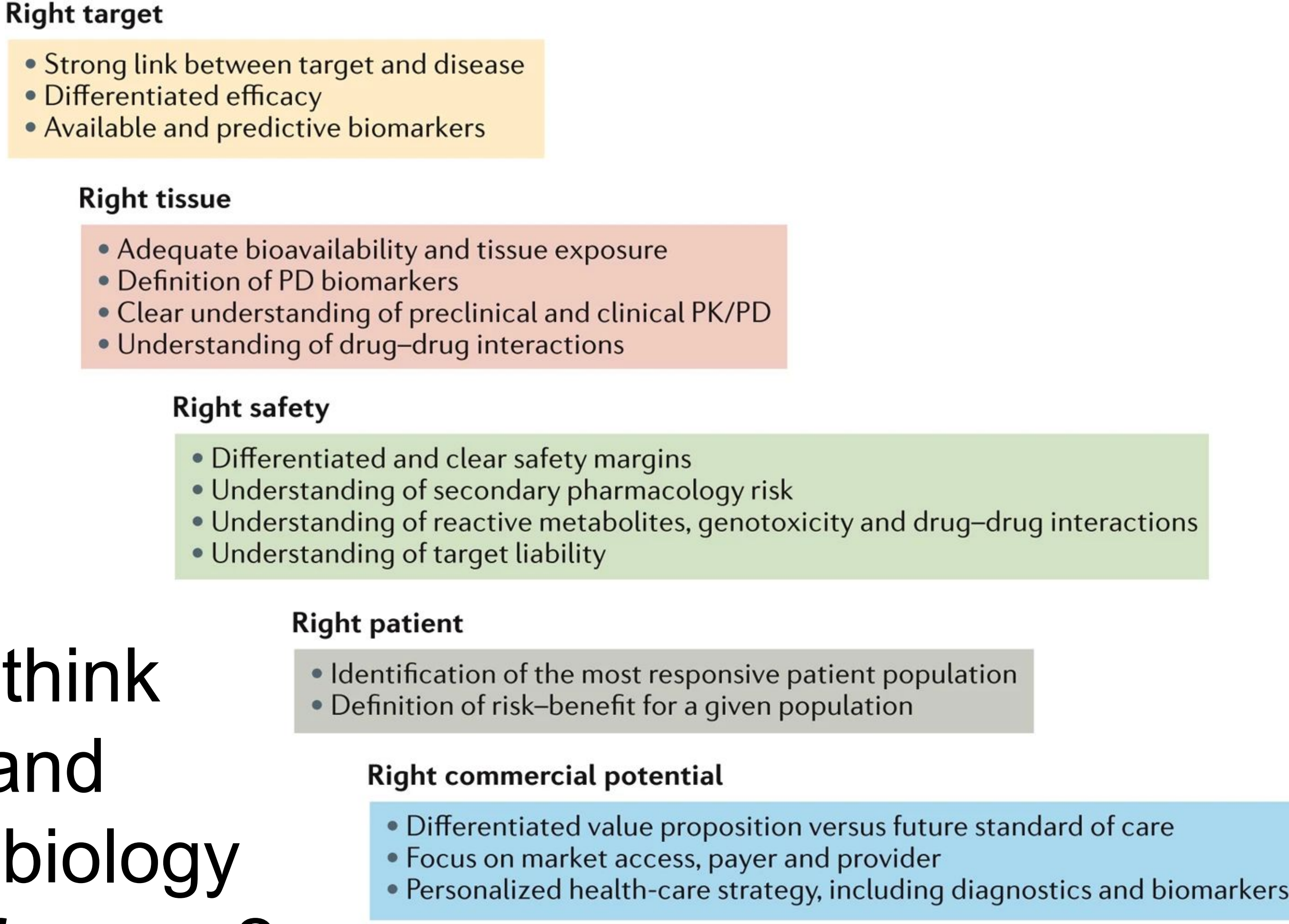
We want to work on hepatocarcinoma (liver cancer) and have the following information about a potential target X:

- X is a receptor expressing on the surface of most cell types;
- Upon binding ligands, X activates innate immune response;
- Gene sequence of X is conserved in primates but *not* in rodents;
- Protein X interacts with protein Y, which is essential, namely Y knockout causes lethal embryos;
- Asian population has a unique genetic variant in the non-coding region of *X*;

Discussion: what are the consequences of having these information?

