Validation of Integrative Models

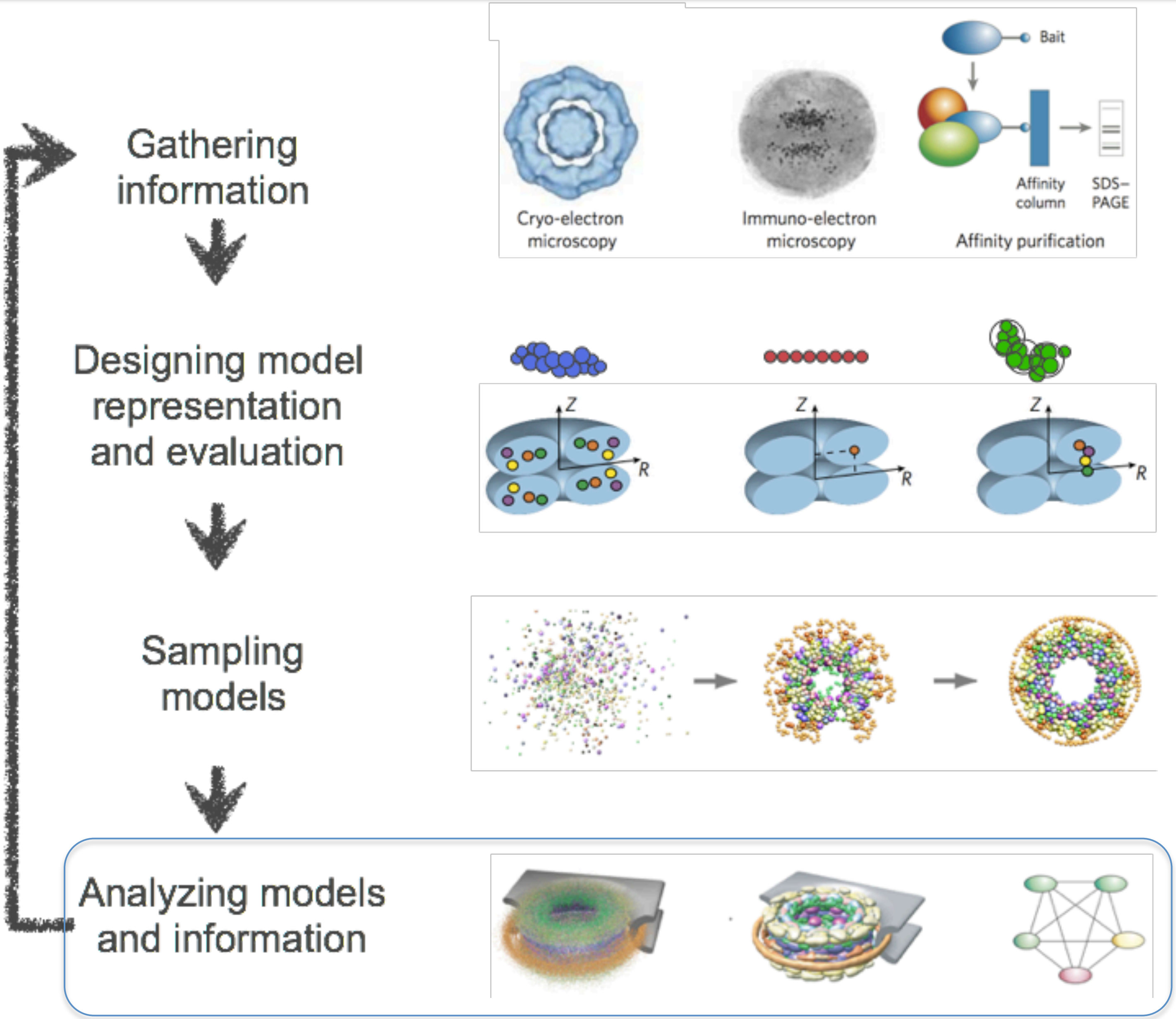
Advanced Analysis Methods

IMP workshop
Dec. 16, 2016

Daniel Saltzberg
saltzberg@salilab.org

Shruthi Viswanath
shruthi@salilab.org

Four Stages of Modeling



Gathering information

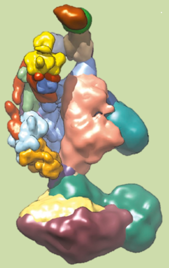
Designing model representation and evaluation

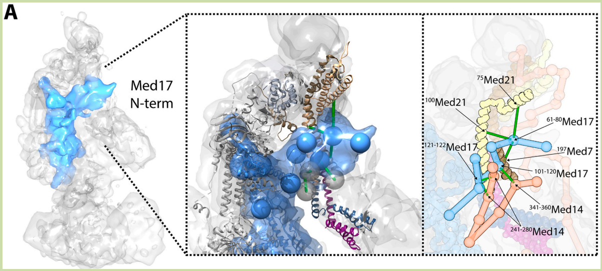
Sampling models

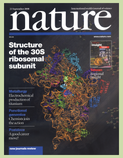
Analyzing models and information

Outcomes of structural modeling

Many models are wrong.
Some models are useful. -Andrej Sali





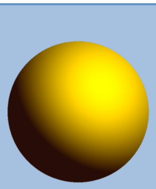


*

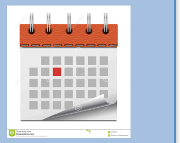
Useful Model!

Biological insight!

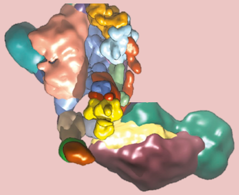
Robinson, Trnka et. al. 2015. eLife



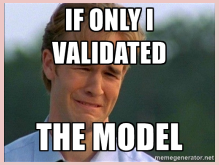
¯_(ツ)_/¯



Unuseful Model



Incorrect claims

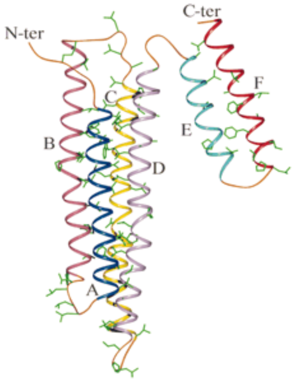


Wrong Model

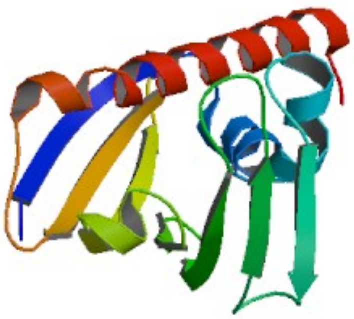
** The IMP developers make no guarantees of Nobel prizes based on use of the software*

Yes, there are bad models…

Fraud



Apolipoprotein A1
(2005)



Birch Pollen Allergen
(2010)

Mistakes

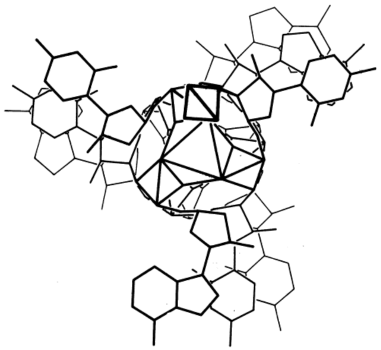
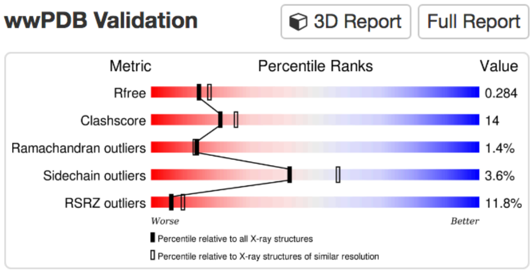


FIGURE 6
Plan of the nucleic acid structure, showing several nucleotide residues.

Nucleic Acid
(1953)

Validating / interpreting integrative models

- Methodology under development

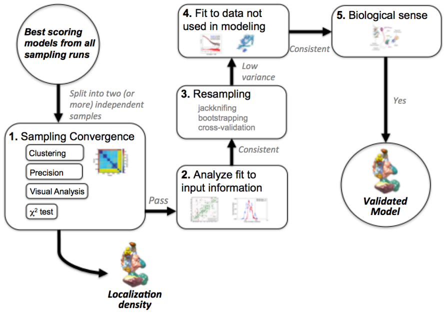


wwPDB Validation

Metric	Percentile Ranks	Value
Rfree		0.284
Clashscore		14
Ramachandran outliers		1.4%
Sidechain outliers		3.6%
RSRZ outliers		11.8%

▮ Percentile relative to all X-ray structures
▯ Percentile relative to X-ray structures of similar resolution

