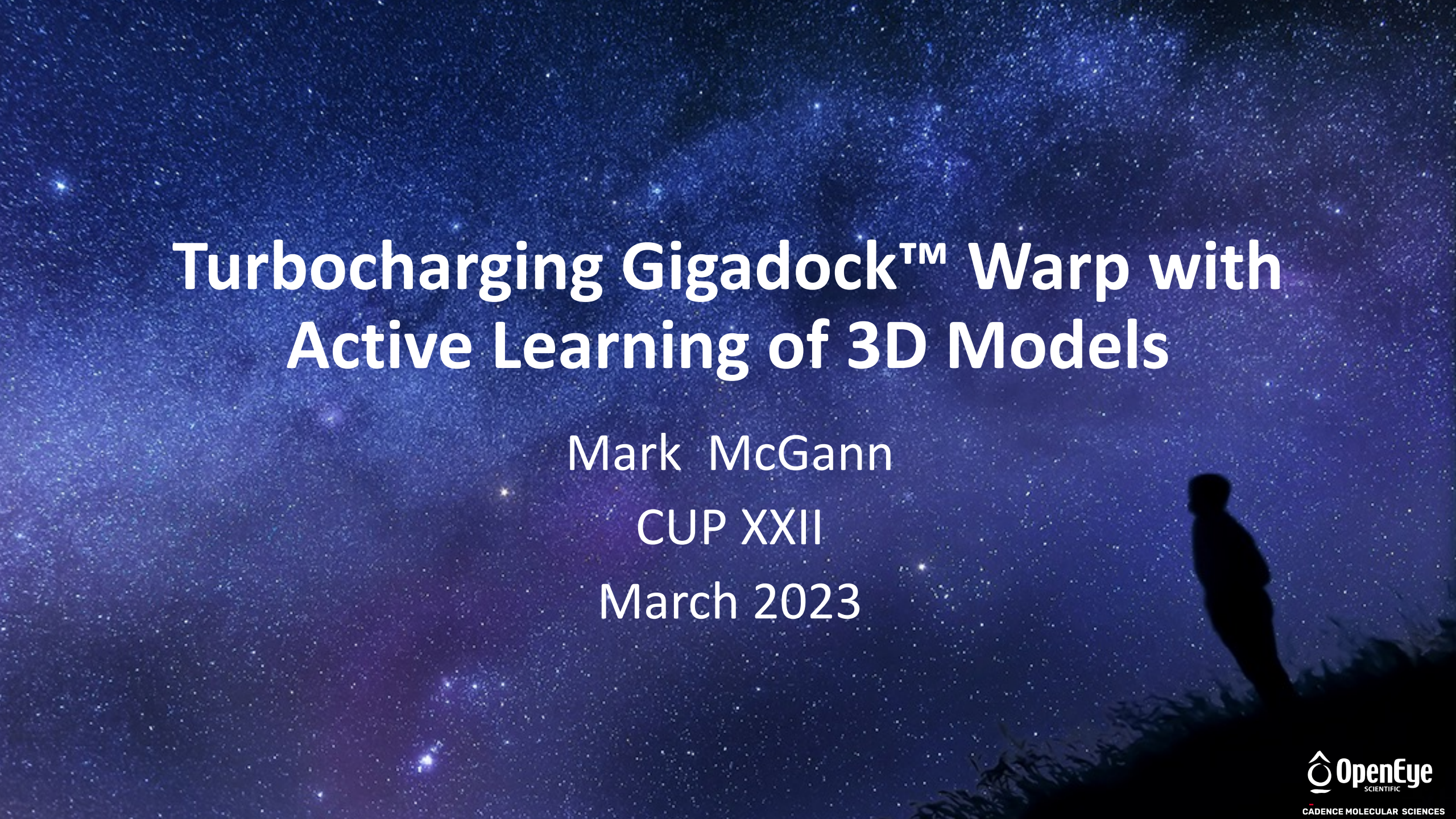
Turbocharging Gigadock™ Warp with Active Learning of 3D Models

Mark McGann

CUP XXII

March 2023



OpenEye SCIENTIFIC

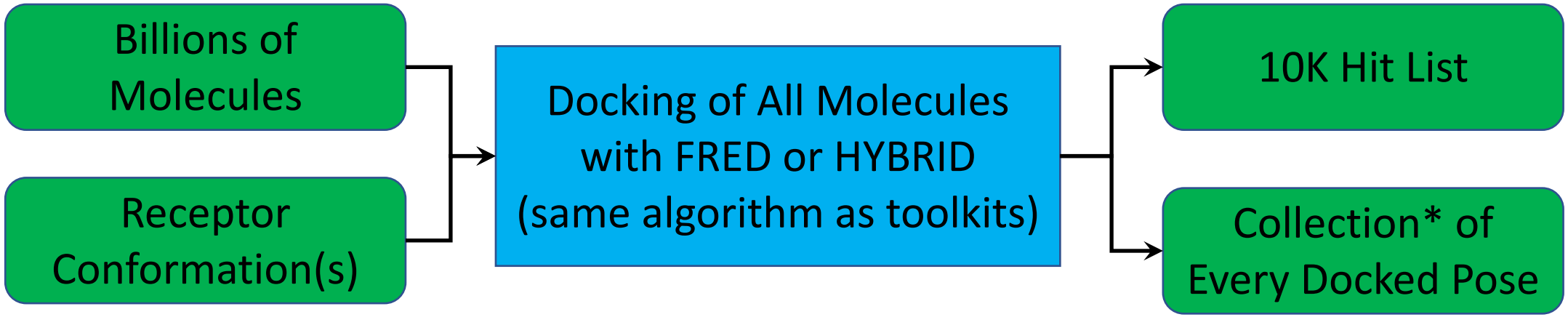
CADENCE MOLECULAR SCIENCES

Overview

- **Current Gigadock Warp**
- Upgrading Gigadock Warp
- Testing Results

Gigadock™ Floe

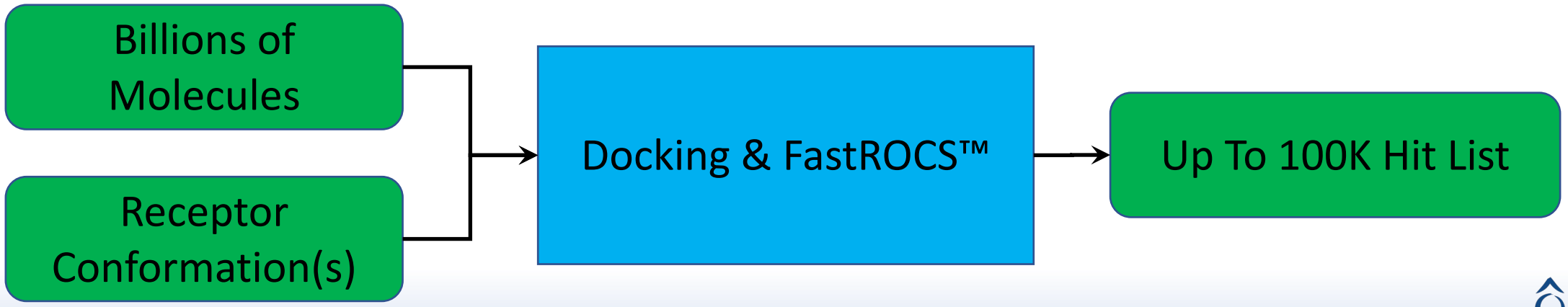
- Floe for complete docking of billions of molecules
- Part of Orion® since initial Orion release in March 2019



* A collection is a data storage mechanism in Orion® for Billions of Molecules

Gigadock Warp Floe

- Drop-in replacement of the Gigadock Floe
- Part of Orion since December 2021
- Goal : Produce same hit list as Gigadock at lower cost
  - Current release gets ~70% identical hitlist when docking Billions*



