CREATIVE PROBABILISTIC PROGRAMMING FOR BIOLOGY

Miriam Shiffman

shiffman@{mit.edu
broadinstitute.org

"MEANINGFULNESS" of a learned representation in biology can only be measured w.r.t. a particular biological CONTEXT or question.

MODELING is the structure that provides this CONTEXT and endows latent representations with meaning.

PROBABILISTIC MODELING is often the best choice
- interpretability / decision theory
- coherent framework for hierarchies, noise
- biology itself is PROBABILISTIC !

PROBABILISTIC PROGRAMMING LANGUAGES are one tool missing from widespread adoption in biology, w/ potential to more naturally + holistically meld the modeling process w/ the process of wet lab science.

how can experimental biology be restructured around probabilistic modeling ?
(as an ongoing part of data collection & experimental design, beyond *post hoc* analysis)

how can PPLs be extended to meet the particular challenges of biology ?
(and promote model-tinkering in new & creative ways)

PROBABILISTIC PROGRAMMING LANGUAGES (PPLs)...

add `random variables` to the list of types we expect in a language: `str`, `int`, ... → fundamental operations in probability = fundamental (automated) features: `sample`, `condition`, `infer`

minimize edit distance b/w writing down the mathematical / coding up the executable model

promote EXPERIMENTATION & CREATIVITY in generative modeling
- complex architectures out of legolike abstractions
- tweak the model but not the algorithm
- concise, intuitive, human-readable

...just as differentiable languages have done for neural networks.

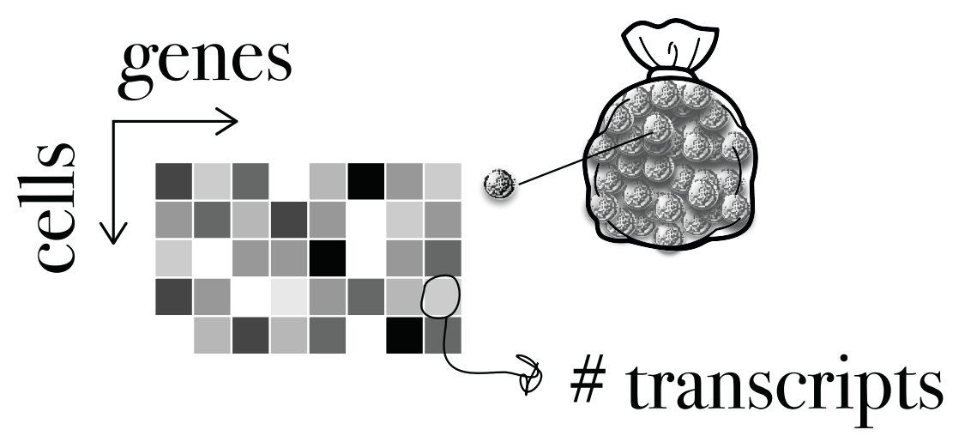
THE FUTURE...?

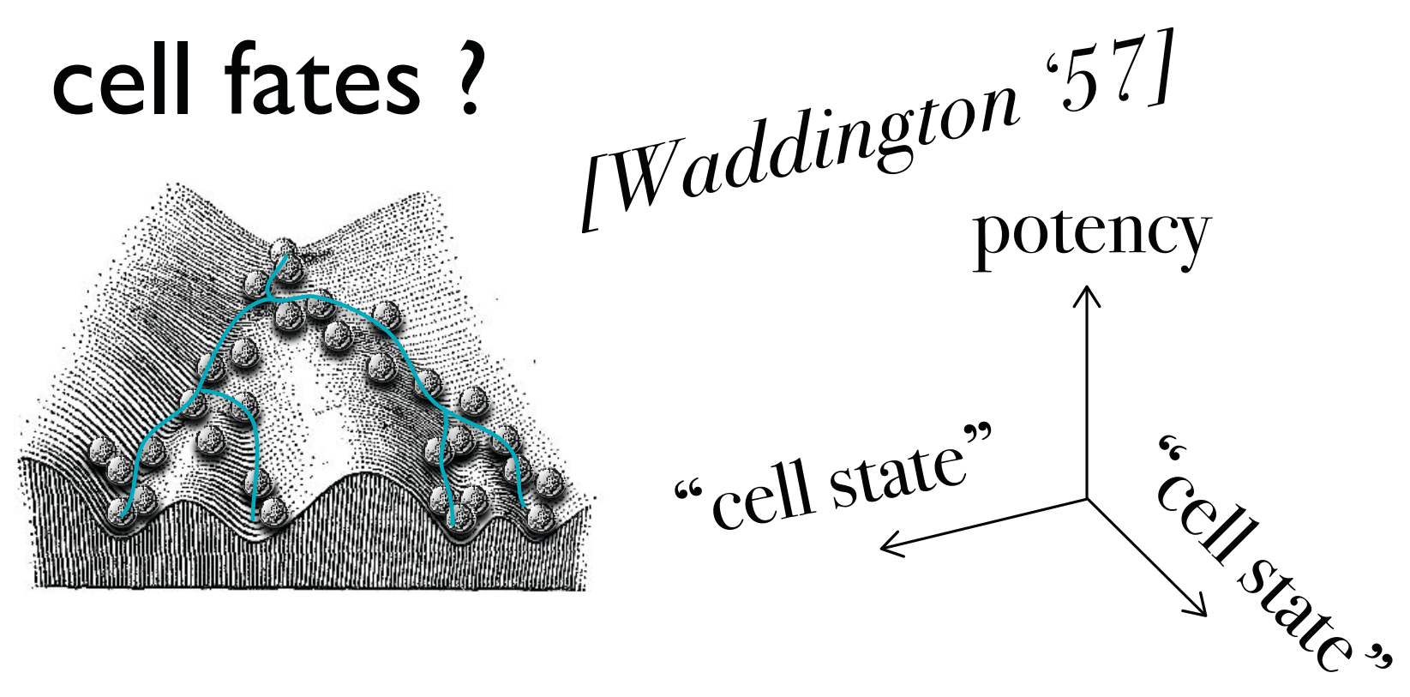
how to:
- visualize uncertainty for high-D, multimodal posteriors ?
- integrate PPLs more intimately into wet lab, like optimizing experimental protocols ?
- extend support for discrete structures LIKE TREES, a common regime in biology ? *(when posterior not diff'able w.r.t. its params, precluding VI & HMC)*