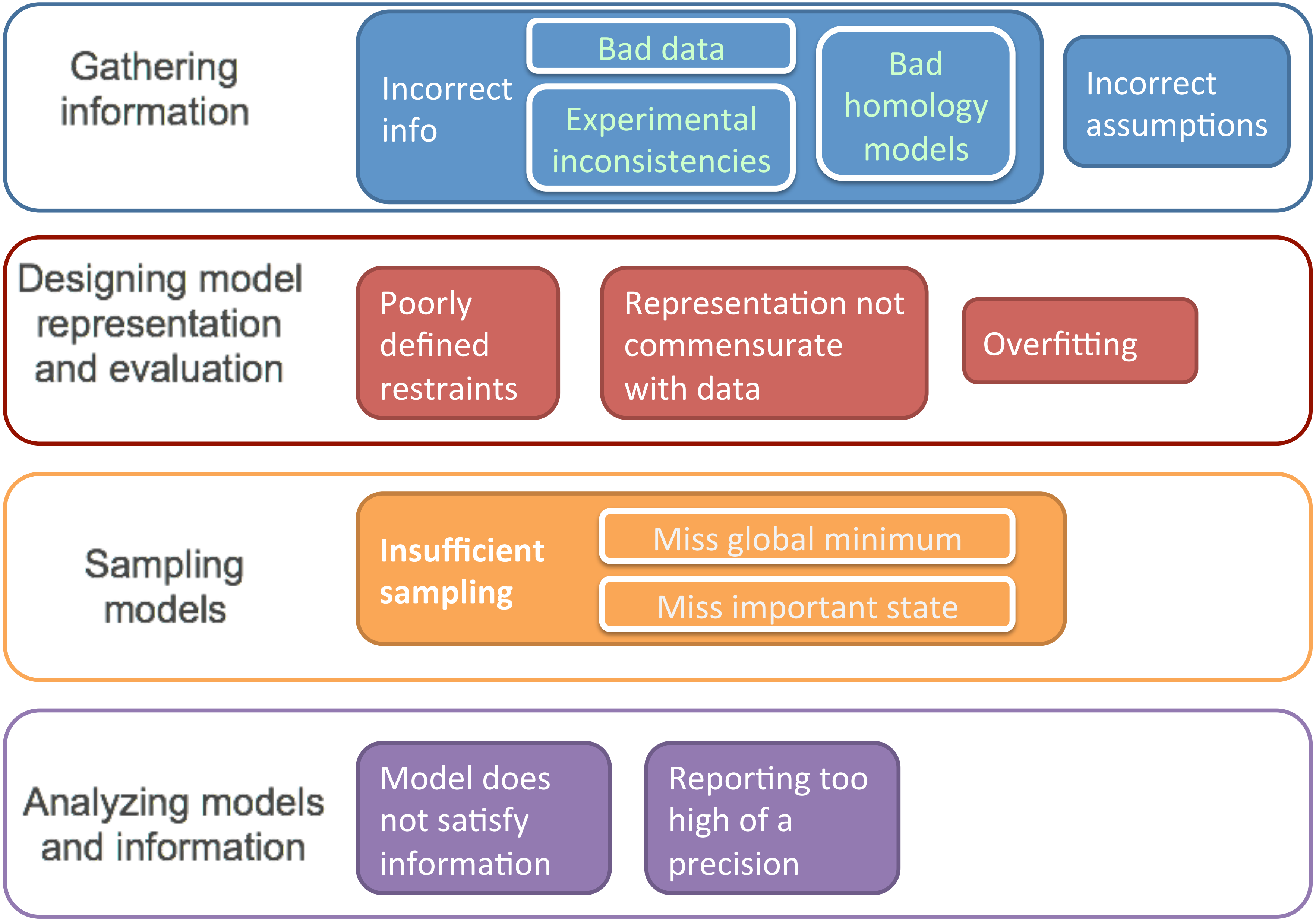
Crystallography validation protocols are fairly well estabilished



This workflow is current as of last week

- Complex analysis in IMP requires customized scripts
  - We’re developing pipelines to perform these methods

A subset of where can modeling go wrong



Gathering information

Incorrect info

Bad data

Experimental inconsistencies

Bad homology models

Incorrect assumptions

Designing model representation and evaluation

Poorly defined restraints

Representation not commensurate with data

Overfitting

Sampling models

Insufficient sampling

Miss global minimum

Miss important state

Analyzing models and information

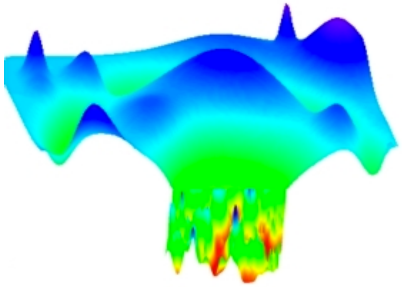
Model does not satisfy information

Reporting too high of a precision

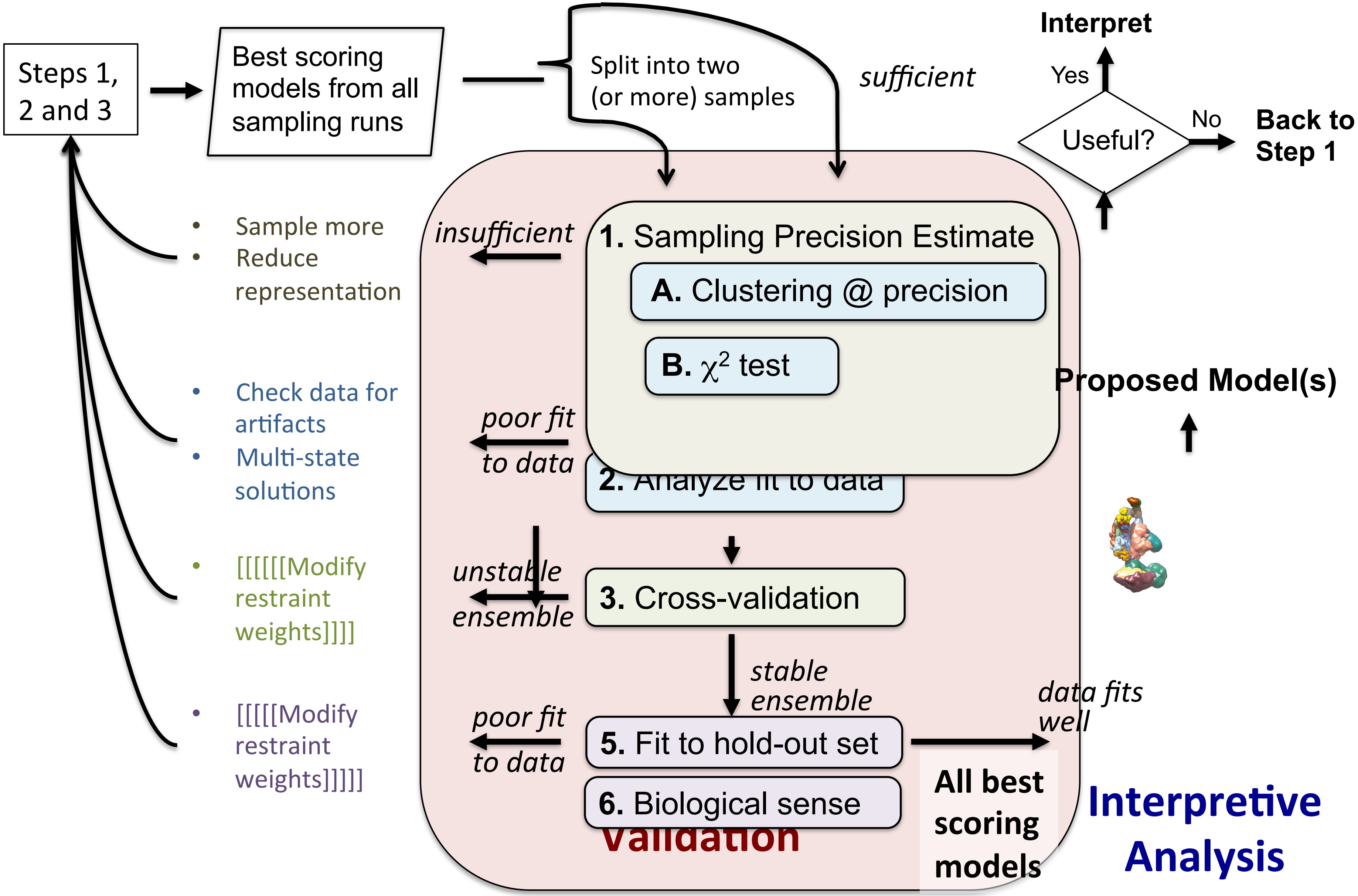
What to validate?

- Sampling Exhaustiveness
  - Possible sampling missed a subset of good scoring models
- Fit to Data/Restraints
  - Poorly fit data may indicate problem with data/modeling
- Jackknifing
  - Guard against overfitting
  - Complete cross-validation
  - Like a composite omit map
- Validation against other data

- How to proceed:
  - All models



Step 4: Practical Analysis Flowchart



Steps 1, 2 and 3

Best scoring models from all sampling runs

Split into two (or more) samples

sufficient

Interpret

Yes

Useful?

No

Back to Step 1

- Sample more
- Reduce representation

- Check data for artifacts
- Multi-state solutions

- [[[[[[Modify restraint weights]]]]

- [[[[[Modify restraint weights]]]]]

insufficient

1. Sampling Precision Estimate

A. Clustering @ precision

B. χ^2 test

Proposed Model(s)