Skipped in the class

Exercise



Right target

- Strong link between target and disease
- Differentiated efficacy
- Available and predictive biomarkers

Right tissue

- Adequate bioavailability and tissue exposure
- Definition of PD biomarkers
- Clear understanding of preclinical and clinical PK/PD
- Understanding of drug–drug interactions

Right safety

- Differentiated and clear safety margins
- Understanding of secondary pharmacology risk
- Understanding of reactive metabolites, genotoxicity and drug–drug interactions
- Understanding of target liability

Right patient

- Identification of the most responsive patient population
- Definition of risk–benefit for a given population

Right commercial potential

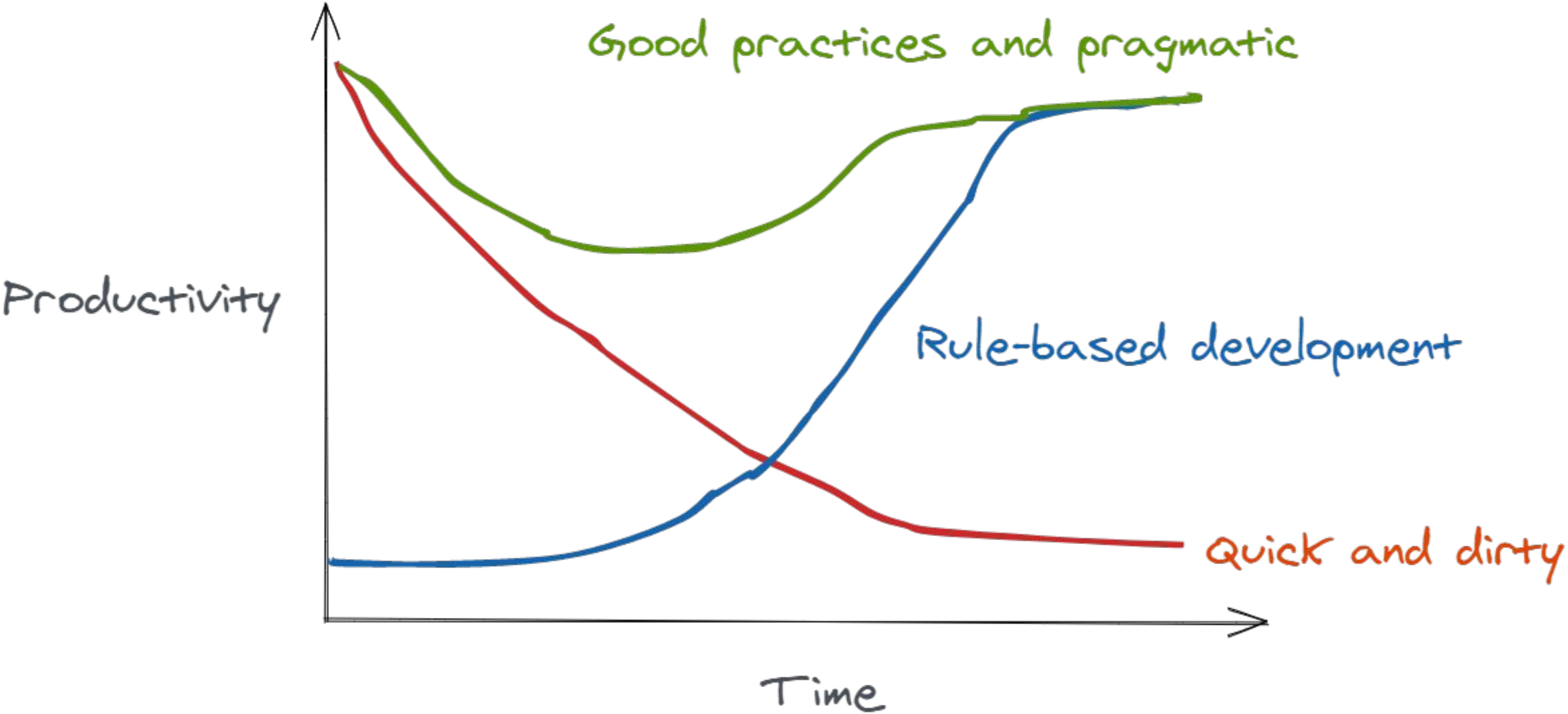
- Differentiated value proposition versus future standard of care
- Focus on market access, payer and provider
- Personalized health-care strategy, including diagnostics and biomarkers

Where do you think mathematical and computational biology will make a difference?

Nature Reviews | Drug Discovery

Nine steps toward reproducible research

1. Version control (*git*)
2. Don’t Repeat Yourself (DRY)
3. Keep It Simple, Stuipid (KISS)
4. Automatic testing (*pytest/Hypothesis*, *testthat*, *GitHub Actions*)
5. Documentation (*sphinx, pkdown*)
6. Dependency Management (*conda*, *packrat*)
7. Containerization (*Docker/Singularity*, *Bioconda/conda-forge*)
8. Pipelining (*Snakemake*, *NextFlow*, *drake*)
9. Self-reporting analysis (*Jupyter Notebook*, *Rmarkdown*)


Good practices and pragmatic
Productivity
Rule-based development
Quick and dirty
Time

Arguments for reproducible research

- Egoism and altruism
- *You will have to do it again*
- Sustainable long-term work

道

Tao, Path, or Way

术

Shu, Technique, or Art

Learn more about reproducible research

- The Missing Semester of Computer Science
- Software Carpentry (Unix Shell, Git, Python & R)
- Genomics Workshop of Data Carpentry
- *Clean Code* by Robert C. Martin
- Open-source tutorials of respective tools, such as sphinx, Snakemake, conda, or docker. Videos or podcasts work just as fine.

Contact the author