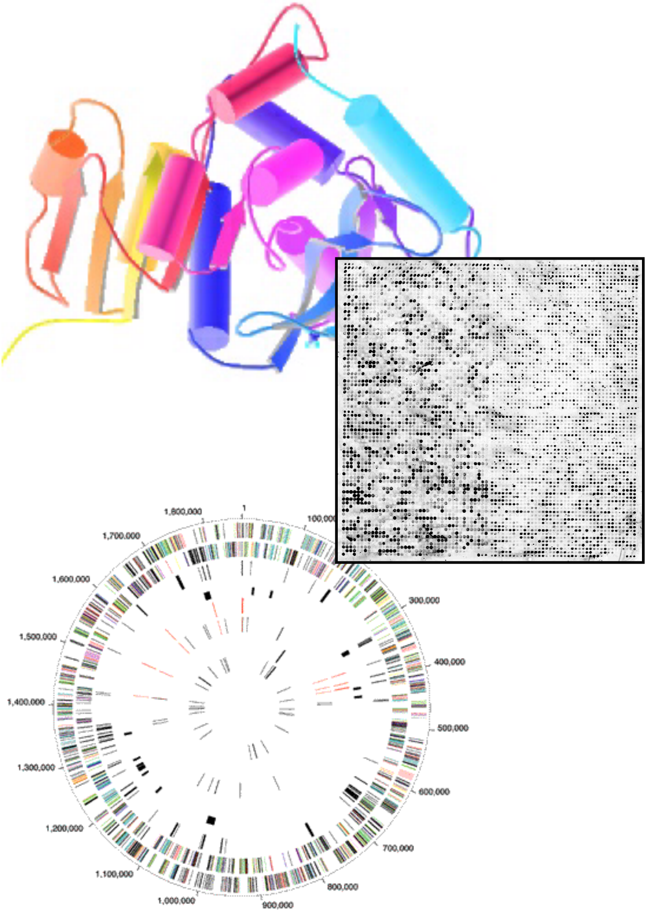
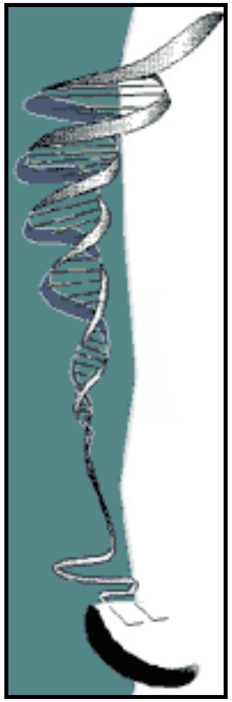
Biomed. Data Science:

Transition from Mining to Modeling





Mark Gerstein, Yale University

gersteinlab.org/courses/452

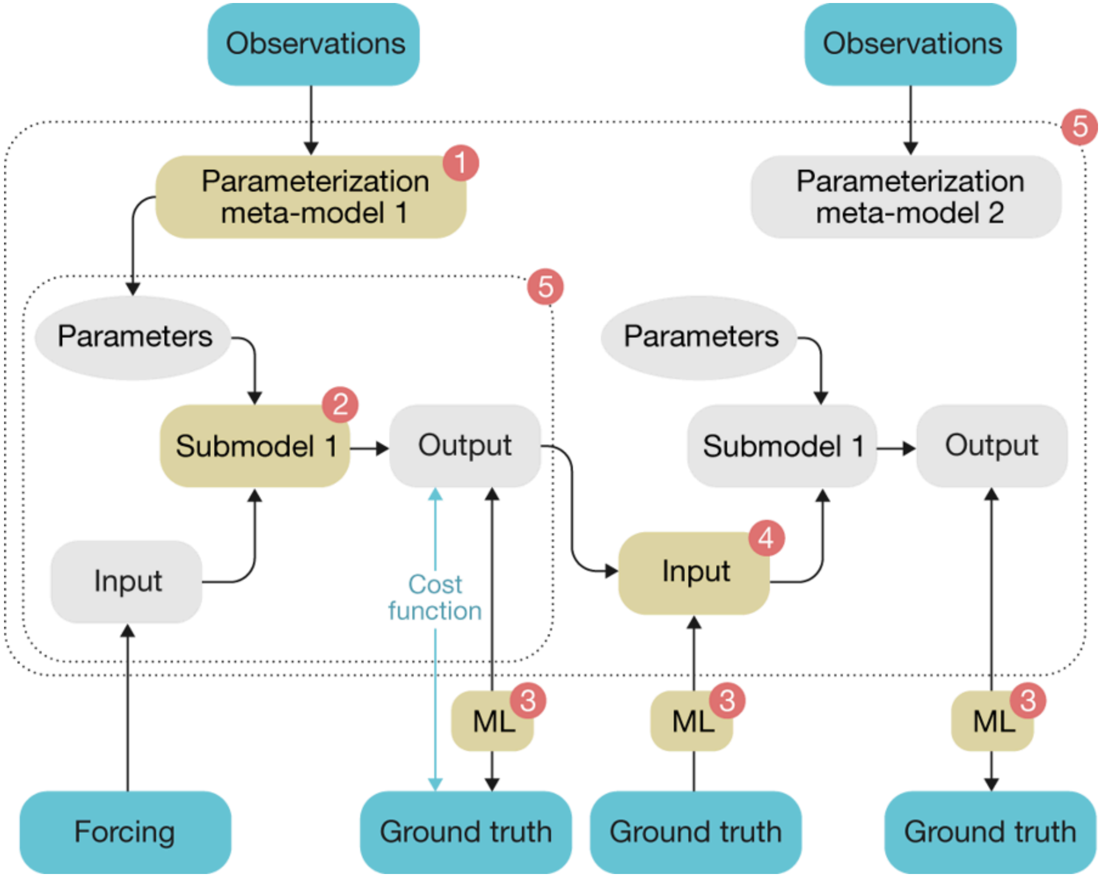
(last edit in spring ’19, pack #14)

Combining Mining & Modeling

- Complementarity of physical & ML approaches
  - "Physical approaches in principle being directly interpretable and offering the potential of extrapolation beyond observed conditions, whereas data-driven approaches are highly flexible in adapting to data"
- Hybrid #1: ML into physical
  - e.g. Emulation of specific parts of a physical for computational efficiency
  - More..
- Hybrid #2:
  Physical knowledge can be integrated into ML framework
  - Network architecture
  - Physical constraints in the cost function
  - Expansion of the training dataset for under sampled domains (ie physically based data augmentation) [More….]

https://www.nature.com/articles/s41586-019-0912-1
2019 Nature volume 566, pages195–204 (2019)