could:
- uncertainty quantification inform next gene to perturb or tissue to sequence ?
- probabilistic programs of biological processes be synthesized from data ?
- useful biological structures be encoded as PPL primitives ? *(Gene Ontology, KEGG pathways, genome coordinates, ...)*

PROBABILISTIC DIFFERENTIATION TREES, w/o PPLs...

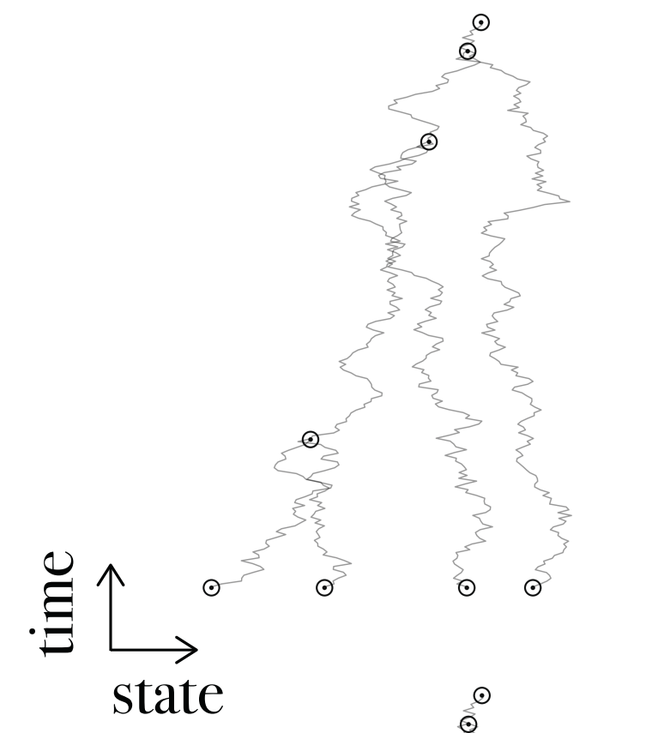
how do stem cells give rise to specialized / divergent / reproducible cell fates ?





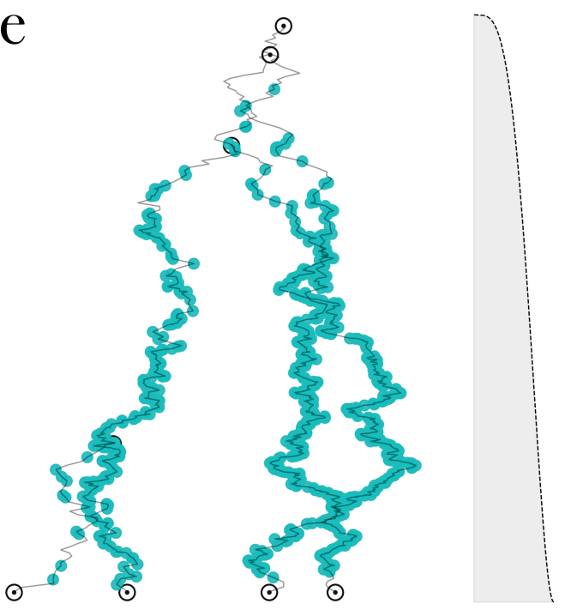
many static snapshots (scRNA-seq) → branching dynamics (continuous, probabilistic tree)

```
      #leaves ~ Poisson + 1
tree | #leaves ~ DirichletDiffusionTree
```