*Result comparing to HSP90 Gigadock

Why Gigadock Warp

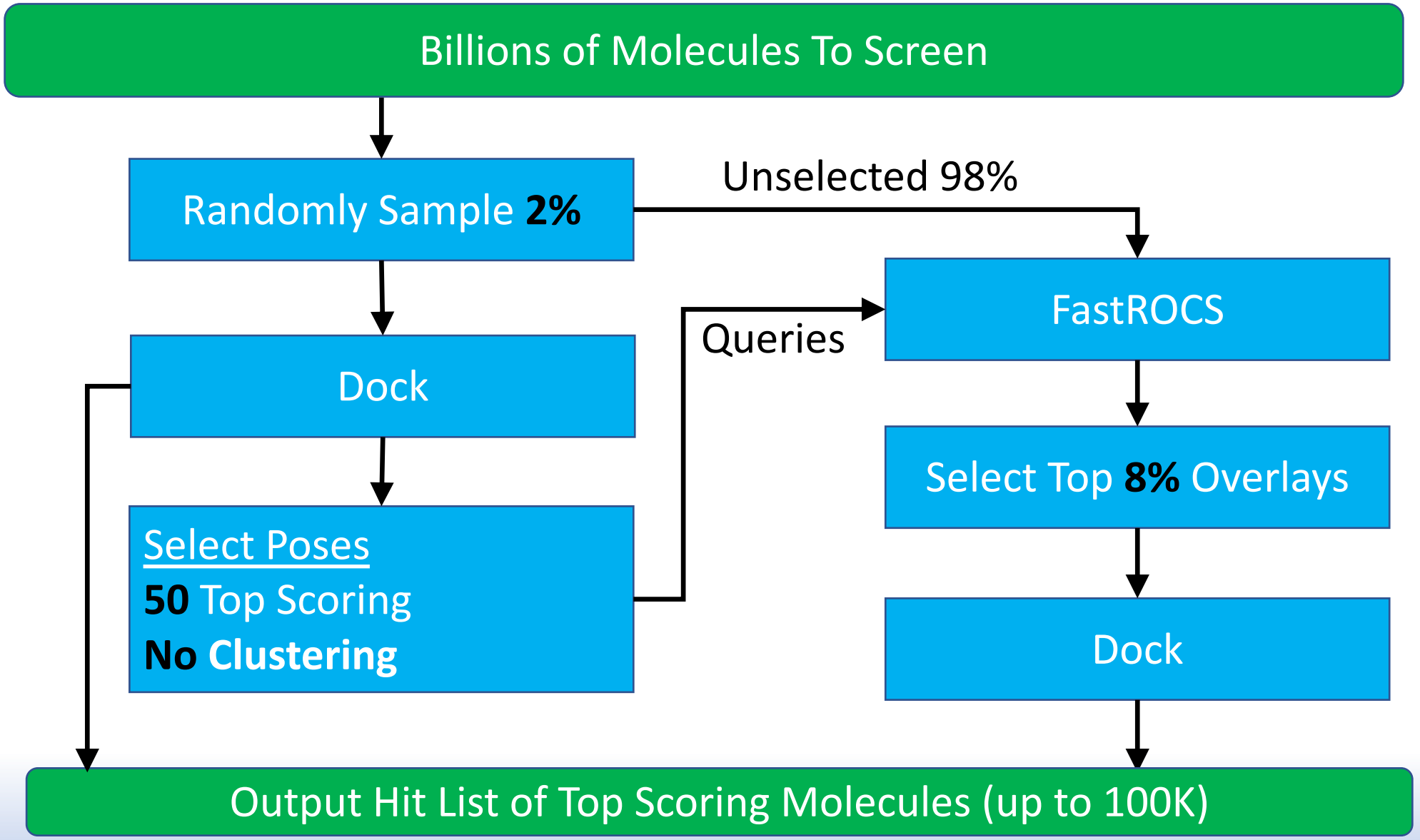
- Latest Enamine collection ~12 Billion Molecules
  - In 2019 enamine was 1.4 Billion
  - Expect size of collections to continue to increase over time

- Cost to dock 12 Billion
  - Gigadock Cost* : **~$10K/Billion** → ~$120K
  - Gigadock Warp Cost** : **~$1.2K/Billion**** -> ~$14K

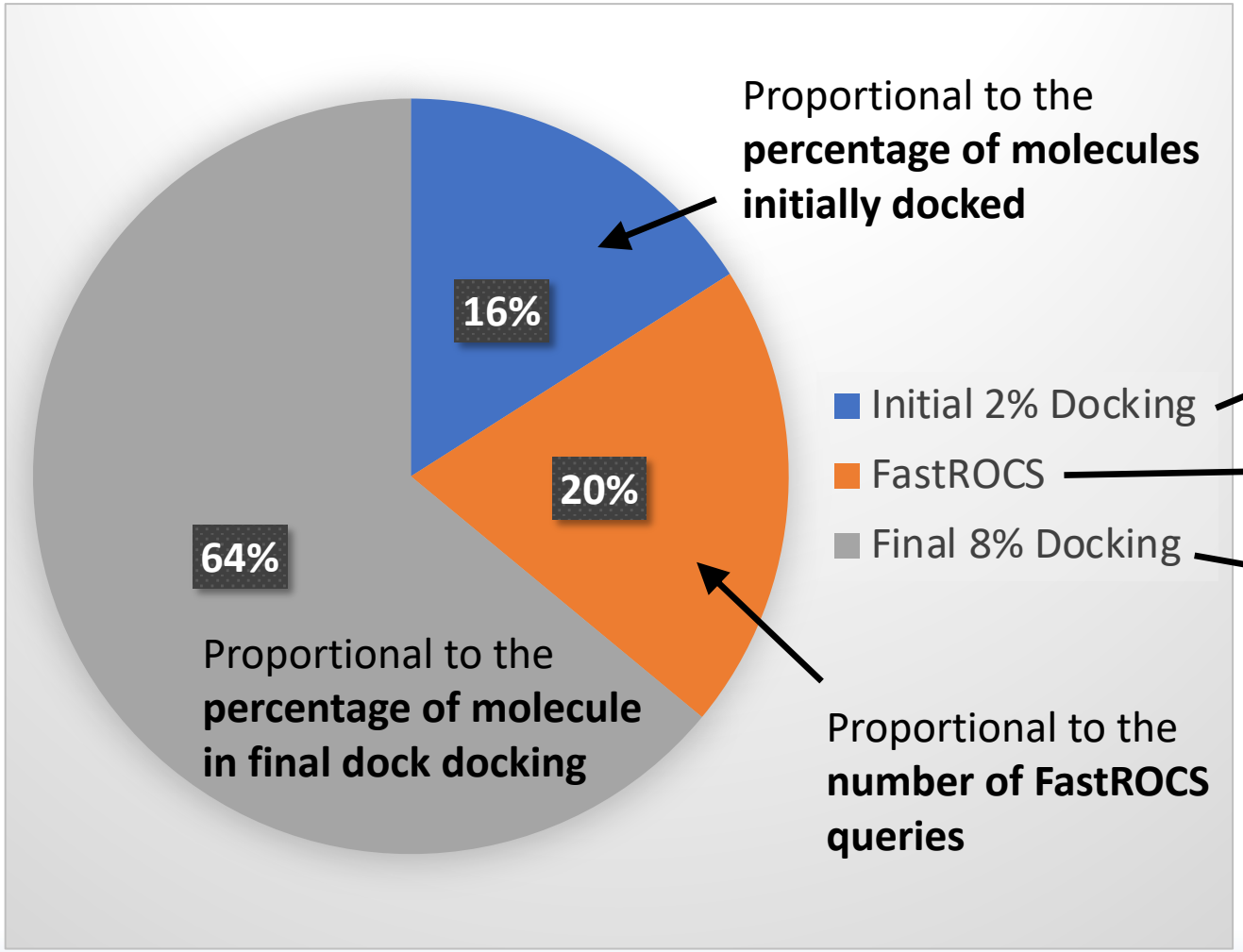
** Cost varies strongly with size of the active site*