poor fit to data

2. Analyze fit to data

unstable ensemble

3. Cross-validation

stable ensemble

data fits well

poor fit to data

5. Fit to hold-out set

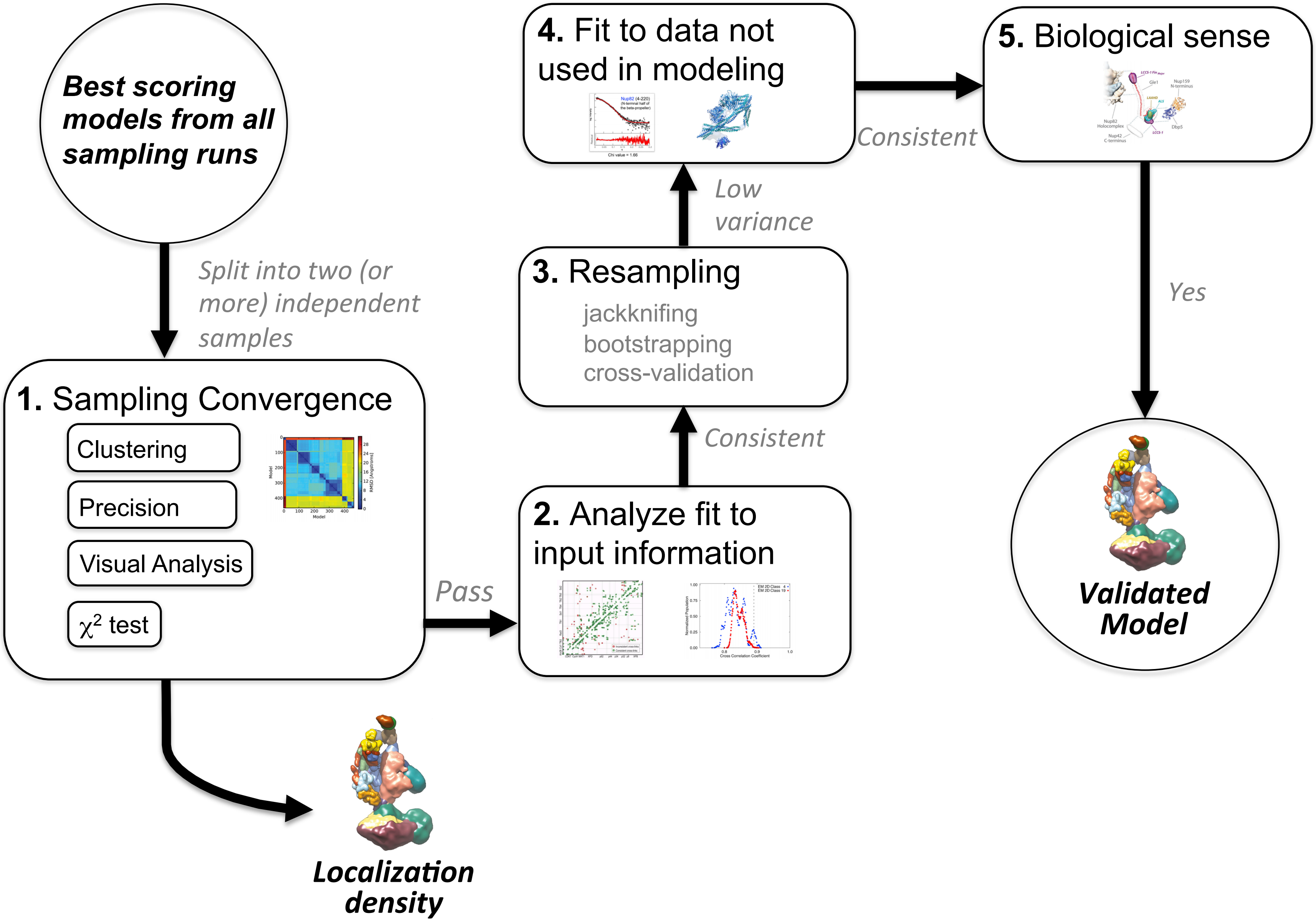
6. Biological sense

Validation

All best scoring models

Interpretive Analysis

Step 4: Analysis



Best scoring models from all sampling runs

Split into two (or more) independent samples

1. Sampling Convergence

Clustering

Precision

Visual Analysis

χ^2 test

Pass

2. Analyze fit to input information

Consistent

3. Resampling

jackknifing

bootstrapping

cross-validation

Low variance

4. Fit to data not used in modeling

Consistent

5. Biological sense

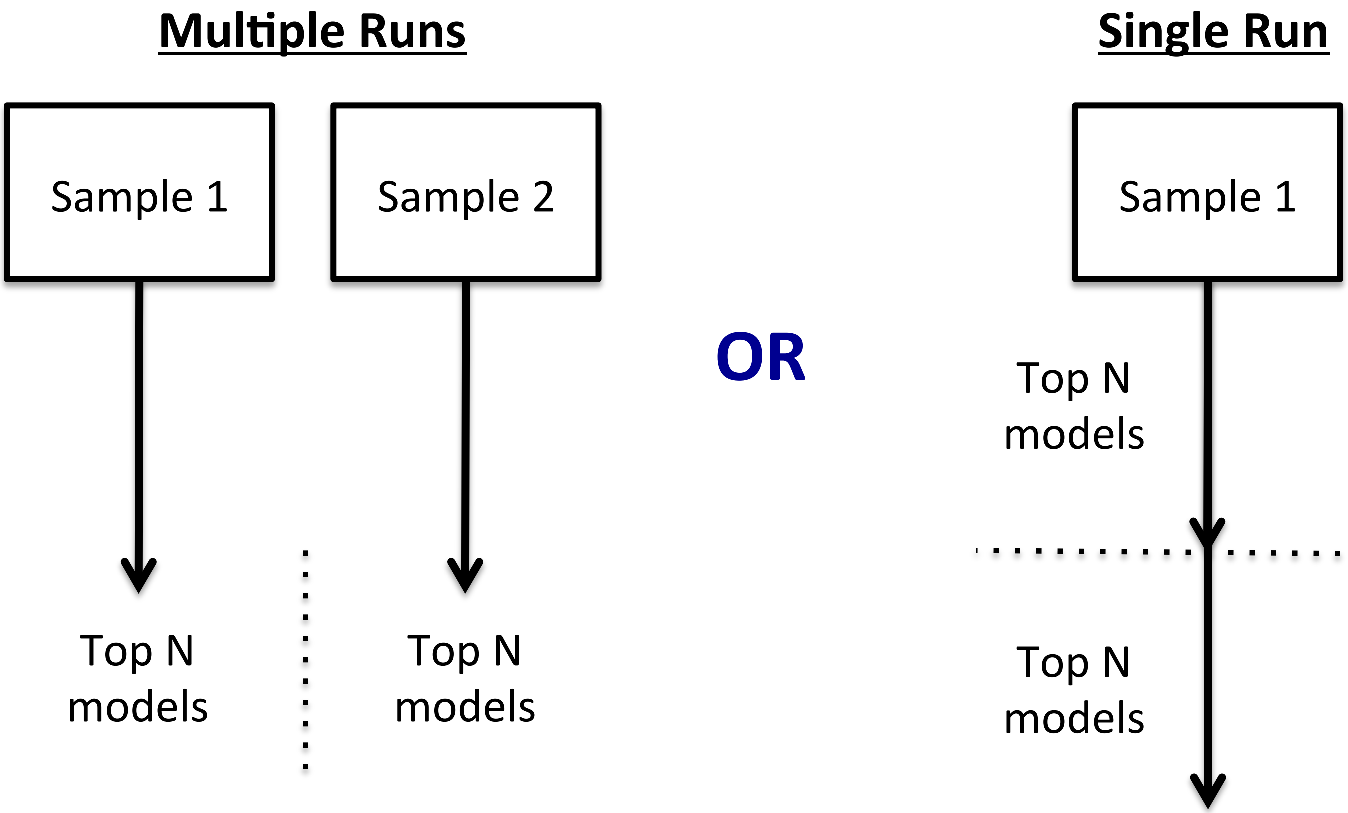
Yes

Validated Model

Localization density

0. Pre-processing

- **Split sampling into multiple independent sets**

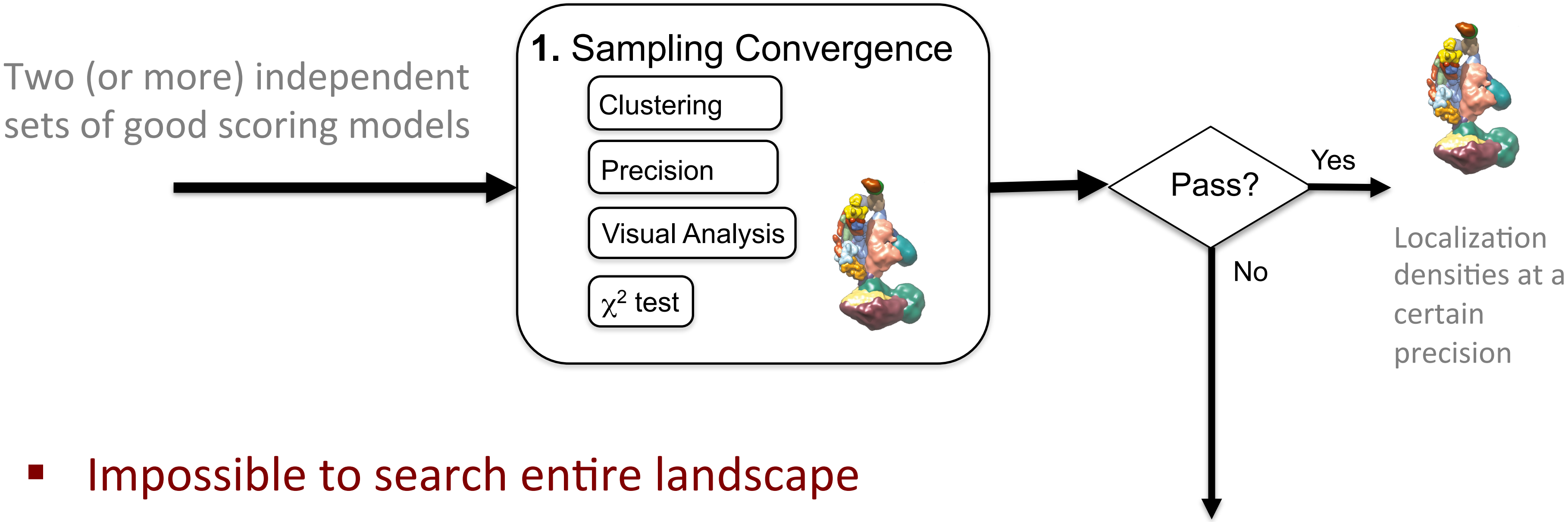


0. Pre-processing

- Split sampling into multiple independent sets
- Gather best scoring models

```
# Must be run in same directory as “output” folder

import IMP
import IMP.pmi
import IMP.pmi.macros

num_models = 100

model = IMP.Model()
are = IMP.pmi.macros.AnalysisReplicaExchange0(model)

are.clustering(score_key='Total_Score',
           feature_keys=[],
           rmsd_calculation_components=None,
           alignment_components=None,
           number_of_best_scoring_models=num_models,
           skip_clustering=True,
           first_and_last_frames=(0,100) # values are percentages…
             )                           # …use to split a single trajectory
```

1. Assessing Sampling Exhaustiveness



Two (or more) independent sets of good scoring models

1. Sampling Convergence

Clustering

Precision

Visual Analysis

χ^2 test

Pass?

