

# CONNECTION OF SMALL EXCESSES IN SUSY SEARCHES USING MACHINE LEARNING

## Outline:

- Intro
- Idea
- Past
- Present
- Future

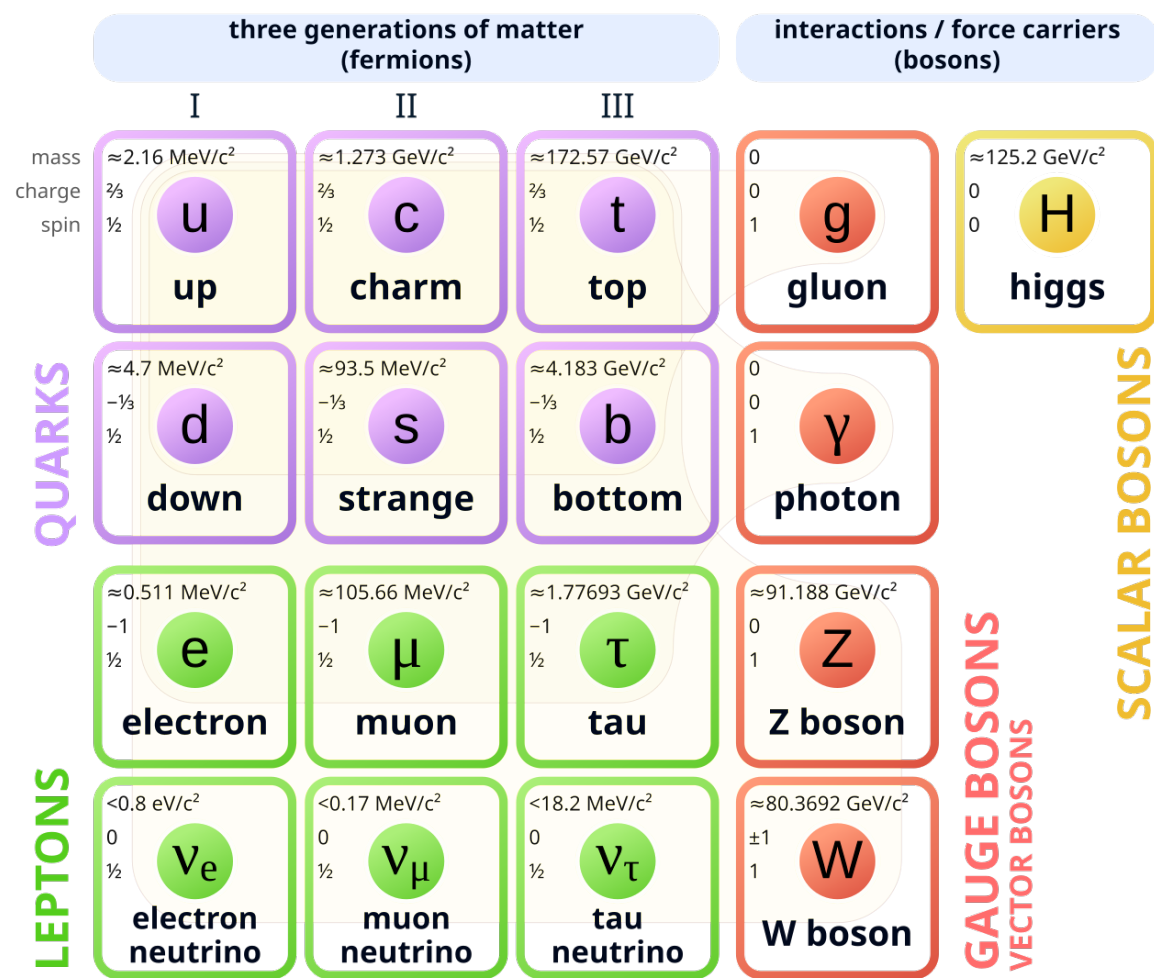
PRELIMINARY

Judita Mamužić  
SMASH Research Fellow  
Jožef Stefan Institute, Slovenia

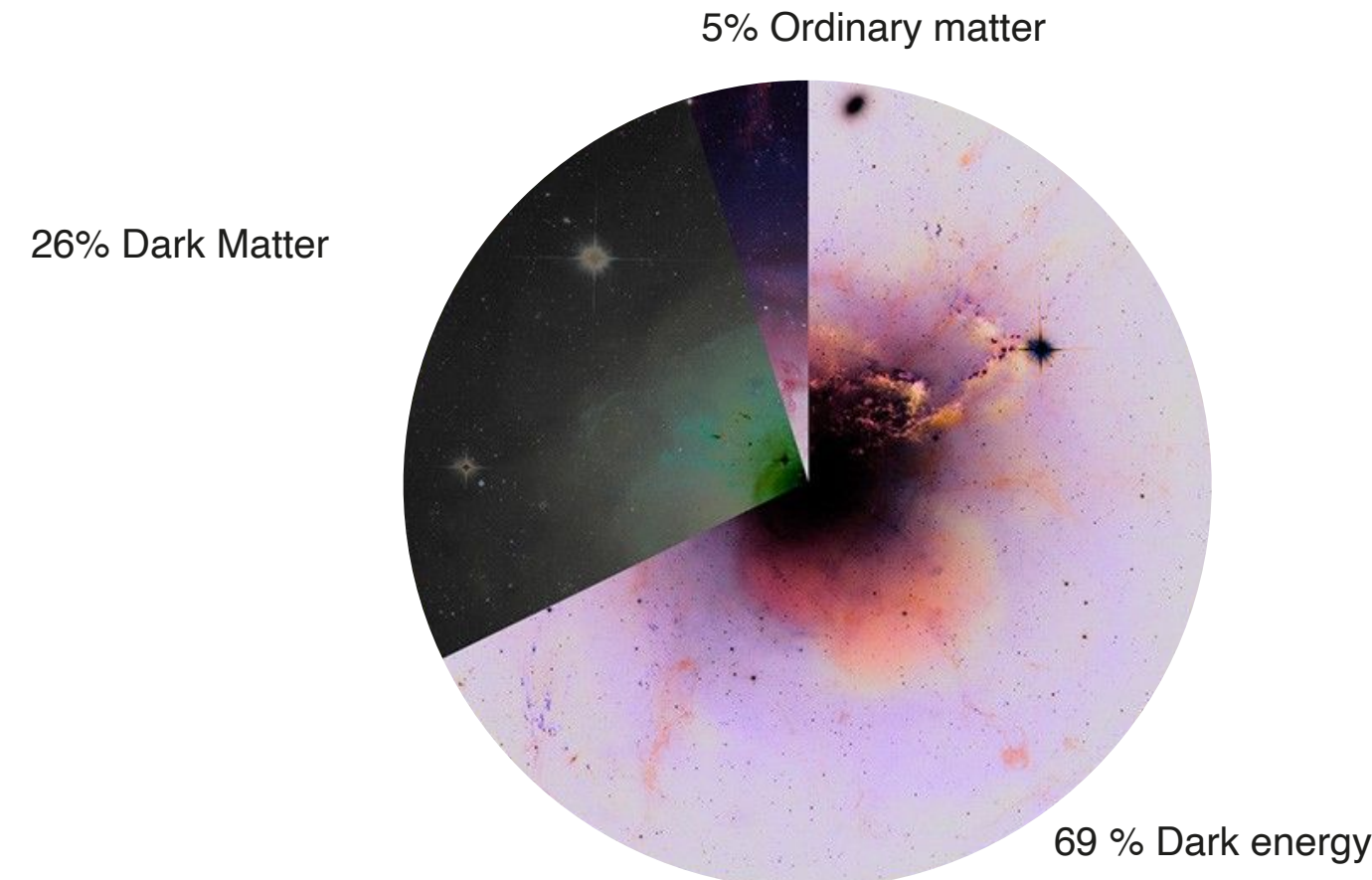


# INTRO: STANDARD MODEL AND PHYSICS BSM

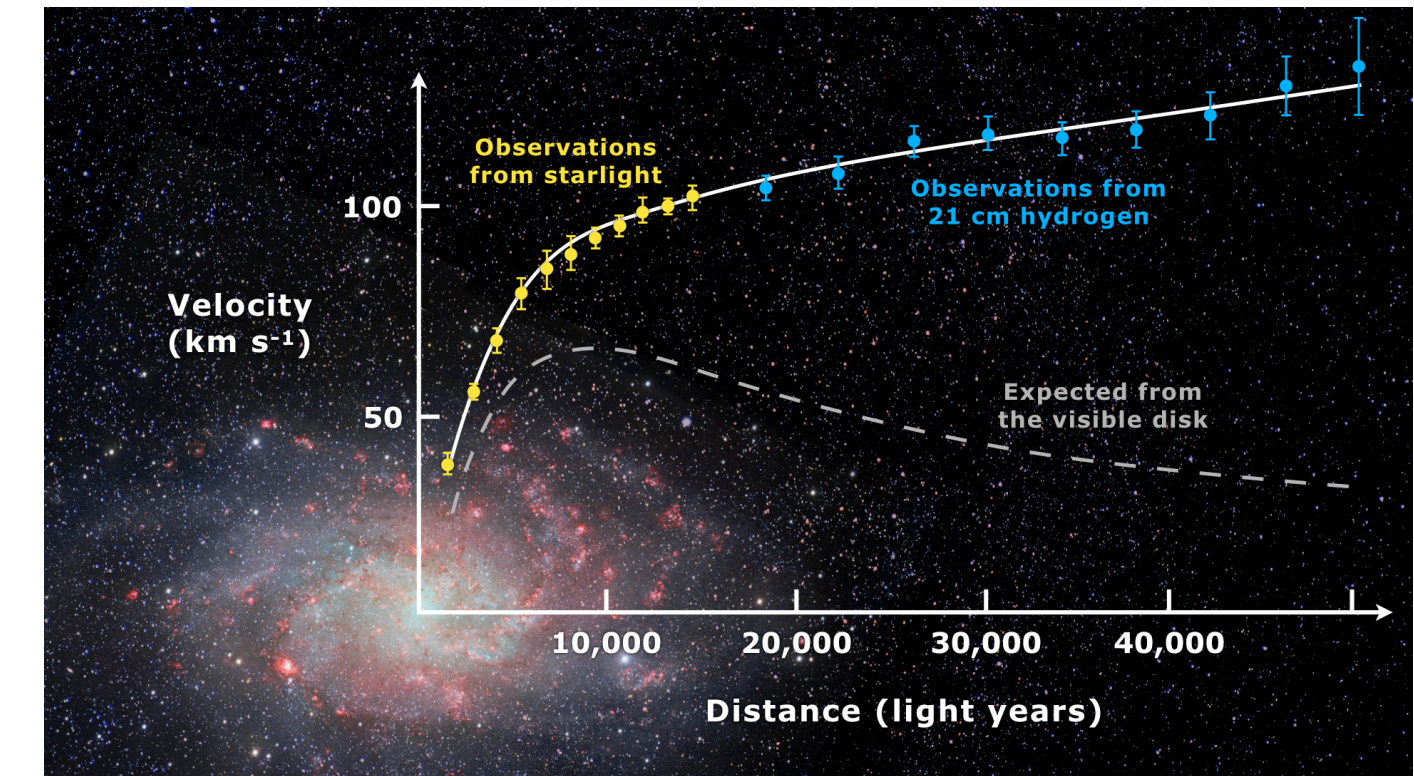