*** Estimated cost for current release version*

Gigadock Warp - Algorithm



Gigadock Warp Compute Cost Breakdown



Cost Relative to Full Gigadock

$$Cost_{Initial\ Docking} = 2\%\ Cost_{Gigadock}$$

$$Cost_{FastROCS} = 2.5\%\ Cost_{Gigadock}$$

$$Cost_{Final\ Docking} = 8\%\ Cost_{Gigadock}$$

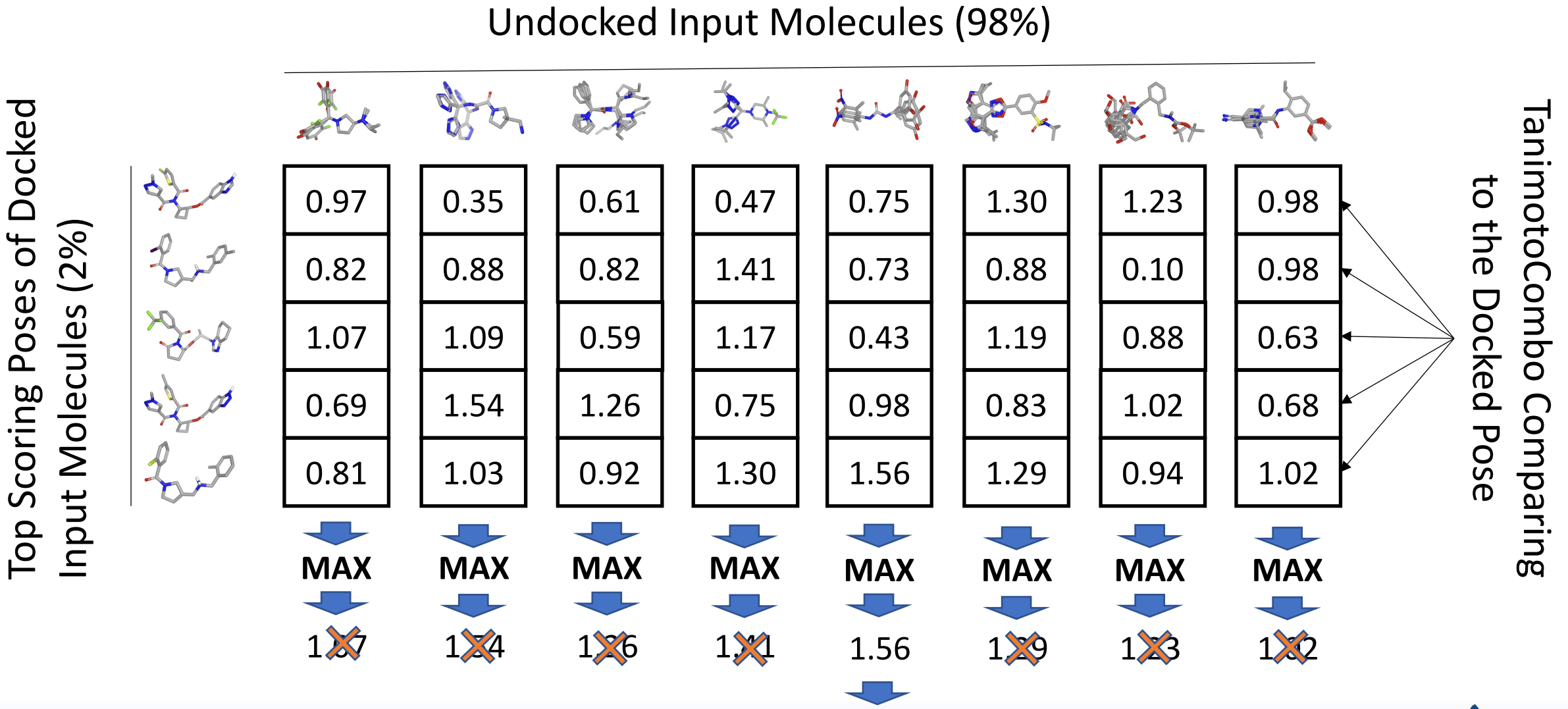
$$\boldsymbol{Cost_{Gigadock\ Wrap} = \frac{Cost_{Gigadock}}{8}}$$

OpenEye SCIENTIFIC
CADENCE MOLECULAR SCIENCES

Overview

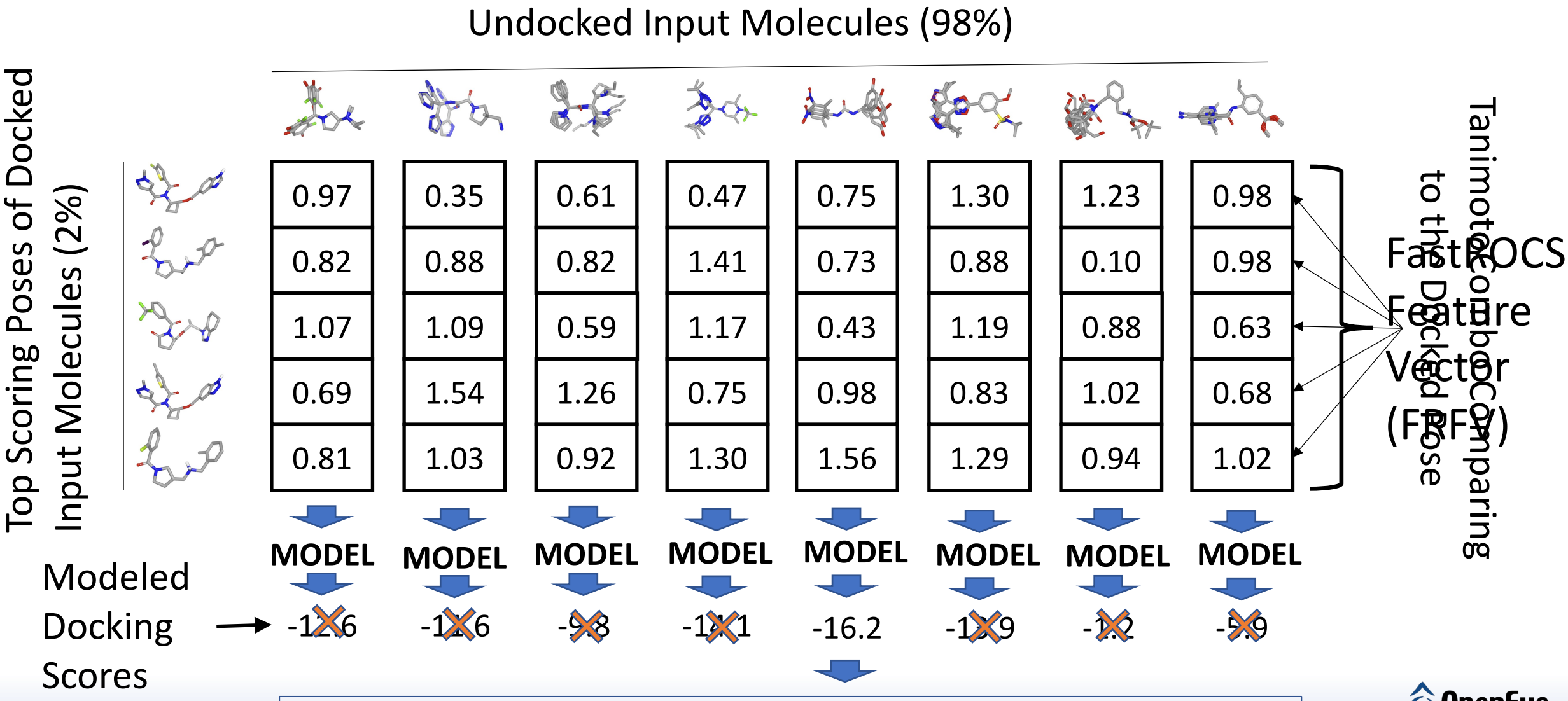
- Current Gigadock Warp
- **Upgrading Gigadock Warp**
- Testing Results

