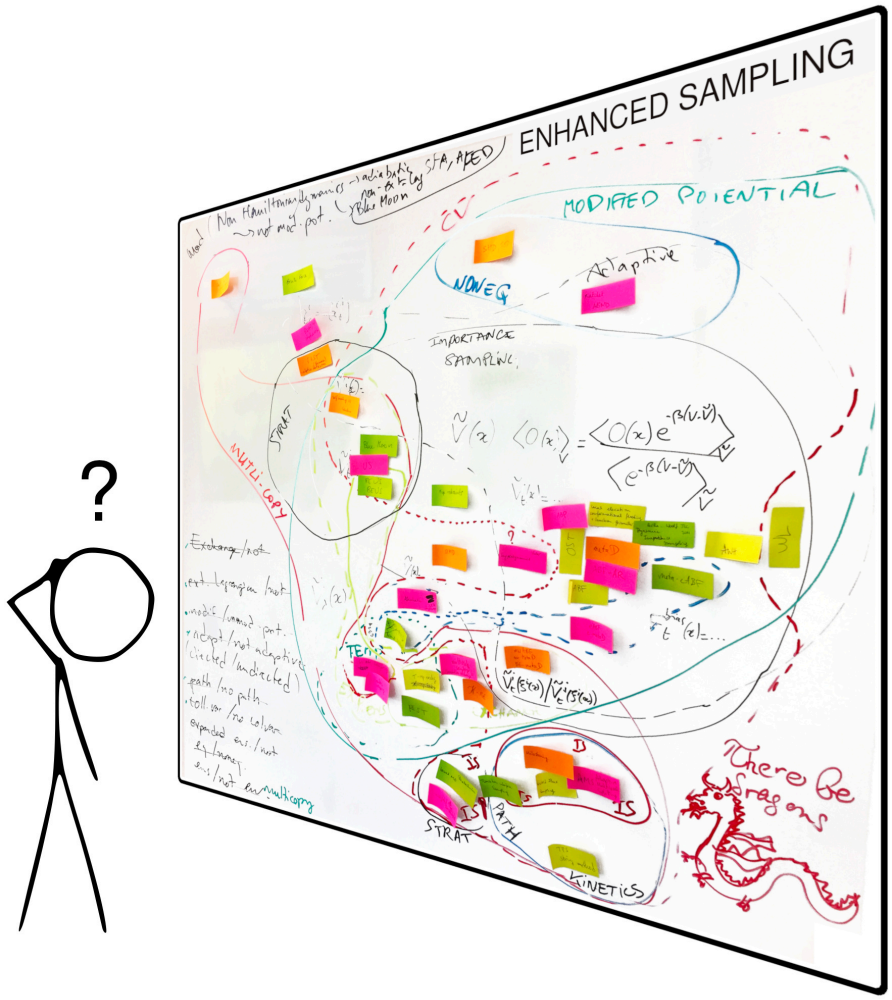
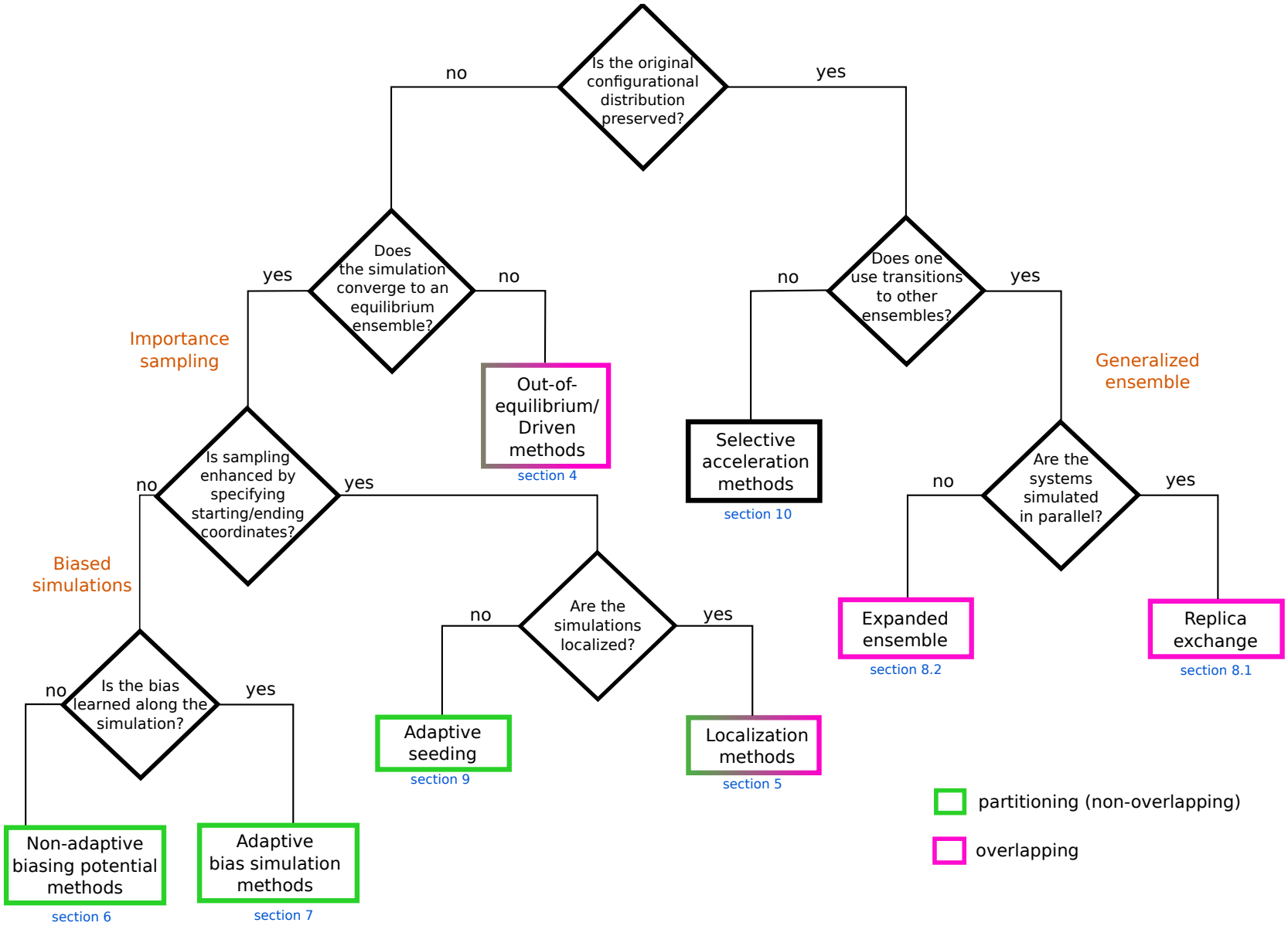
How does it all fit together?