Hybrid #1: ML into physical models



(1) Improving parameterizations

(2) Replacing a 'physical' sub-model with a machine learning model

(3) Analysis of model–observation mismatch

(4) Constraining submodels

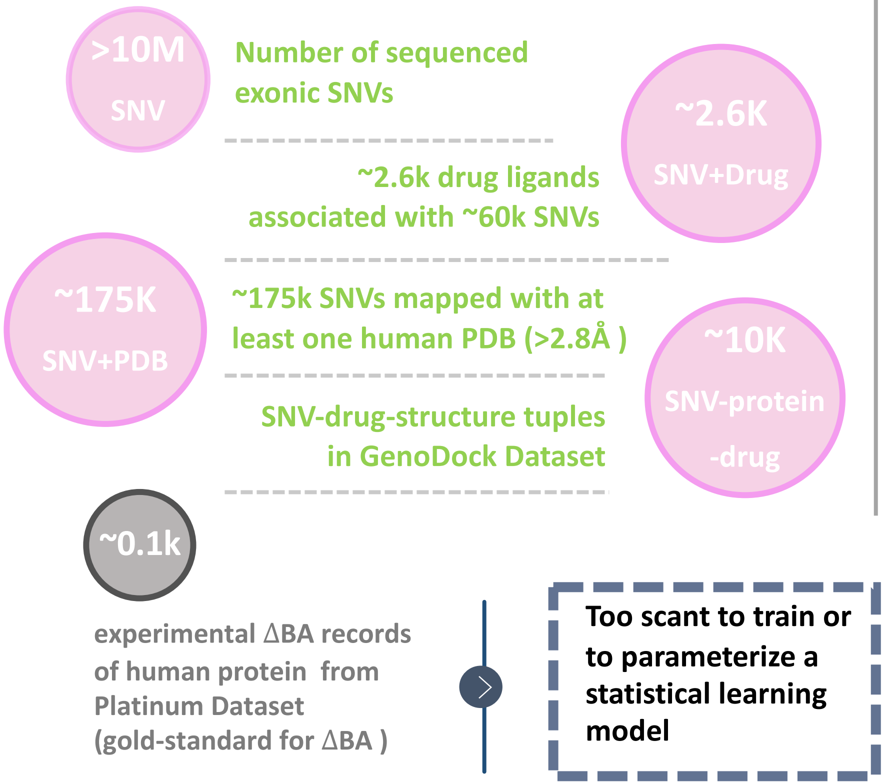
(5) Surrogate modelling or emulation

https://www.nature.com/articles/s41586-019-0912-1
2019 Nature volume 566, pages195–204 (2019)

Example of Hybrid #2: Integrating Physical Knowledge into Machine Learning

Physical Data Augmentation for Hybrid Physical-Statistical Model Construction

The Major Hurdle:
Highly Scant Ligand Binding Assay Data for ΔBA



Number of sequenced exonic SNVs

~2.6k drug ligands associated with ~60k SNVs

~175k SNVs mapped with at least one human PDB (>2.8Å)

SNV-drug-structure tuples in GenoDock Dataset

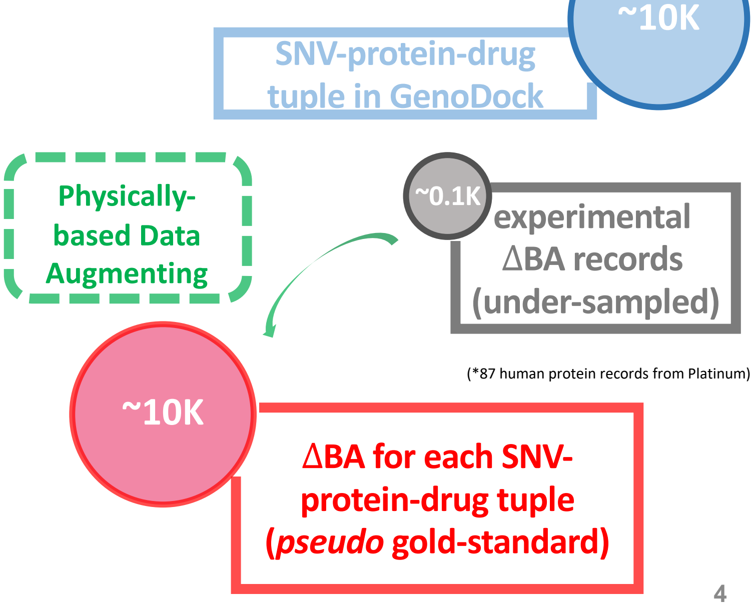
experimental ΔBA records of human protein from Platinum Dataset (gold-standard for ΔBA)

Too scant to train or to parameterize a statistical learning model

The Physically-based Data Augmentation Approach:
Leveraging Physical Calculations of ΔBA to Fill the Gap

(Reichstein *et al.*, **Nature**, 2019 & Xie et al., preprint, 2018)

- Expansion of the training dataset for under sampled domains
- Data augmentation is crucial to avoid overfitting