Current FastROCS Selection in Gigadock Warp



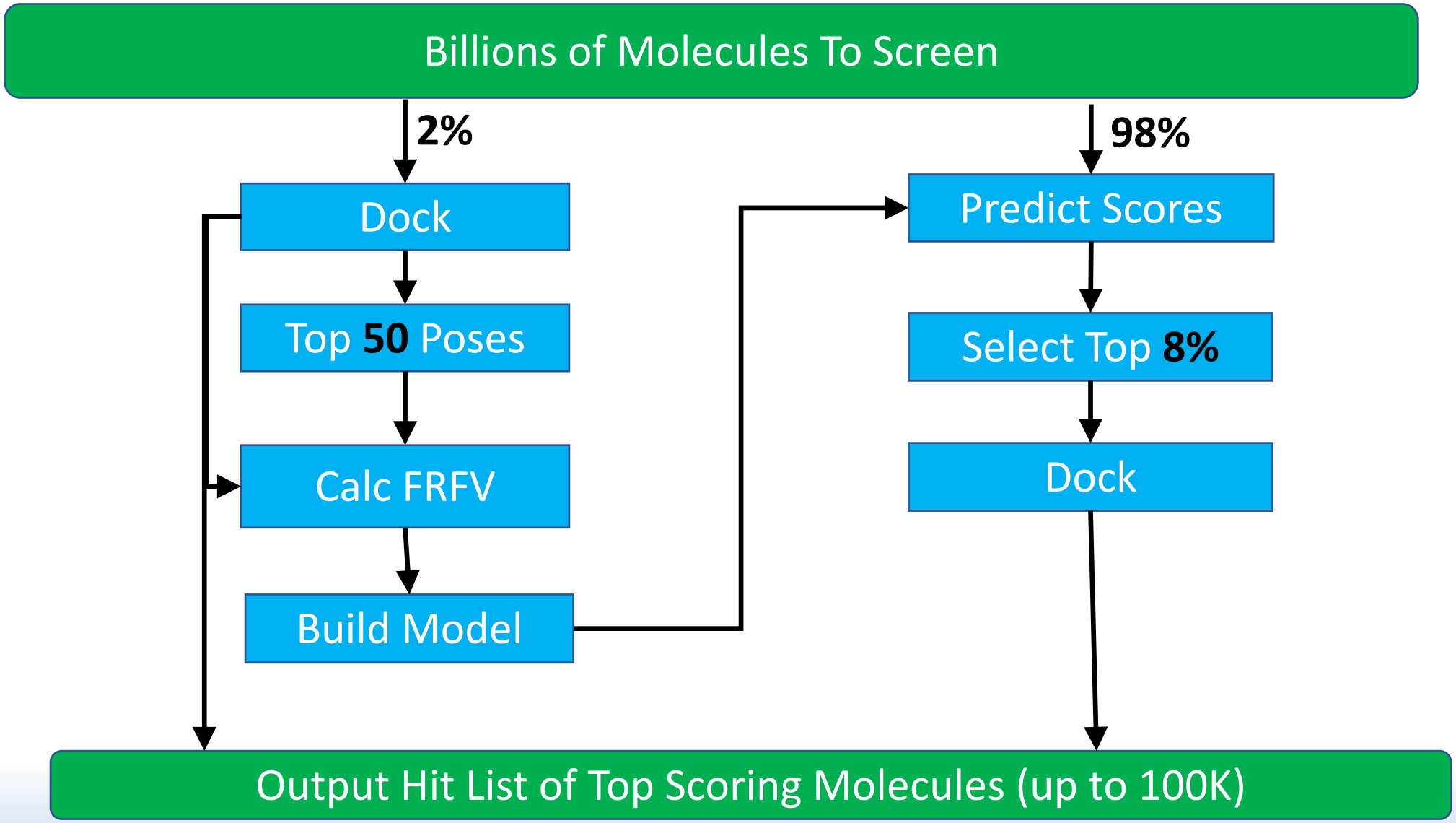
OpenEye SCIENTIFIC
CADENCE MOLECULAR SCIENCES

Modeling with FastROCS Feature Vectors (FRFV)