Standard Model of Elementary Particles



## Cosmic microwave background measurements

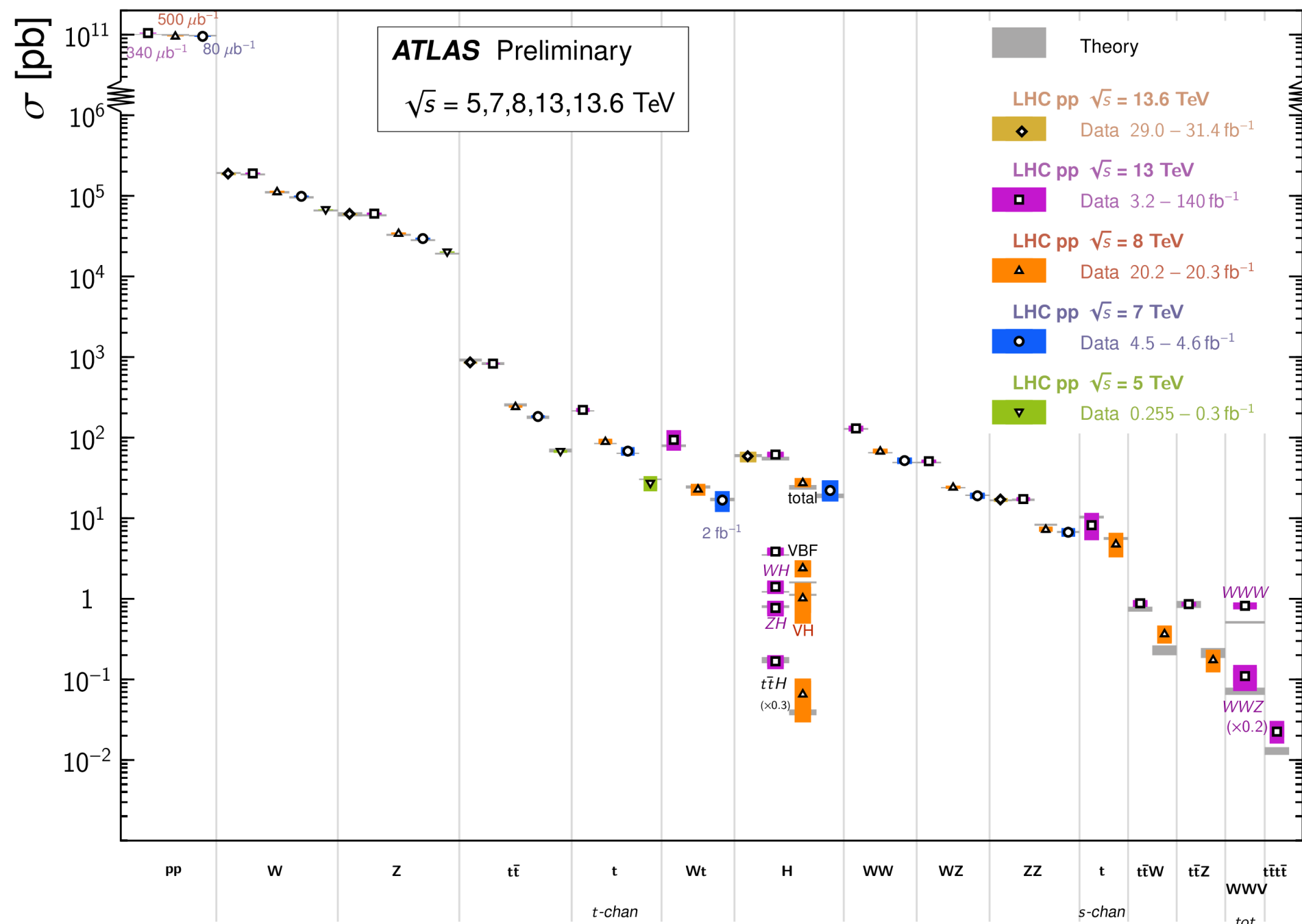


## Galactic rotation curves



Standard Model Total Production Cross Section Measurements

Status: June 2024



[Image modified from NASA Dark Matter](#)

[doi:10.1046/j.1365-8711.2000.03075.x](https://doi.org/10.1046/j.1365-8711.2000.03075.x)

- **Standard Model** describes well the building blocks of matter (quarks and leptons), and their interactions (electromagnetic, weak and strong), with force carriers (photons, gluons and W/Z bosons), acquiring mass via the Braun-Englert-Higgs mechanism
- Good agreement of experimental results with the theory prediction
- A number of phenomena cannot be explained by the SM
- Discrepancies can be explained by the existence of physics **Beyond the Standard Model** (e.g. Dark Matter candidate)

ATL-PHYS-PUB-2024-011



# INTRO: LHC BSM SEARCHES STRATEGY

## Analysis optimization

Full model

Simplified model

## Statistical interpretation

Model independent

Model dependent

$$P(D | B)$$

$$P(D | S + B)$$
$$CL_S = \frac{p_{S+B}}{1 - p_B}$$

How significant is the excess?