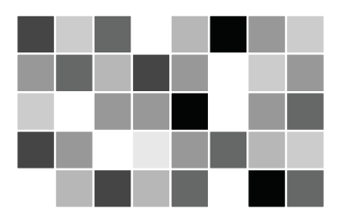


cell positions

```
                       times ~ Beta
        branches | tree,times ~ richGetRicher
states | branches,tree,times ~ GaussianDiffusion
```



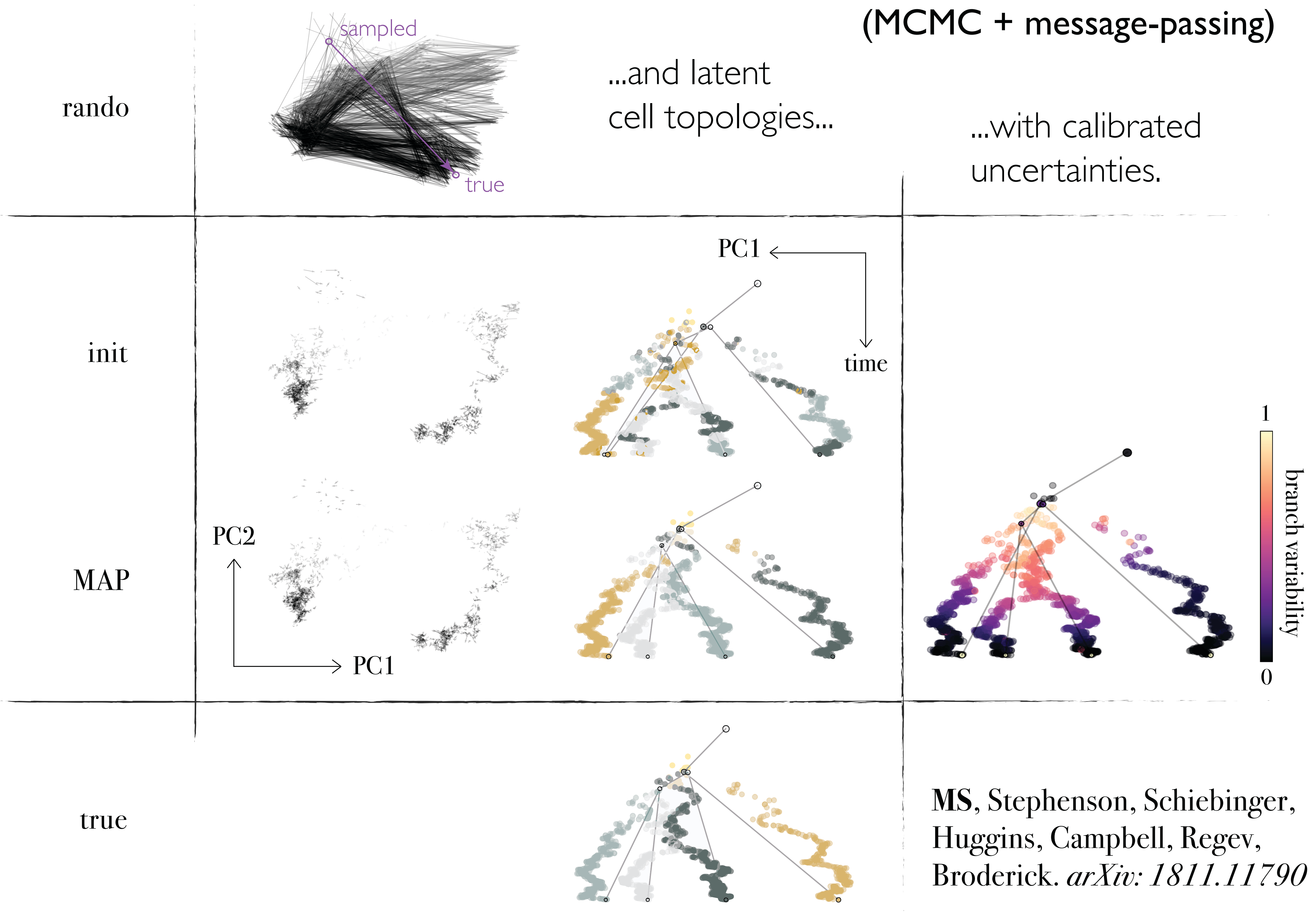
```
expressionProfiles | states ~ Binom(N_UMI, σ(states))
```



① *simulate:* 2000 cells, 10 genes; fixed topology / times ("time course")

recovery of latent cell states...

② *infer:* posterior over trees (MCMC + message-passing)



...and latent cell topologies...

...with calibrated uncertainties.

MS, Stephenson, Schiebinger, Huggins, Campbell, Regev, Broderick. *arXiv: 1811.11790*