Yes

No

Localization densities at a certain precision

- Impossible to search entire landscape

- **Method:** Compare independent samples of models
  - **Visual analysis:**
    Compare localization densities.
  - **Statistical (in)significance:**
    Show no statistically significant differences between clustering results

- **Sample more**
- **Reduce sampling space**
  - add more information
- **Reduce DOF**
  - reduce representation
  - impose symmetry

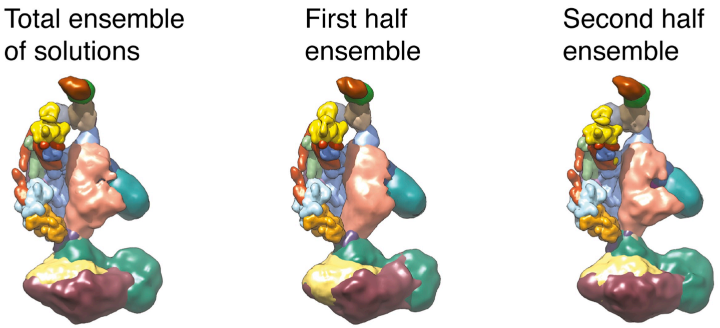
** No method gives proof of convergence*

1. Assessing Sampling Exhaustiveness

- Visual Analysis
  - Get clusters and localization densities for each independent cluster

```
import IMP
import IMP.pmi
import IMP.pmi.macros

rmf_dir = ./rmfs/     # path to the rmf directory
num_rmfs = 4          # number of rmfs in the directory
num_clusters = 1

# Setup macro
model = IMP.Model()
mc = IMP.pmi.macros.AnalysisReplicaExchange0(model)

rmsdc = {"B":"B"} # compo
alignc = None

densityc = {"Spc97":["Spc97"],"Spc98":["Spc98"],"Tub4":
["Tub4"],"Spc110":["Spc110"]}
#densityc = None

mc.clustering (rmsd_calculation_components=rmsdc,
           number_of_clusters=num_clusters,
            display_plot=True,
           number_of_best_scoring_models=num_rmfs,
           exit_after_display = False,
            rmfsdir=rmf_dir,
           density_custom_ranges = densityc)
```

Yeast Mediator Complex



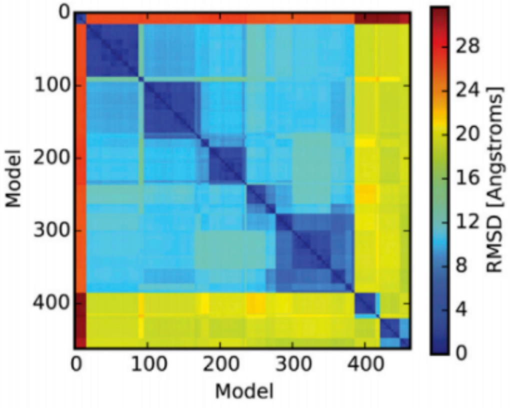
Robinson, Trnka et. al. 2015. eLife

1. Assessing Sampling Exhaustiveness

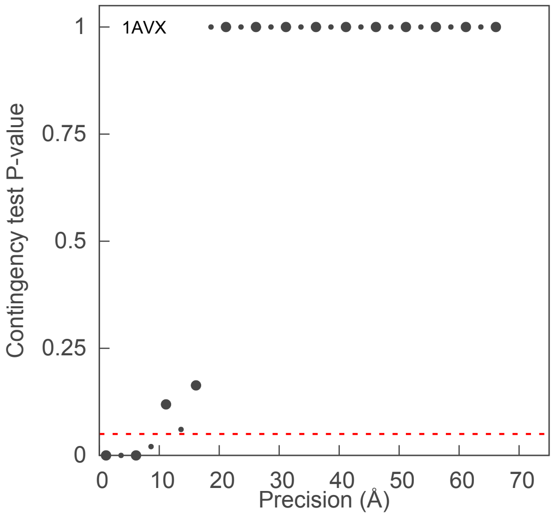
Clustering and Precision

- Distance matrix is determined by pairwise C_α RMSD calculation
- **k-means** is used to separate into clusters based on RMSD
  - Must specify the number of clusters
- How many clusters to choose?
  - Visual analysis
  - Clustering metrics
- Clustering choices determine precision of your models
  - Many clusters – high precision
  - Fewer clusters – low precision

Distance Matrix

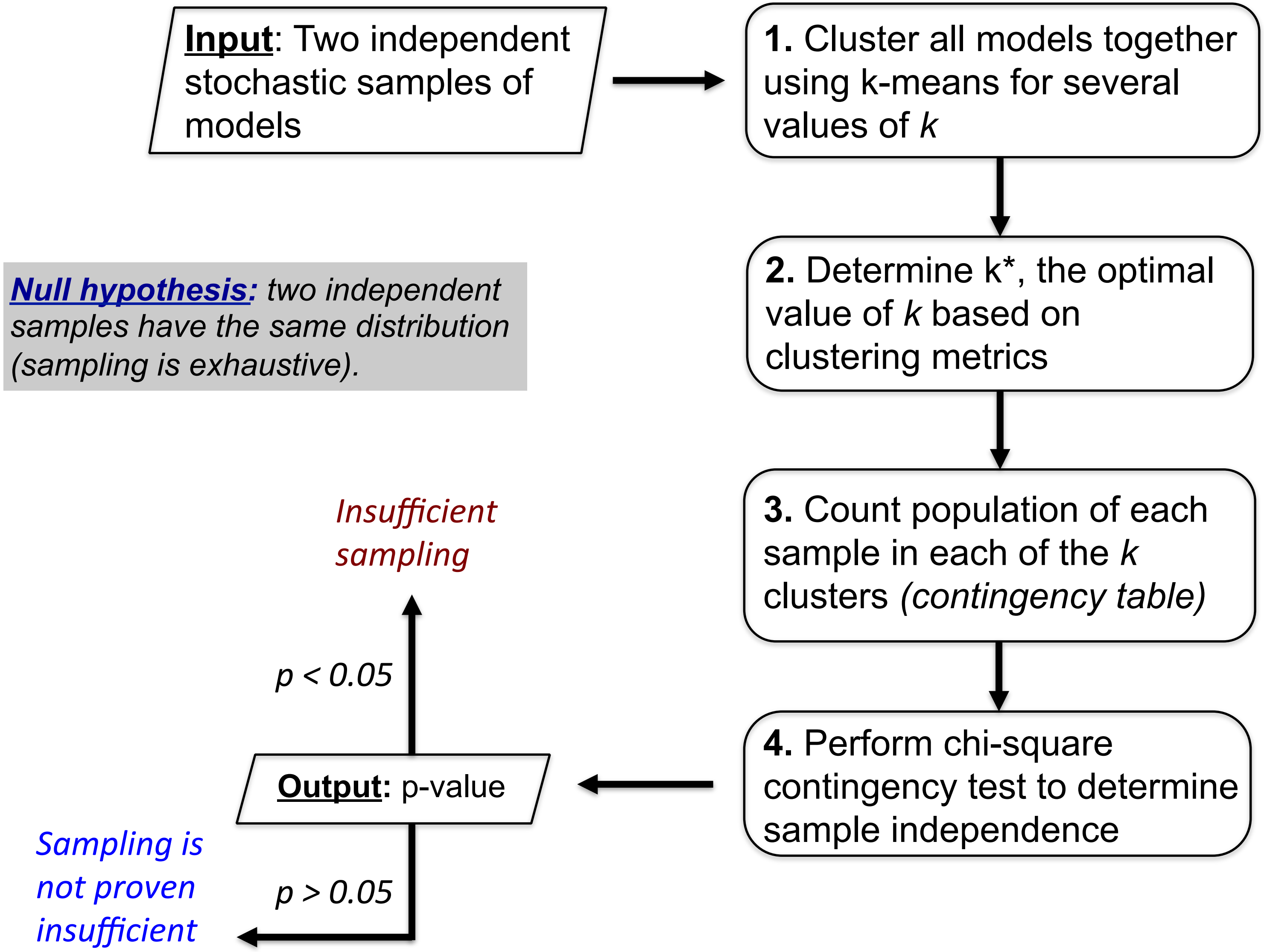


Nup82 – Top 463 models



1. Assessing Sampling Exhaustiveness

Chi² Sampling Test Flowchart



Input: Two independent stochastic samples of models

1. Cluster all models together using k-means for several values of *k*

2. Determine k*, the optimal value of *k* based on clustering metrics

3. Count population of each sample in each of the *k* clusters *(contingency table)*

4. Perform chi-square contingency test to determine sample independence

Output: p-value

Null hypothesis: *two independent samples have the same distribution (sampling is exhaustive).*

$p < 0.05$

Insufficient sampling

$p > 0.05$

Sampling is not proven insufficient

Chi-squared convergence test

- **INPUT:** Get *N* top scoring models for each run from the output of sampling

  `get_top_models_each_run.py <N>`

- **1. Clustering:** Perform k-means clustering on the combined set of models

  `cluster_kn.py`

  `precision_rmsf.py`

- **2. Determine k*:** Determine the optimal value of k using clustering metrics

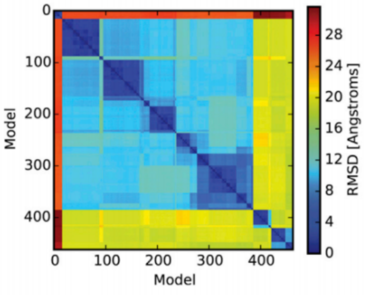
  `metric_wrapper.sh`

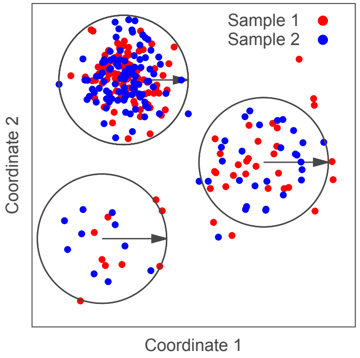
  - **Dunn Index:** ratio of minimum inter cluster precision to maximum intra cluster precision.