Types of ingredients we have seen

- Thinking about probabilities instead of energies
  - Helps us to apply methods in PATH space as well as configuration space, and to remember free energies and normalizations constants
- Using Markov Chain Monte Carlo to find high-probability outcomes
- Using ensembles of simulations
- Biasing simulations to make them do what you want them to
  - Add an energy term, or start a path part to the goal
- Removing biases
  - Reweighting or other associated techniques

One overall classification: see LiveCoMS article



WARNING!! WARNING!!! DANGER!!

- Most simulations are wrong the first 2-3 times; the key is to identifying the errors quickly!

- How do I know if I've sampled enough?
- What if I choose bad directions to bias along?
- What are the uncertainties in my estimates?
- How do I know what method to use?

How do I know if I've sampled enough?

- You can't see . . . what you don't see . . .
- Enhanced sampling will give you better sampling along the degrees of freedom you FORCED it to sample better.
- It will NOT necessarily help you sample degrees of freedom that you did not force
- "orthogonal" degrees of freedom.
- You need to monitor those!

What if I chose the wrong path to bias along?

- It might not be the slow degree of freedom
  - Internal vs. external degrees of freedom
  - Pulling two proteins apart
    - Use some enhanced sampling in the direction of the COM of the protein.
    - However, perhaps the proteins need to rearrange in order to come apart.
  - Pulling them apart does not necessarily speed up the rearrangement
    - If you pull hard enough, you may be able to rip open a door without turning the handle. . . .
- Monitor other configurational degrees of freedom that are motivated by the problem
- See if there are physical reasons to expect other degrees of freedom (pdb/cryo-EM structures, etc)

What are the uncertainties in my method?

- Simple error propagation often doesn't work that well
  - Data is often not Gaussian
  - Data is almost always highly correlated
- As a minimum, you probably want to repeat with as independent a set of starting coordinates/paths as you can.
  - "The person who has one clock knows what time it is, the person with more than one clock never knows"
  - That's OK, these are all stochastic methods - there is a probability distribution of methods

Which method should I use?

- Don’t rely on any one method
  - All methods have strengths and weaknesses,
  - So don't take anything at face value!
- Easily availability of software may “fool” you into thinking that the methods are easy to use
  - But there are a lot of hyperparameters (settings) that can affect the results
  - They often need careful tuning to make them work right.
- Toy models are great to visualize what is going on
  - but they also could fool you into thinking a method is effective in more general problems

Simple things you can do to avoid (many types of) error

- Are my results consistent if I use different starting configurations or paths?
- Are my results consistent if I use different settings for the method?
- Are my results consistent if I use different methods?
- Do I get results that are clearly explainable in the limits?
- Do I get results that are consistent with physics?

Next steps

- We have MOST of the information up, and will work to finish posting it over the next few days.
- Look at the web page!

Q & A

- Questions about enhanced sampling methods!
- Questions about molecular modeling in research in general!