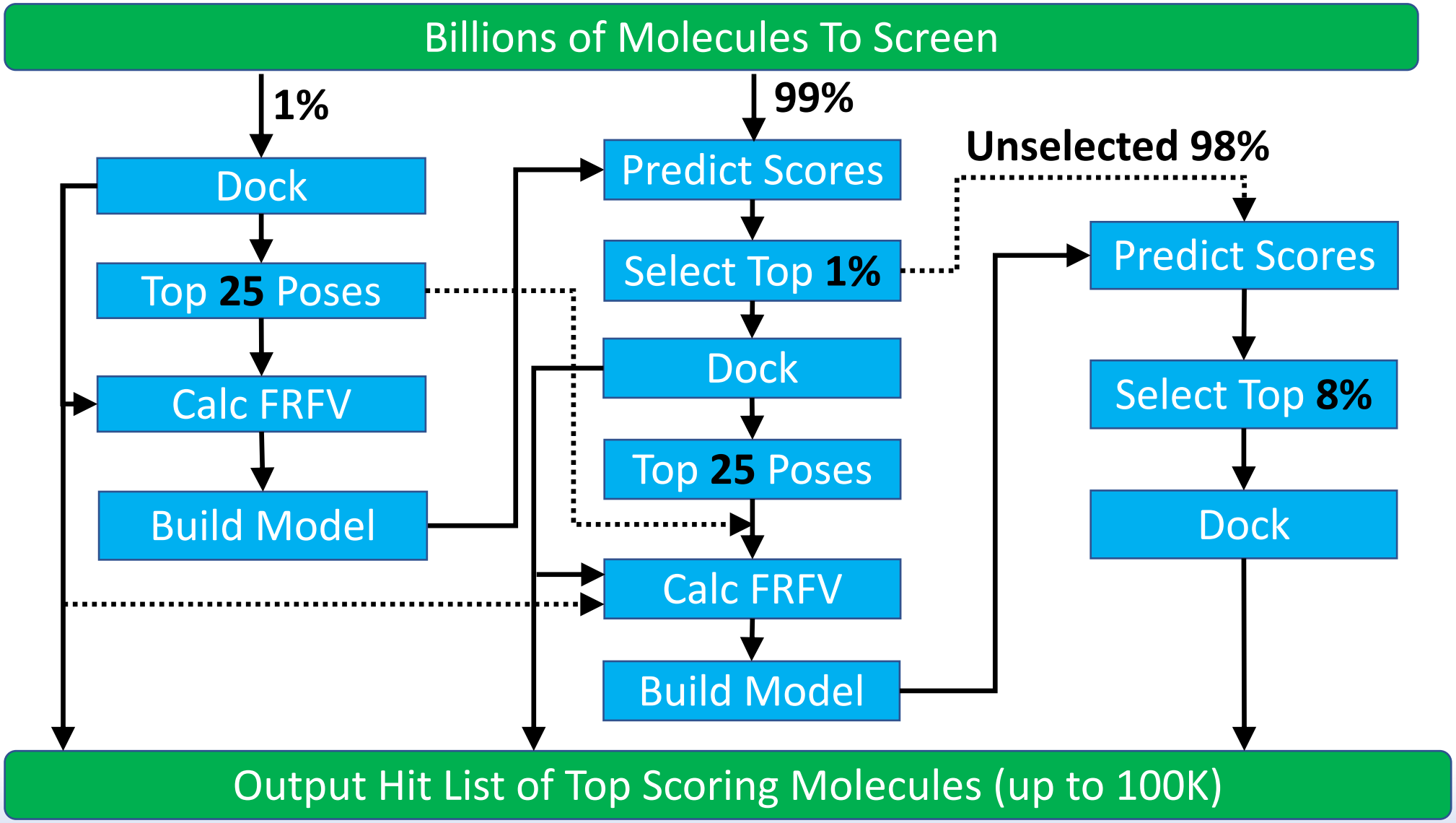


OpenEye SCIENTIFIC
CADENCE MOLECULAR SCIENCES

FastROCS Feature Vector (FRFV) – 1 Stage Model



FastROCS Feature Vector (FRFV) – 2 Stage Model



OpenEye SCIENTIFIC
CADENCE MOLECULAR SCIENCES

Overview

- Current Gigadock Warp
- Upgrading Gigadock Warp
- **Testing Results**

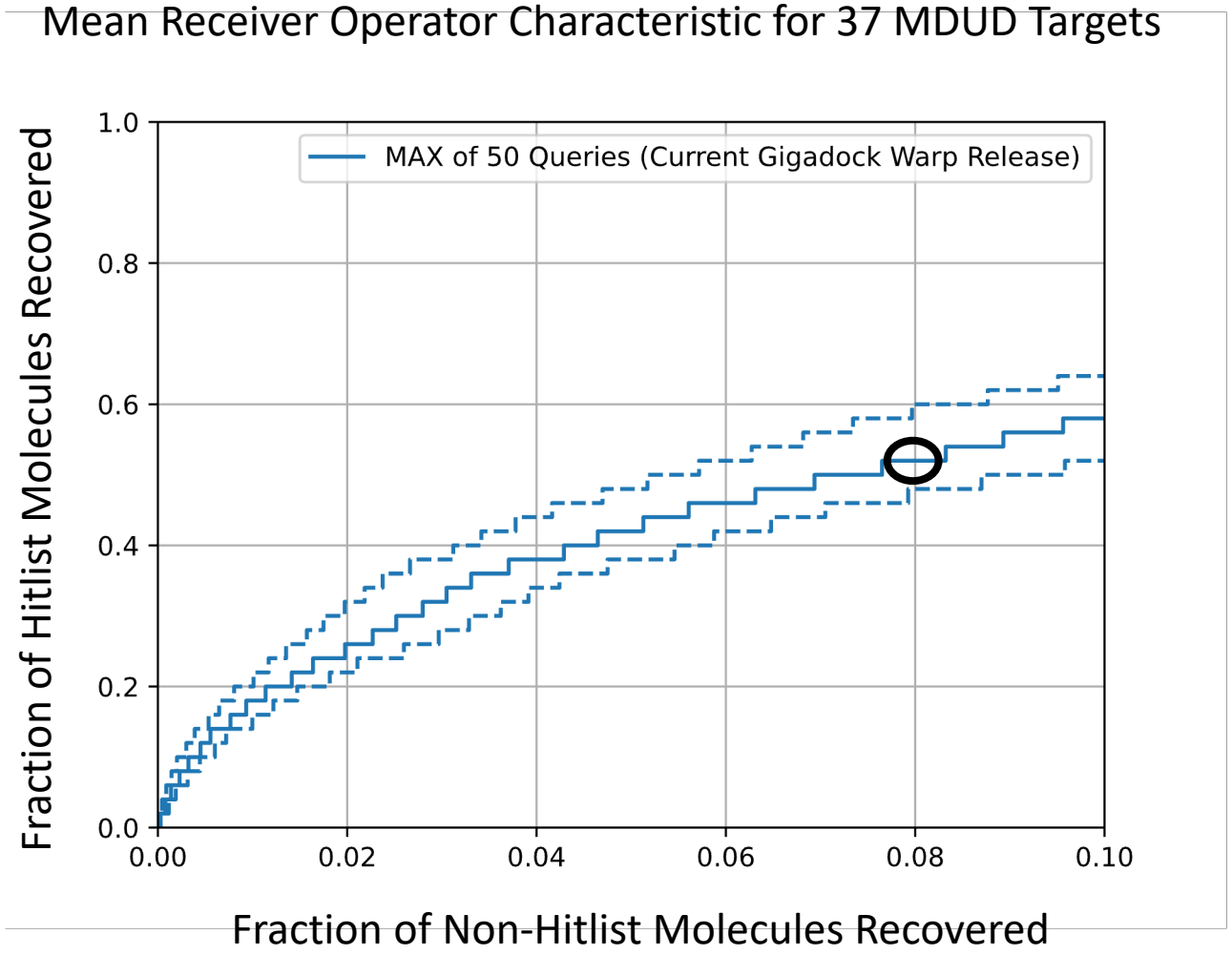
Testing

- Targets – 37 Receptors from MDUD* Dataset
- Molecules – 5 Million Random Enamine Molecules

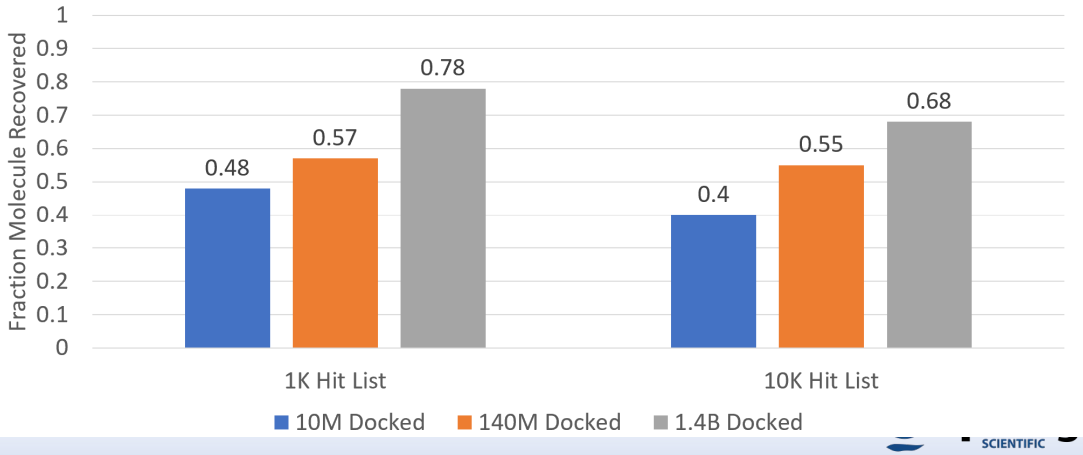
- Analysis
  1. For each target designate 50 Top scoring molecules as ‘hit list’
     - Equivalent to 100K hit list when docking 1 Billion
  2. 1% Test, 99% Training Split
  3. Construct Model with Test Data
     - Linear Regression
  4. Construct receiver operating characteristic curve calculate AUC
     - Molecules that would be in the hit list are ‘actives’
     - Molecules that would not be in the true hit list are ‘inactives’