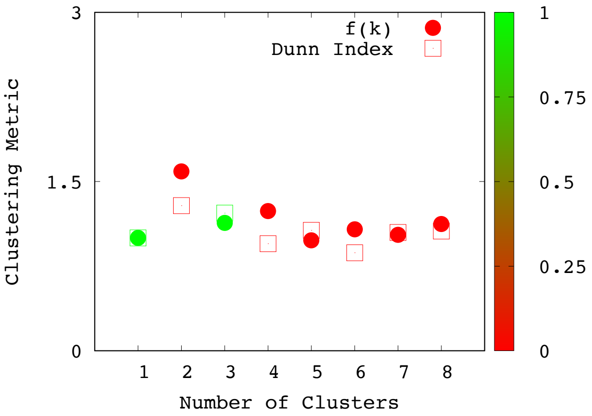
    `metric_dunn.py`

  - **Distortion Index:**, f(k) : does having *k* clusters produce a smaller distortion than having *k-1* clusters?

    `metric_fk.py`







Contingency table and p-value calculation

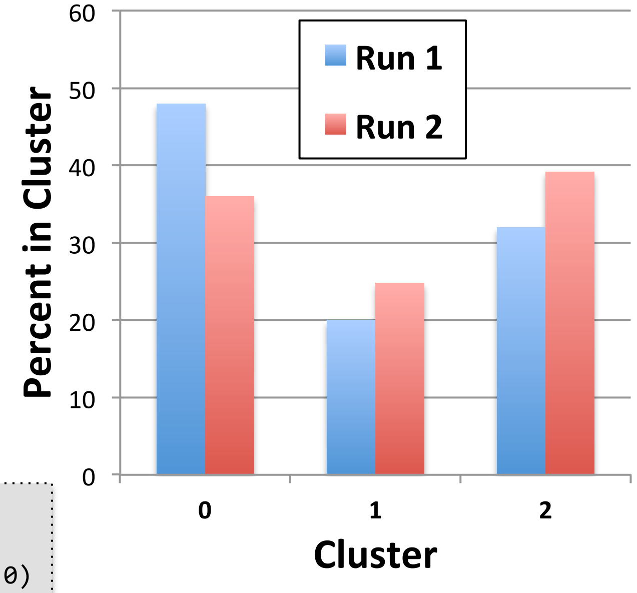
- **3. Population Count:** Calculate number of models from each run in all clusters to form *contingency table*

  `get_models_per_cluster_kmeans.py`

	Pct. of Run in Cluster	
Cluster	**Run 1**	**Run 2**
0	48.0	36.0
1	20.0	24.8
2	32.0	39.2

- **4. Calculate *p*-value:** A p-value < 0.05 indicates a statistically significant difference between populations and incomplete sampling

  `test_sampling_convergence.py`

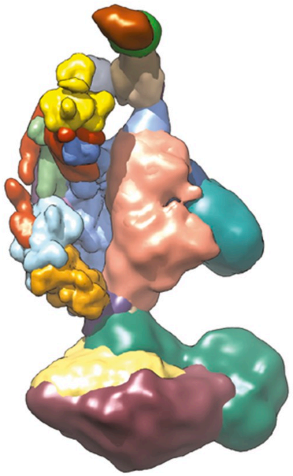


***p*-value = 0.228**

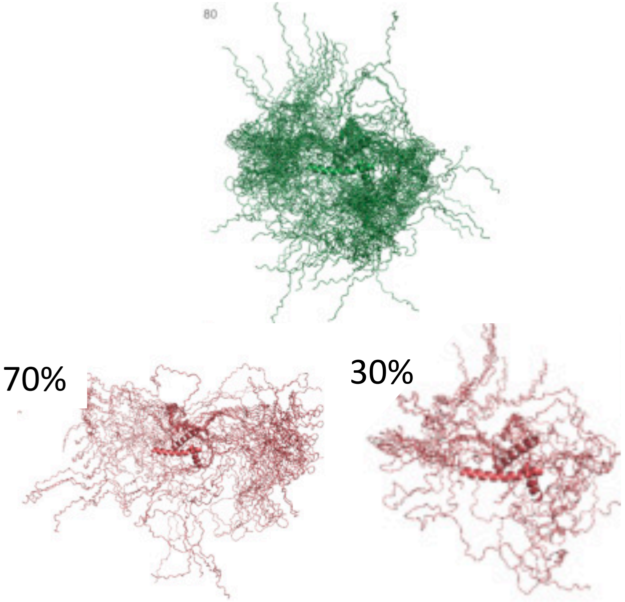
```
numModelsFile = sys.argv[1]  # file with number of models per cluster
modelsArray = numpy.loadtxt(numModelsFile)
percentArray = numpy.transpose((modelsArray/modelsArray.sum(axis=0)) * 100.0)
[chisquare,pvalue,dof,expected]=scipy.stats.chi2_contingency(percentArray)
print "P-value",pvalue
```

1. Assessing Sampling Exhaustiveness

- **Output:**
  - Clusters
    - Localization Density (or Ensemble)
    - Precision



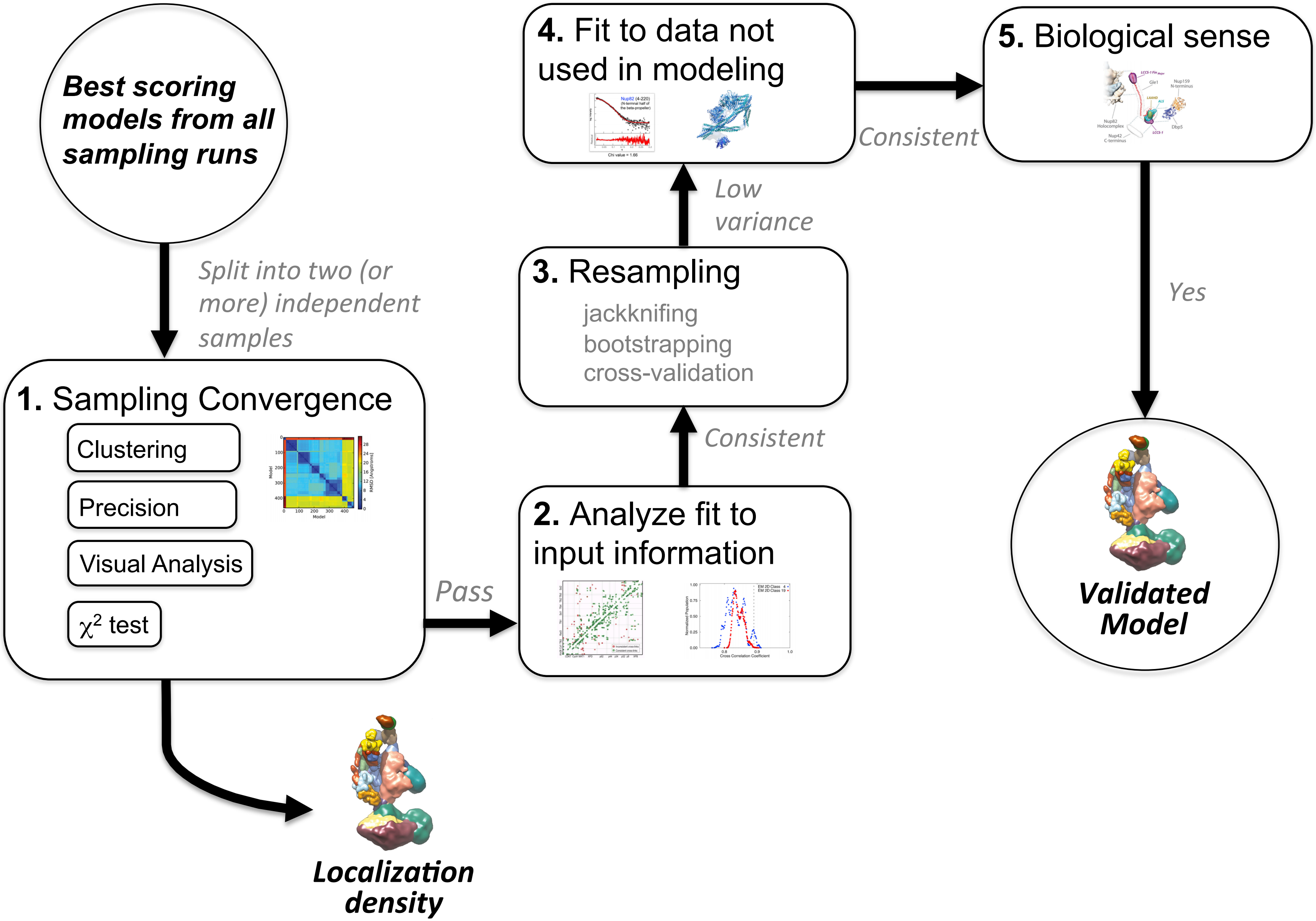
Single cluster ensemble



Comparison of single and multi-state ensembles

Carter, Lester, et al. "Prion Protein—Antibody Complexes Characterized by Chromatography-Coupled Small-Angle X-Ray Scattering." Biophysical journal 109.4 (2015): 793-805.