SNV-protein-drug tuple in GenoDock

Physically-based Data Augmenting

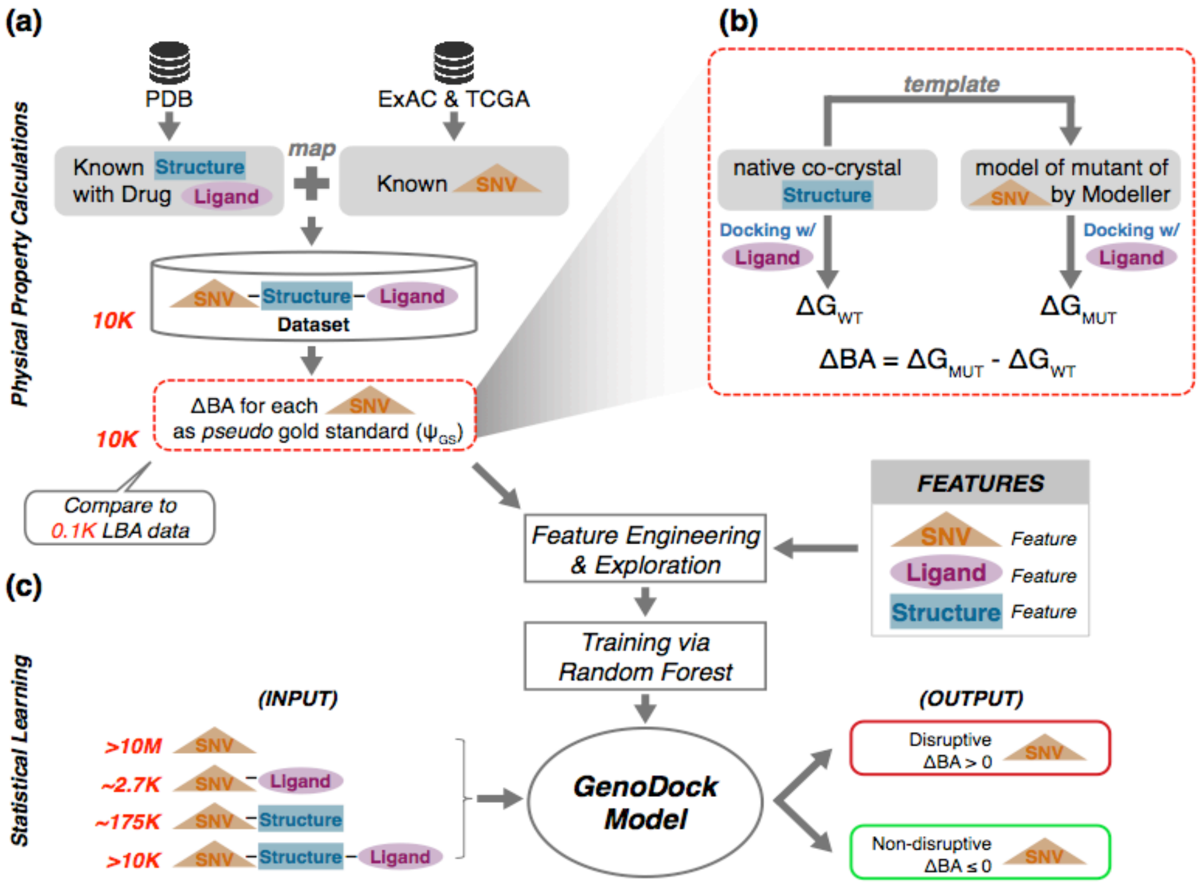
experimental ΔBA records (under-sampled)

(*87 human protein records from Platinum)

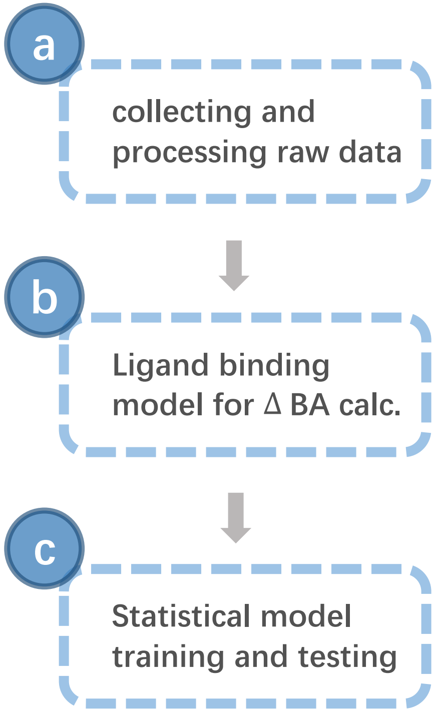
ΔBA for each SNV-protein-drug tuple (*pseudo* gold-standard)

[Wang et al. Structure (‘19, in press)]

Framework of the GenoDock Project - from Dataset Preparation to Model Construction



Wang *et al.*, accepted by *Structure*, 2019



a — collecting and processing raw data

b — Ligand binding model for Δ BA calc.

c — Statistical model training and testing

[Wang et al. Structure (‘19, in press)]