** J Comput Aided Mol Des. 2012 Aug;26(8):897-906.*

Results with Current Gigadock Warp Settings



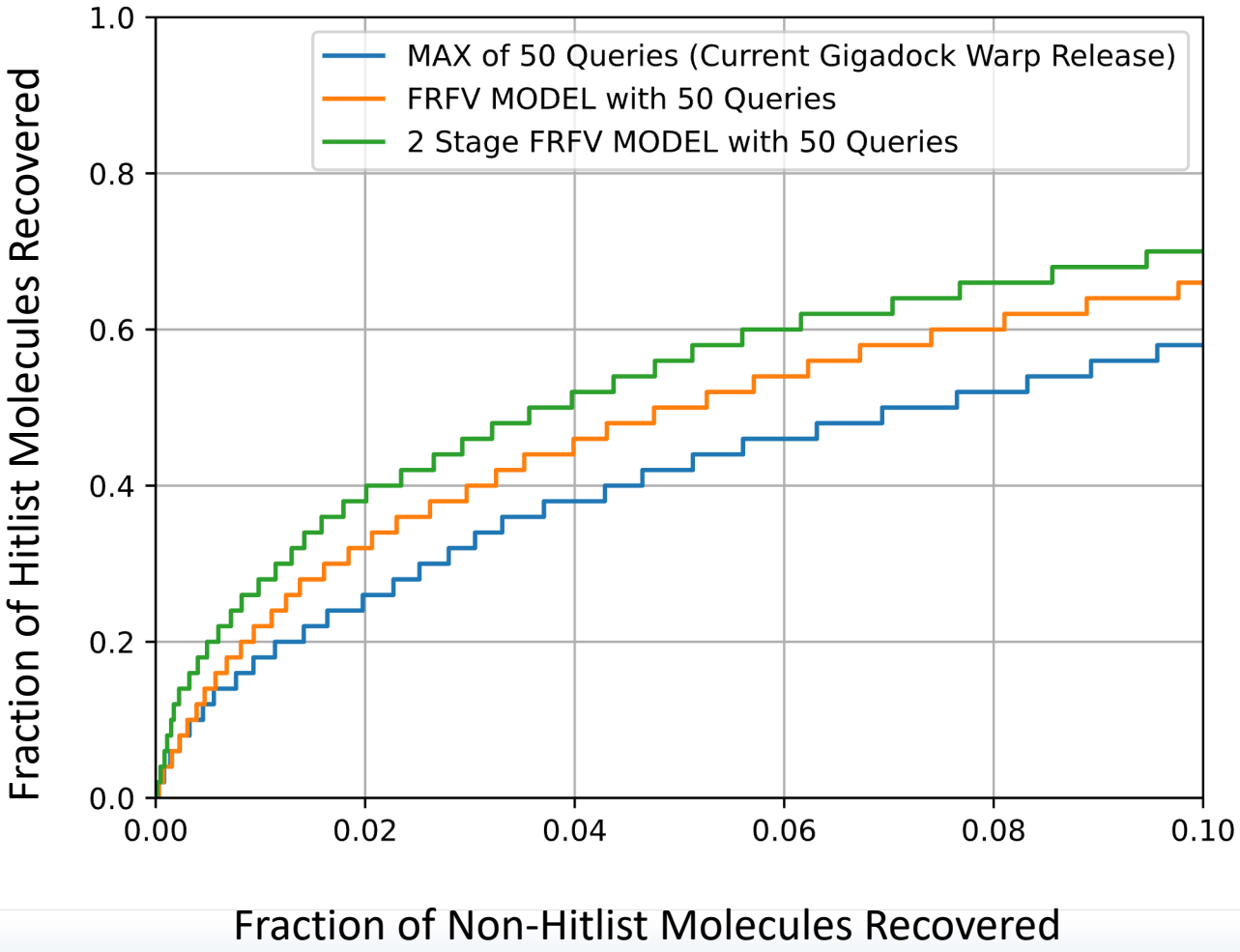
- AUC : 0.86
- 8% ‘Inactives’ → ~50% ‘hit list’
- Previous work indicates performance improves with number of molecules docked



Dashed curves are the upper and lower 95% confidence interval

FastROCS Feature Vectors (FRFV) Models

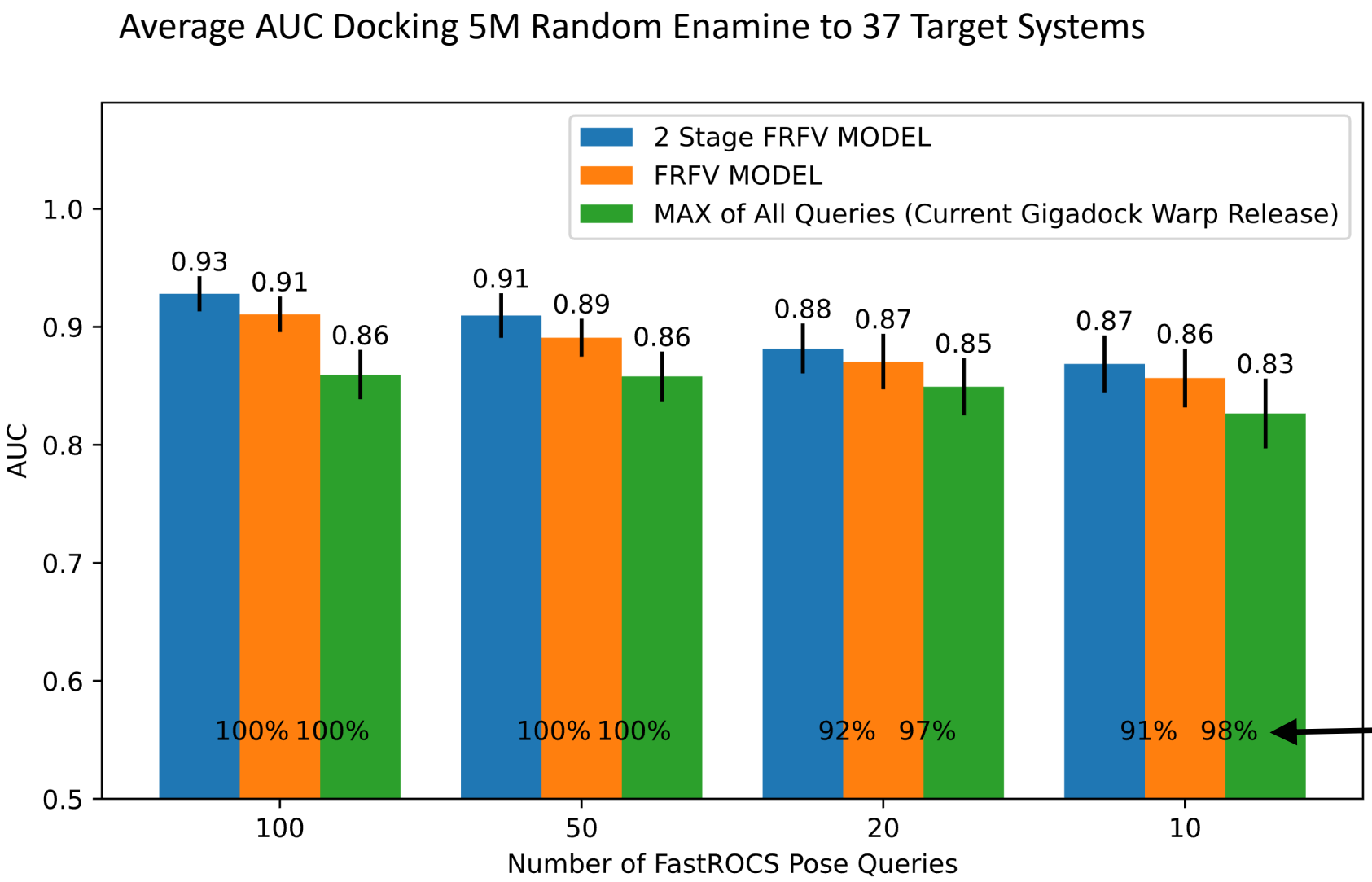
Average Receiver Operator Characteristic Across 37 MDUD Targets



- MAX
  - AUC : 0.86
  - 8% ‘Inactives’ → ~50% ‘hit list’
- MODEL Docking Score
  - AUC : 0.89
  - 8% ‘Inactives’ → ~60% ‘hit list’
- 2 Stage MODEL Docking Score
  - AUC : 0.91
  - 8% ‘Inactives’ → ~65% ‘hit list’

OpenEye SCIENTIFIC
CADENCE MOLECULAR SCIENCES

Number of FastROCS Query Poses vs Performance



- FRFV Models work
- 2 Stage FRFV Models work even better

Probability difference in mean AUC is statistically significant

OpenEye SCIENTIFIC
CADENCE MOLECULAR SCIENCES

2D Models

Graphsim Fingerprints

- Path (4096)
- Tree (4096)
- Circular (4096)
- MACCS166 (166)