Step 4: Analysis



Best scoring models from all sampling runs

Split into two (or more) independent samples

1. Sampling Convergence

Clustering

Precision

Visual Analysis

χ^2 test

Pass

2. Analyze fit to input information

Consistent

3. Resampling

jackknifing

bootstrapping

cross-validation

Low variance

4. Fit to data not used in modeling

Consistent

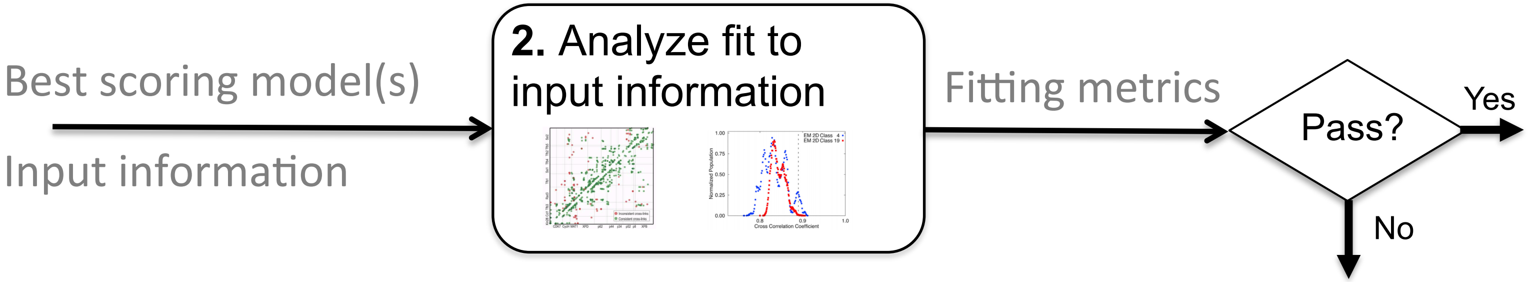
5. Biological sense

Yes

Validated Model

Localization density

2. Assessing Fit to Data


Best scoring model(s)

Input information

2. Analyze fit to input information

Fitting metrics

Pass?

Yes

No

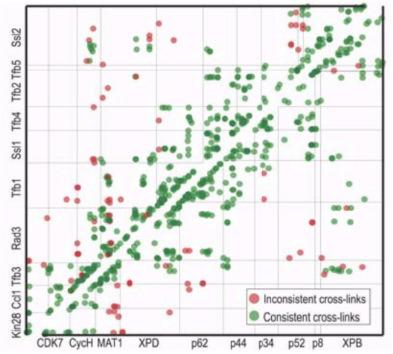
- **Examine restraints that are not satisfied by any model**
  - Artifacts
  - Different experimental conditions
- **Evaluate a multi-state model**
  - Can you satisfy the model with two states simultaneously

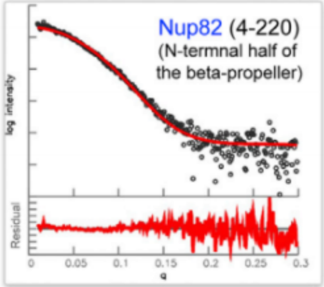
- **Method:** Does the resulting ensemble of best scoring models actually represent the input data?
  - Passing criteria are subjective

2. Assessing Fit to Data

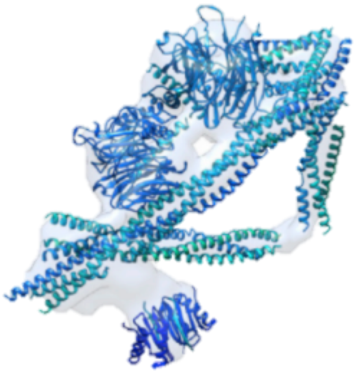
Assessing Violations by Data Type

- **Crosslinks**
  - Distance violations
  - Score violations
- **SAXS**
  - chi^2 value
  - Radius of Gyration
- **EM**
  - Cross Correlation
  - Visual inspection





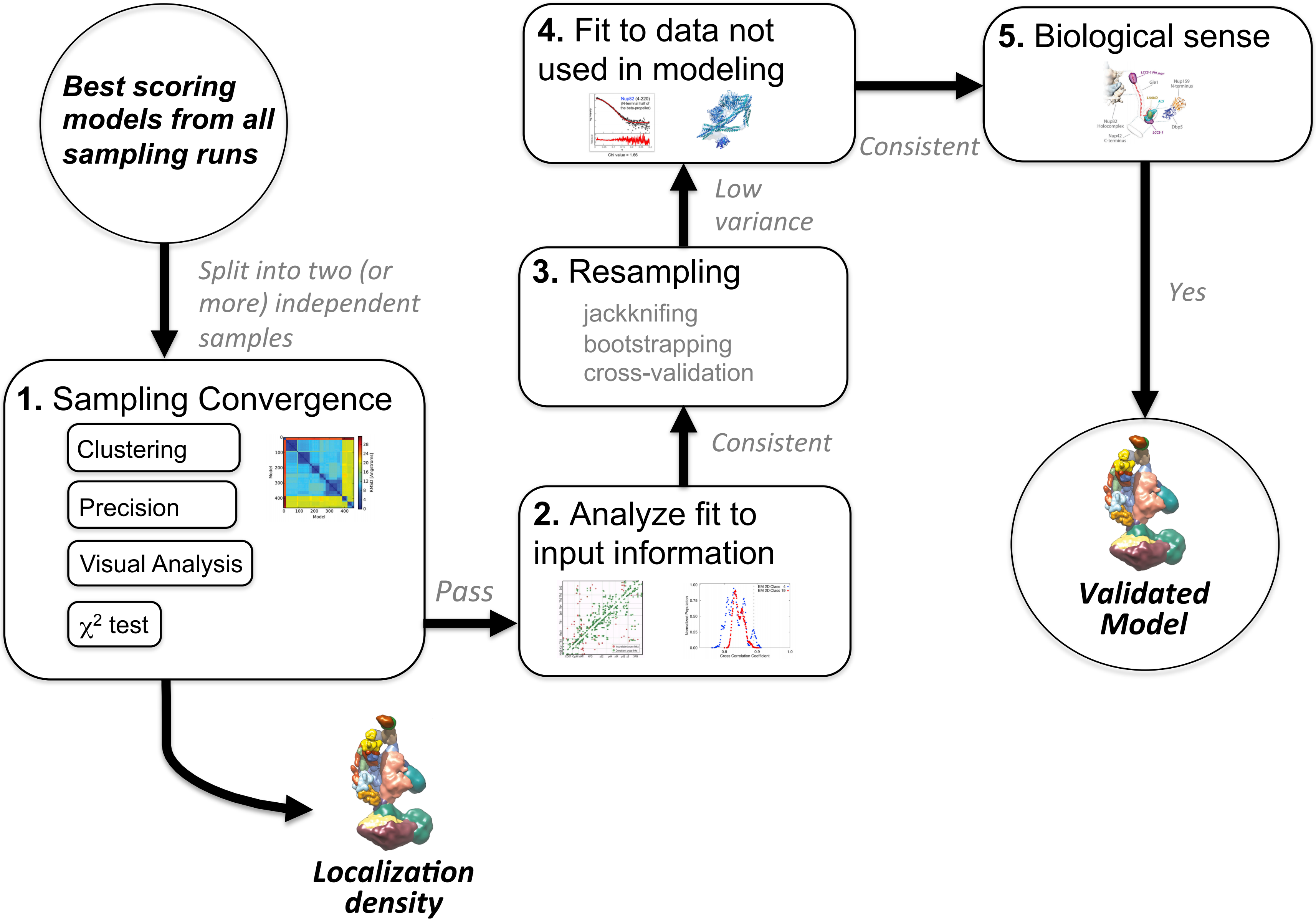
Chi value = 1.66



Subjective Questions:

- **How do we define a violation?**
- **How many violations define a failing model?**

Step 4: Analysis



Best scoring models from all sampling runs

Split into two (or more) independent samples

1. Sampling Convergence

Clustering

Precision

Visual Analysis

χ^2 test

Localization density

Pass

2. Analyze fit to input information

Consistent

3. Resampling

jackknifing

bootstrapping

cross-validation

Low variance

4. Fit to data not used in modeling

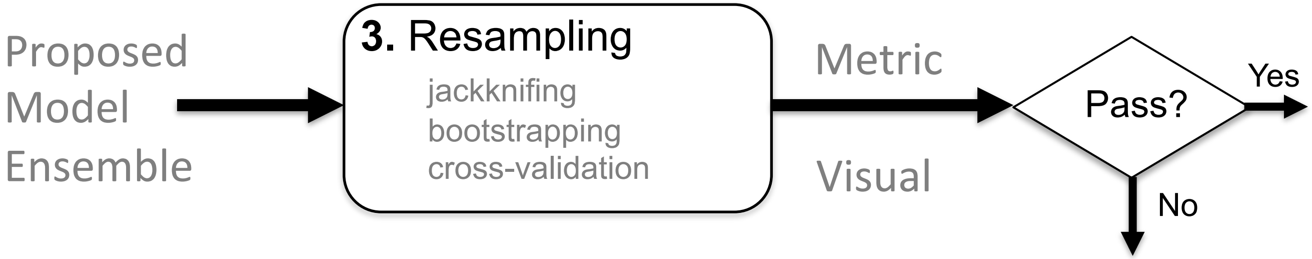
Consistent

5. Biological sense

Yes

Validated Model

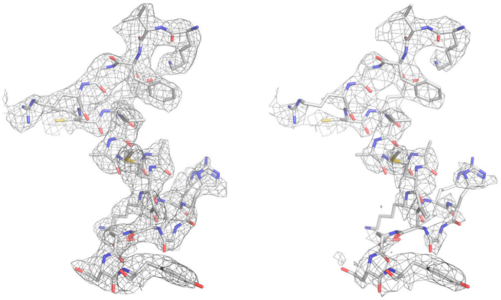
3. Resampling Methods



- **Recalculate models using subsets of the data**
  - **Bootstrapping**
    - Remove random subsets of data
  - **Jackknifing**
    - Remove systematic subsets of data
    - **Cross-validation**
      - Predict values of held-out data
      - Score to original data

- Prevent overfitting to certain data
- Assess the stability of the model ensemble with respect to target data.

- **Model is too dependent on certain data**
  - Reduce weight of the offending data
- **Data is not self-consistent**

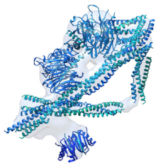

MR Map
with Model

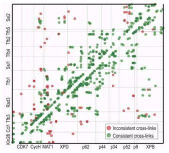
Composite Omit Map
with Model

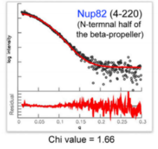
Similar to calculating the composite omit map

3. Resampling Methods

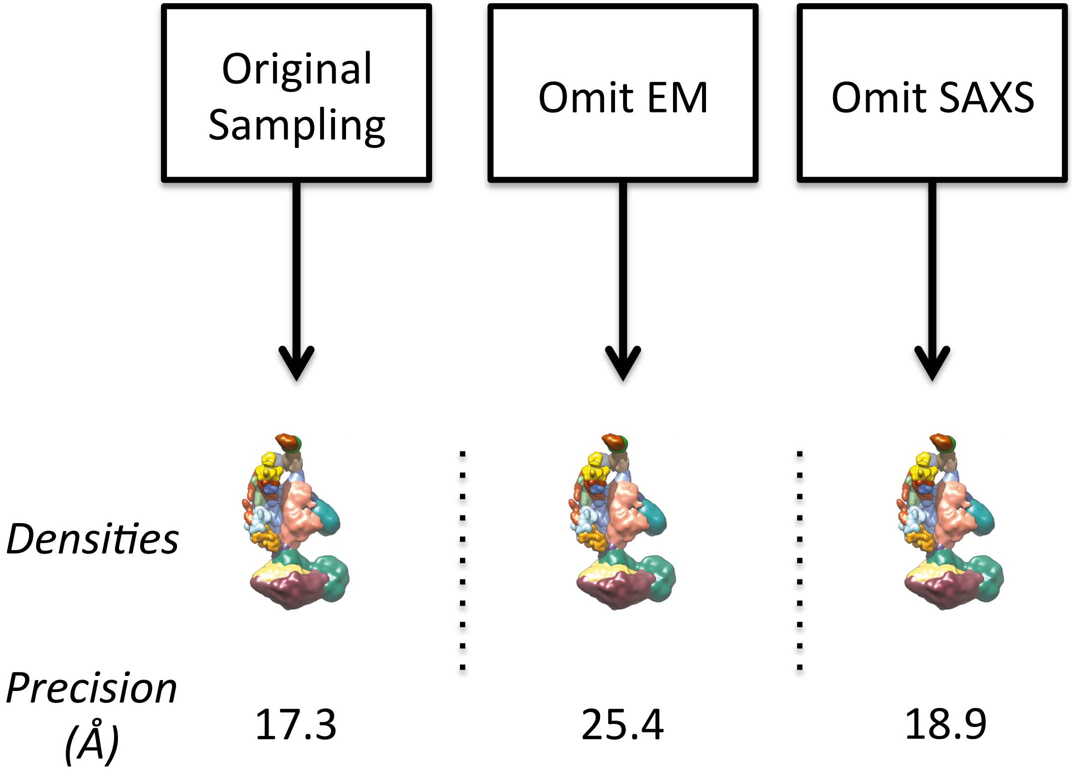
- **Jackknifing**
  - **Omit pieces of data**
    - **Whole sets**
      - EM
      - SAXS
    - **Subsets**
      - XL







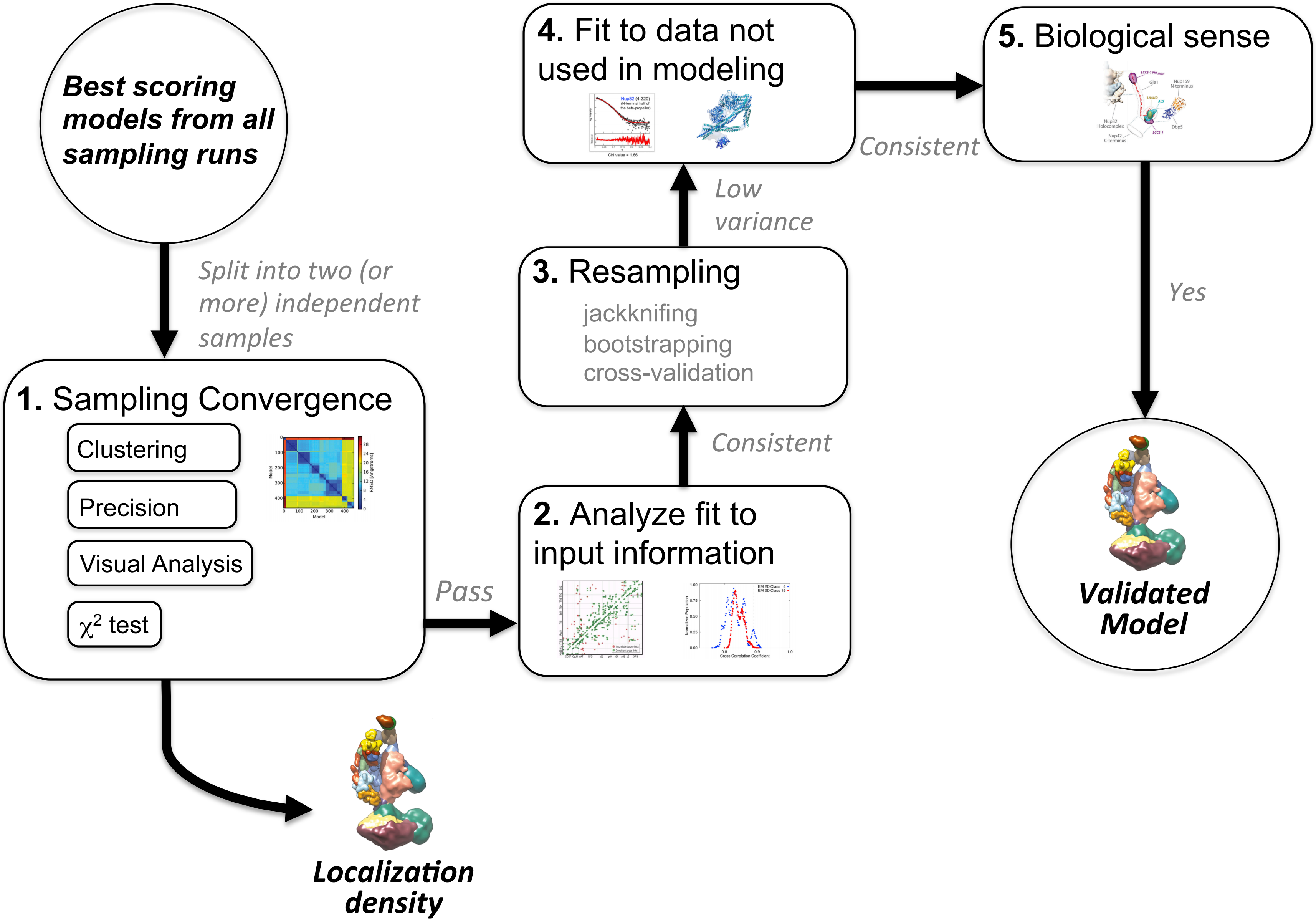
  - **Densities similar?**
  - **Precision similar?**



	Original Sampling	Omit EM	Omit SAXS
Densities			
Precision (Å)	17.3	25.4	18.9

> Practical Considerations:
> Recalculating the entire ensemble is expensive.

Step 4: Analysis



Best scoring models from all sampling runs

Split into two (or more) independent samples

1. Sampling Convergence

- Clustering
- Precision
- Visual Analysis
- χ^2 test

Localization density

Pass

2. Analyze fit to input information

Consistent

3. Resampling

- jackknifing
- bootstrapping
- cross-validation

Low variance

4. Fit to data not used in modeling

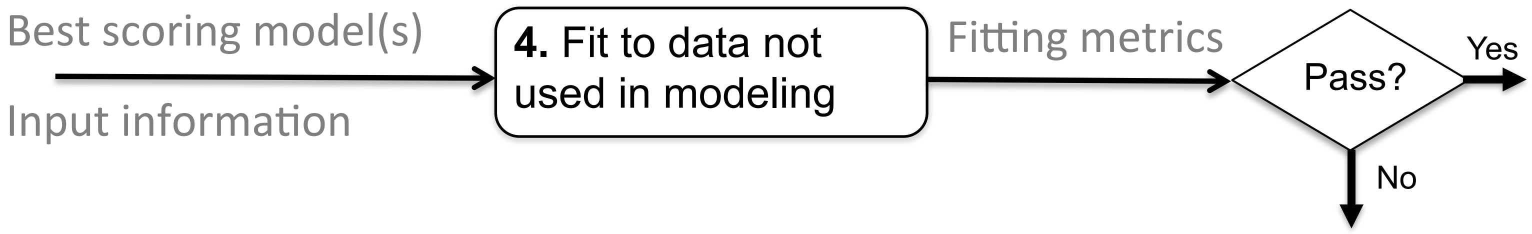
Consistent

5. Biological sense

Yes

Validated Model

4. Fit to Information Not Used in Modeling



Best scoring model(s)

Input information

4. Fit to data not used in modeling

Fitting metrics

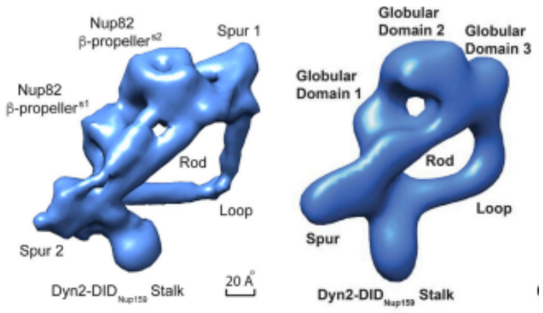
Pass?