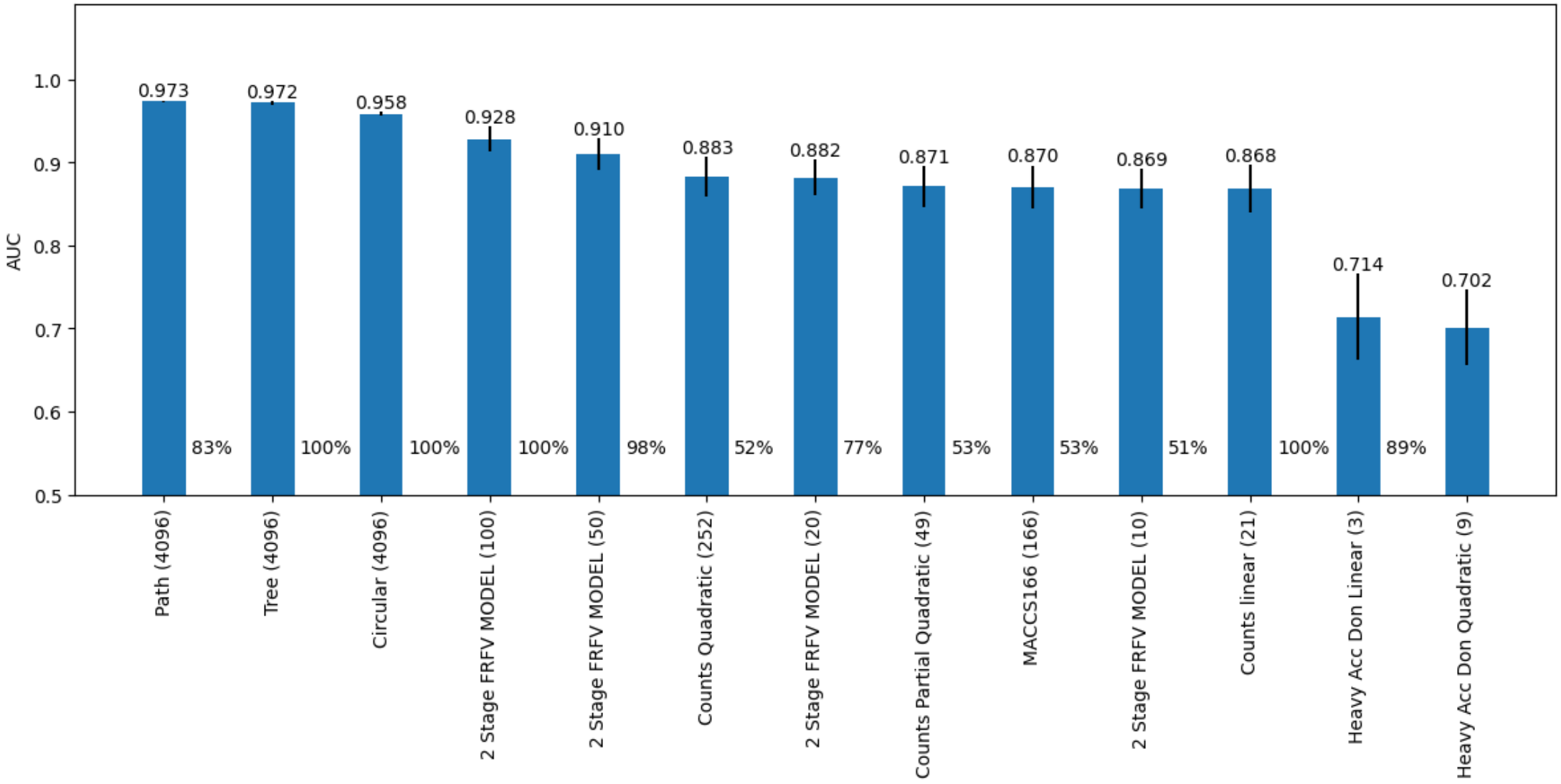
- Number of features in parenthesis
- Feature count matters at scale
  - For 10 Billion docking a 1% training set size is 100 Million

Simple Feature Counts

- Counts Linear (21)
  - Atom Counts : Heavy, Acceptor, Donor, H, C, N, O, F, P, S, Cl, Br, I
  - Bond Counts : Rotatable, All Bonds, Single, Double, Triple, Aromatic
  - Ring Counts : All Rings, Aromatic
- Counts Partial Quadratic (49)
  - Atom Counts : **Heavy, Acceptor, Donor**, H, C, N, O, F, P, S, Cl, Br, I
  - Bond Counts : **Rotatable**, **All Bonds**, Single, Double, Triple, Aromatic
  - Ring Counts **: All Rings, Aromatic**
- Counts Quadratic (252)
  - Atom Counts : **Heavy, Acceptor, Donor, H, C, N, O, F, P, S, Cl, Br, I**
  - Bond Counts : **Rotatable, All Bonds, Single, Double, Triple, Aromatic**
  - Ring Counts : **All Rings, Aromatic**
- Heavy Acc Don Linear (3)
  - Atom Counts : Heavy, Acceptor, Donor
- Heavy Acc Don Quadratic (9)
  - Atom Count : **Heavy, Acceptor, Donor**