Yes

No

- **Same methodology as Step 2**
  - Pre-defined hold-out set
  - Information that is difficult to embed in a restraint
  - Information from a slightly different construct
  - New information collected after modeling

- **Examine restraints that are not satisfied by any model**
  - Artifacts
  - Different experimental conditions

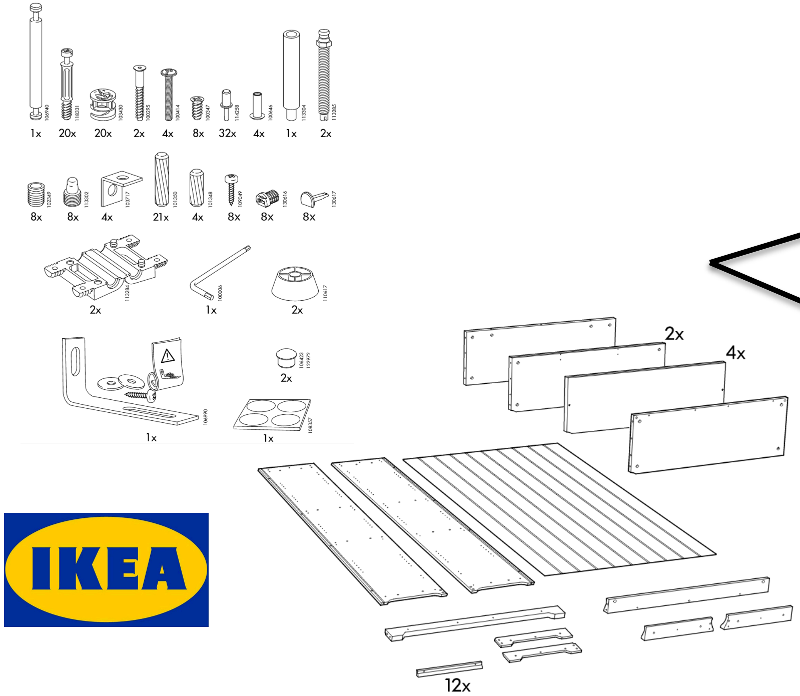


Fernandez-Martinez et al. 2016 *Gaik et al. 2015*

Comparison of Nup82 models to negative stain EM of truncated model

5. Biological Significance

- **The utility of the model is, in itself, a validation.**
  - Satisfaction of patterns unlikely to occur by chance
  - A wrong model is not likely to make sense



Supposed to be a bookshelf

- Poor book holder
- Pretty unstable

Probably incorrect

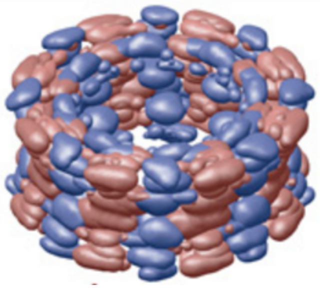
- Can hold books
- Looks like IKEA
- Doesn't fall apart

Probably correct

5. Biological Significance

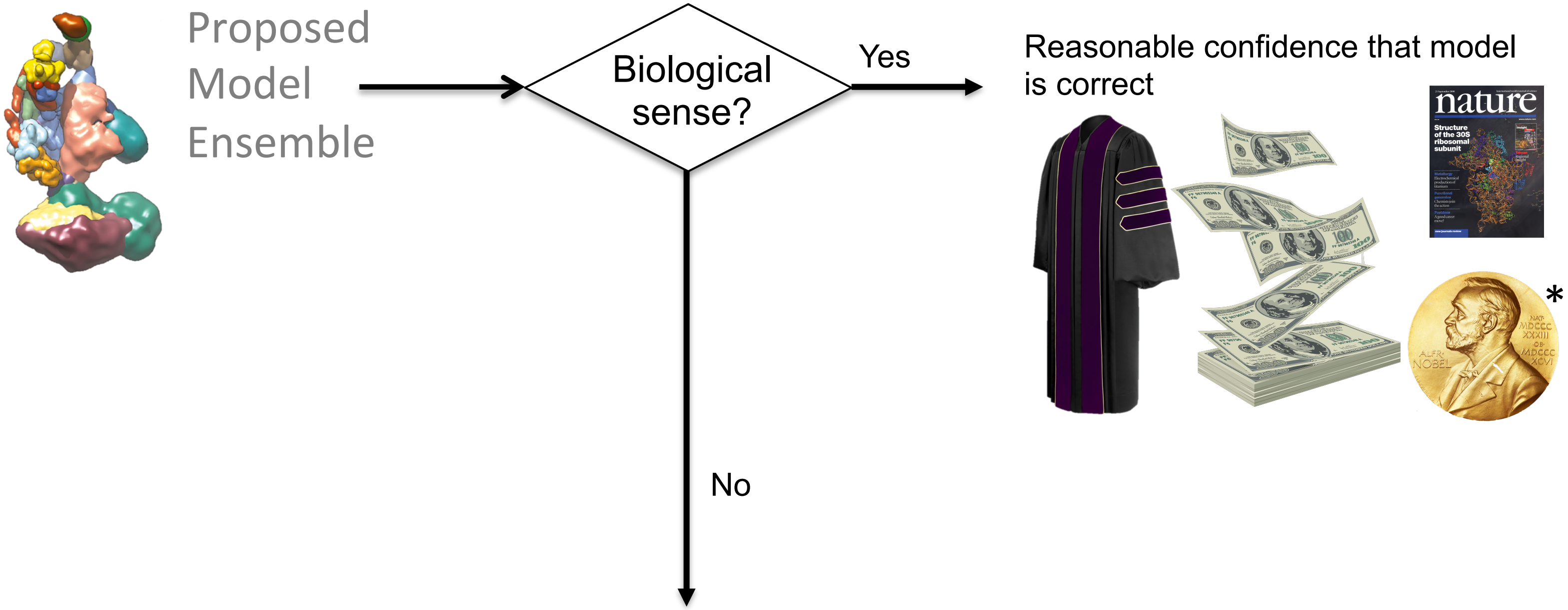
- **The utility of the model is, in itself, a validation.**
  - Satisfaction of patterns unlikely to occur by chance

Observation of suspected 16-fold symmetry in the NPC



Alber, Frank, et al. "The molecular architecture of the nuclear pore complex." Nature 450.7170 (2007): 695-701.

5. Biological Significance


nature
Structure of the 30S ribosomal subunit
ALFR. NOBEL
NAT. MDCCCXXXIII OB. MDCCCXCVI
*

Proposed Model Ensemble

Biological sense?

Yes

Reasonable confidence that model is correct

No

Model is not necessarily wrong, but care must be taken in any new claims

What if I need more information?

- **Look outside of traditional structural biophysical experiments**
  - **CoIP**
  - **Hydrogen/Deuterium Exchange**

  - **Make simple assumptions**
    - Symmetry
    - Interface
    - Oligomerization states
    - Stoichiometry

Communicating model validation

- **Recent examples**

Fernandez-Martinez et al., Structure and Function of the Nuclear Pore Complex Cytoplasmic mRNA Export Platform, Cell (2016), http://dx.doi.org/10.1016/j.cell.2016.10.028

Robinson, Philip J., et al. Molecular architecture of the yeast Mediator complex, Elife 4 (2015), http://dx.doi.org/10.7554/elife.08719

Integration into the WWPDB

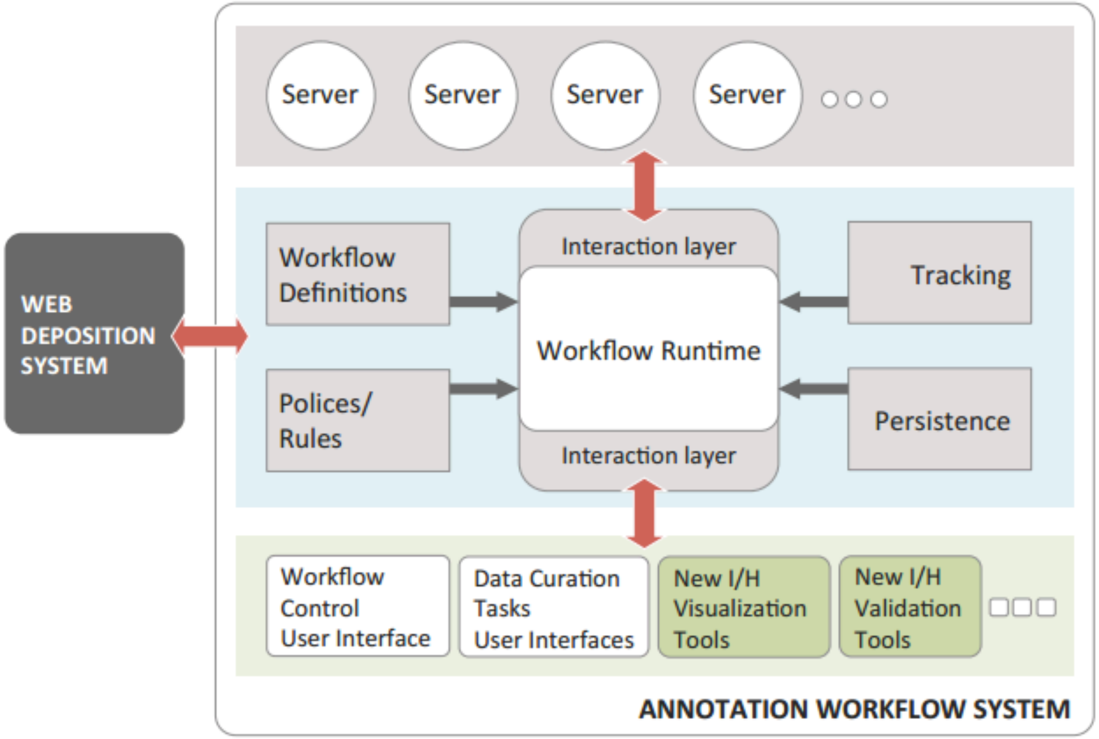


Figure 6. Components of the extensible wwPDB workflow system. It consists of the workflow runtime execution environment, workflow control and curation-task-specific user interfaces, and the supporting compute server infrastructure. The proposed validation and visualization tools for I/H models are highlighted.

Recap

- Validation is a fundamental part of modeling
  - Reduce probability of publishing errors
  - Assessment of the quality of the model and data
- Methods for validating integrative models are under development and not exhaustive
  - Guide using recent examples
  - Watch for future developments / pipelines in IMP

Future of IMP

- IMP is under heavy development
  - 2017 reformulation of the python interface, PMI
  - Check www.integrativemodeling.org
  - Addition of new experimental methods
    - Second Harmonic Generation
    - Hydrogen/Deuterium Exchange
    - Fiber Diffraction
    - ???
- Integration with ChimeraX
- Collaboration pushes IMP forward
  - What interesting problems of yours need solving?