Bolded values include squared value & cross terms with other squared values

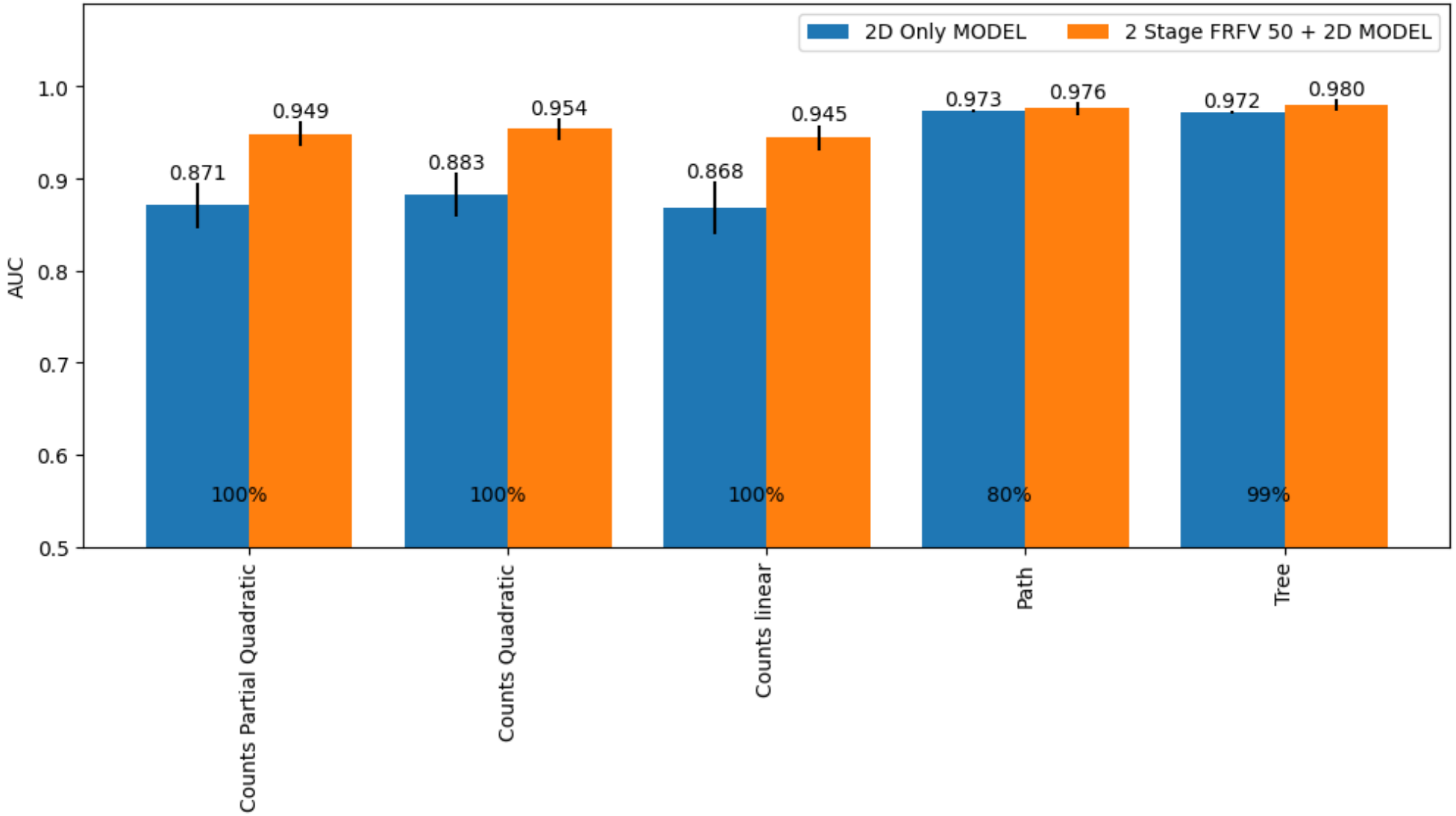
FastROCS Feature Vector Compared to 2D



Type of Feature Modeled

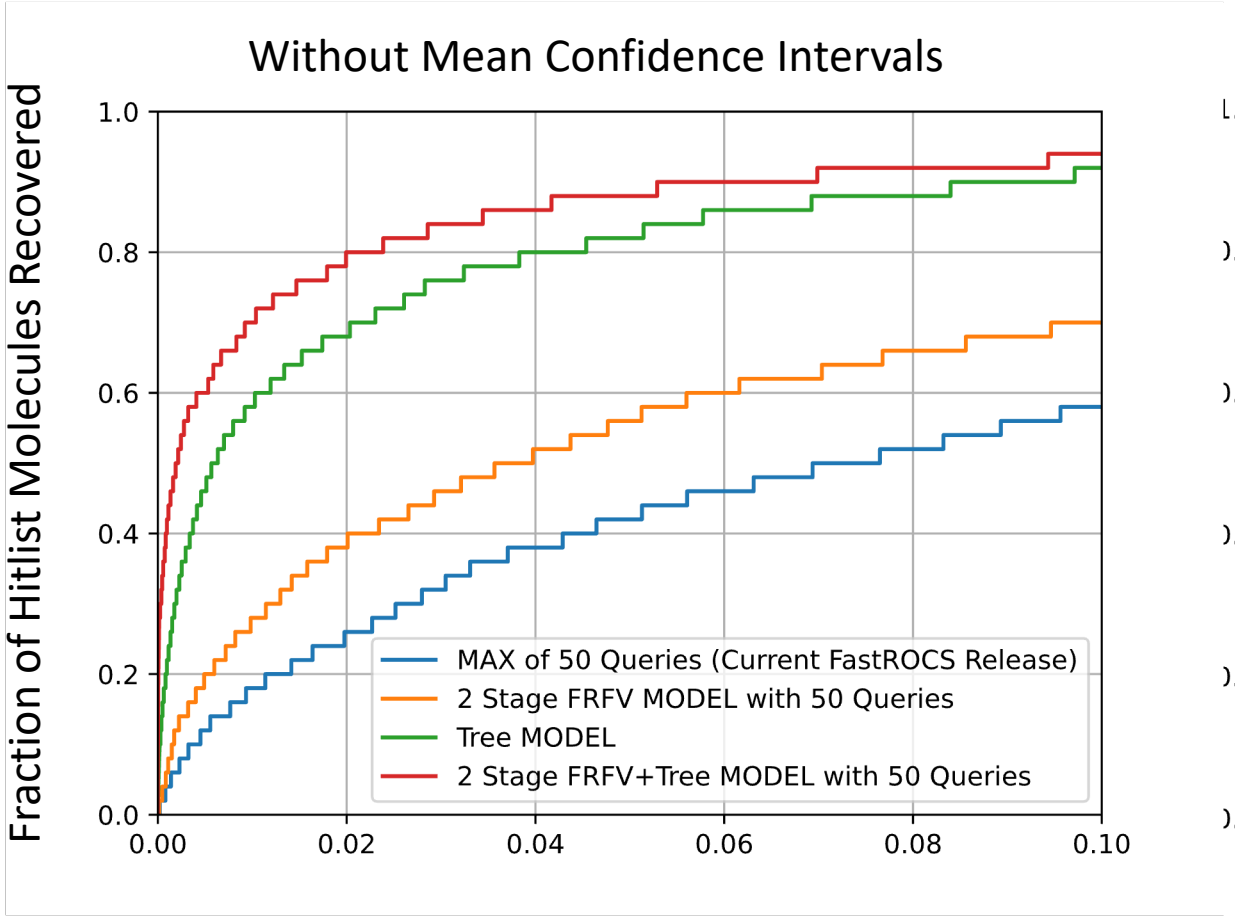
Combining FastROCS Feature Vectors and 2D

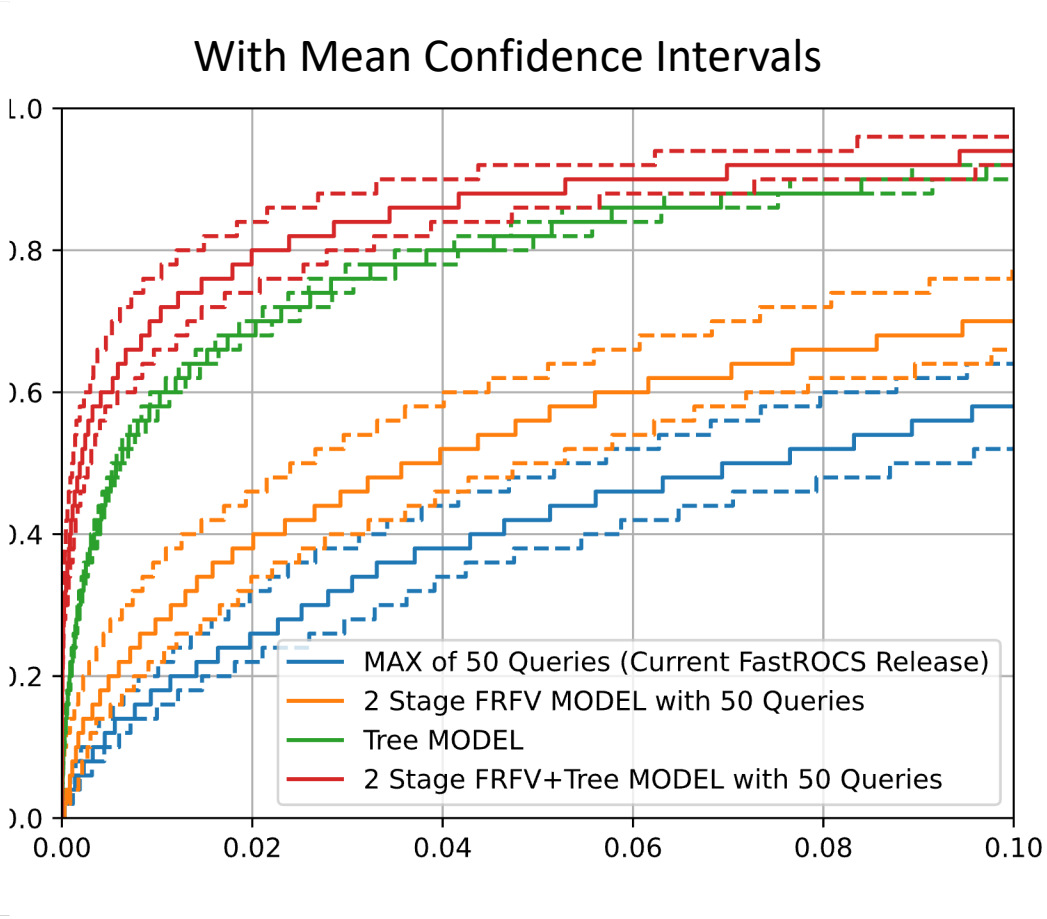
Average Performance Docking 5M Random Enamine to 37 Target Systems



OpenEye SCIENTIFIC
CADENCE MOLECULAR SCIENCES

Receiver Operator Characteristic for FRFV + Tree





FRFV + Tree → Great performance and many features

OpenEye SCIENTIFIC
CADENCE MOLECULAR SCIENCES

Next Steps

- Giga scale performance on multiple target systems
- Alternate Models*
- Multistage optimization with >2 stages
- Good 2D fingerprint with less 4K features?
- Hyperparameter optimization
  - e.g., % initial docked, clustering queries

**See Sayan Mandal's poster*

Conclusions

- FastROCS Feature Vectors (FRFV) work
  - Better that choosing the maximum Tanimoto (current Gigadock Warp)
  - Same compute cost as using maximum Tanimoto

- 4K Graphsim Fingerprints are effective
  - Many features → More difficult to used in models

- Combining FRFV and 2D → better results


Thank You
The End
OpenEye
SCIENTIFIC
CADENCE MOLECULAR SCIENCES