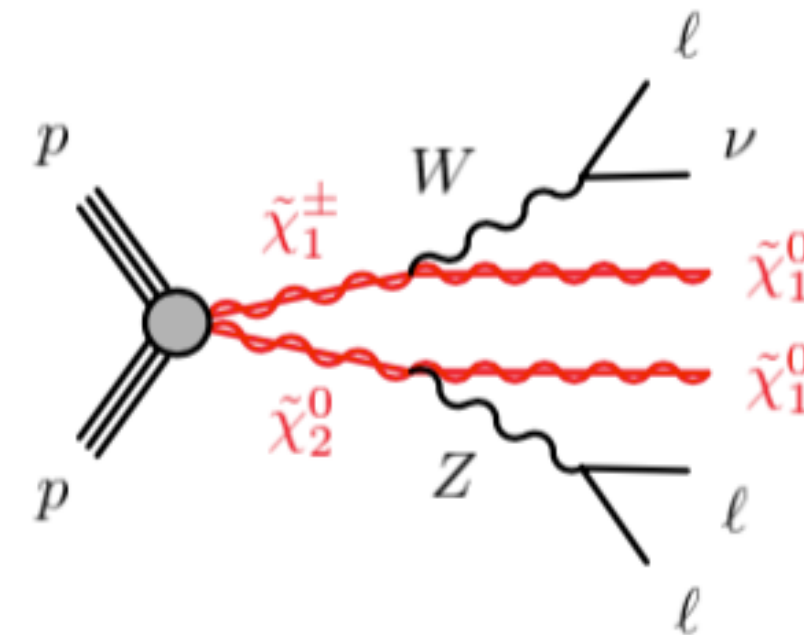
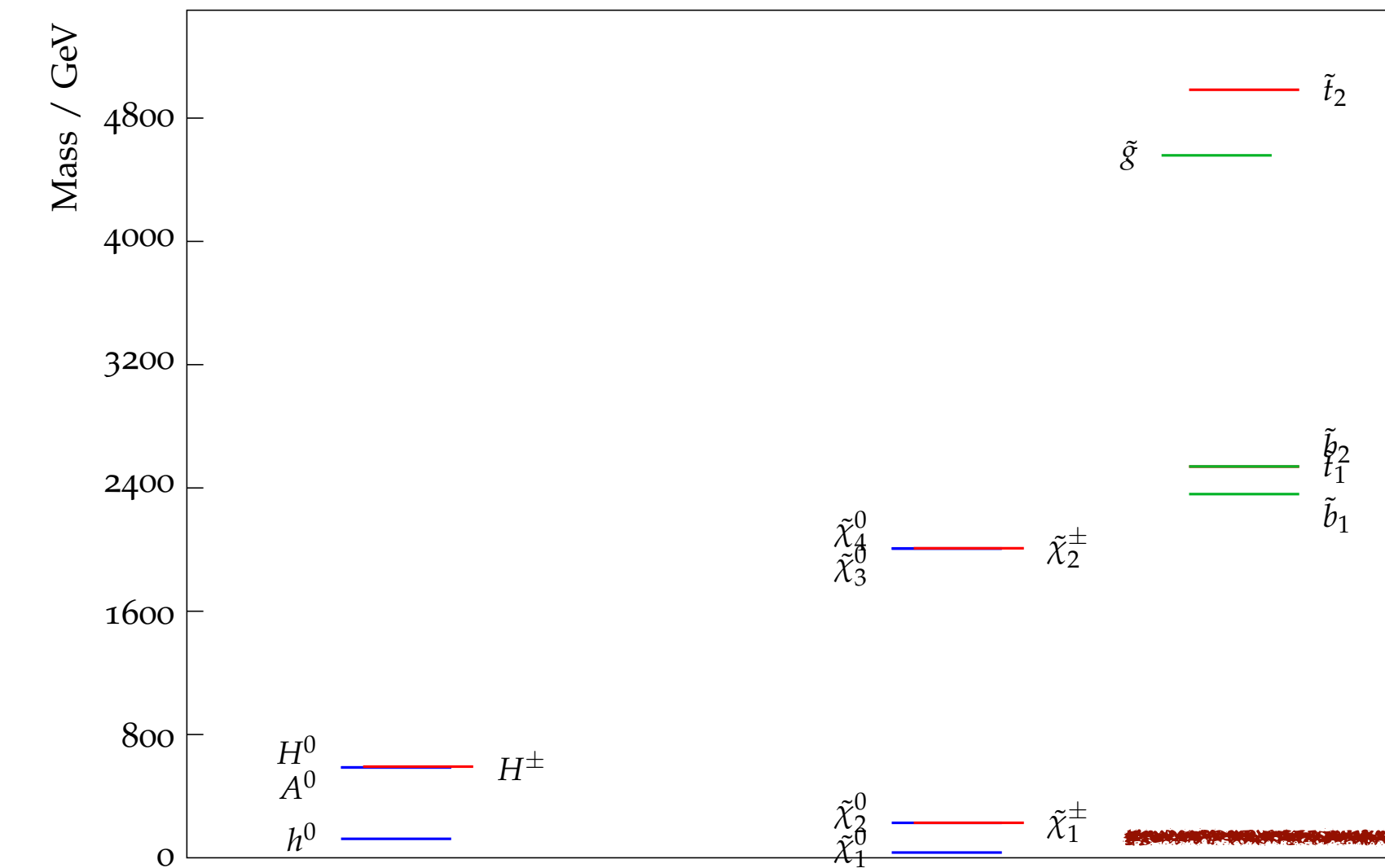
Is the model excluded?

- Rule of thumb in BSM:
  - ~ 3-10 SD: sign of new physics
  - ~ 2 SD: small excess

- Model excluded at 95% CL for  $CLs_{obs} < 0.05$  (exclusion).
- Analyses optimised to maximise parameter space with  $CLs_{exp} < 0.05$  (sensitivity).

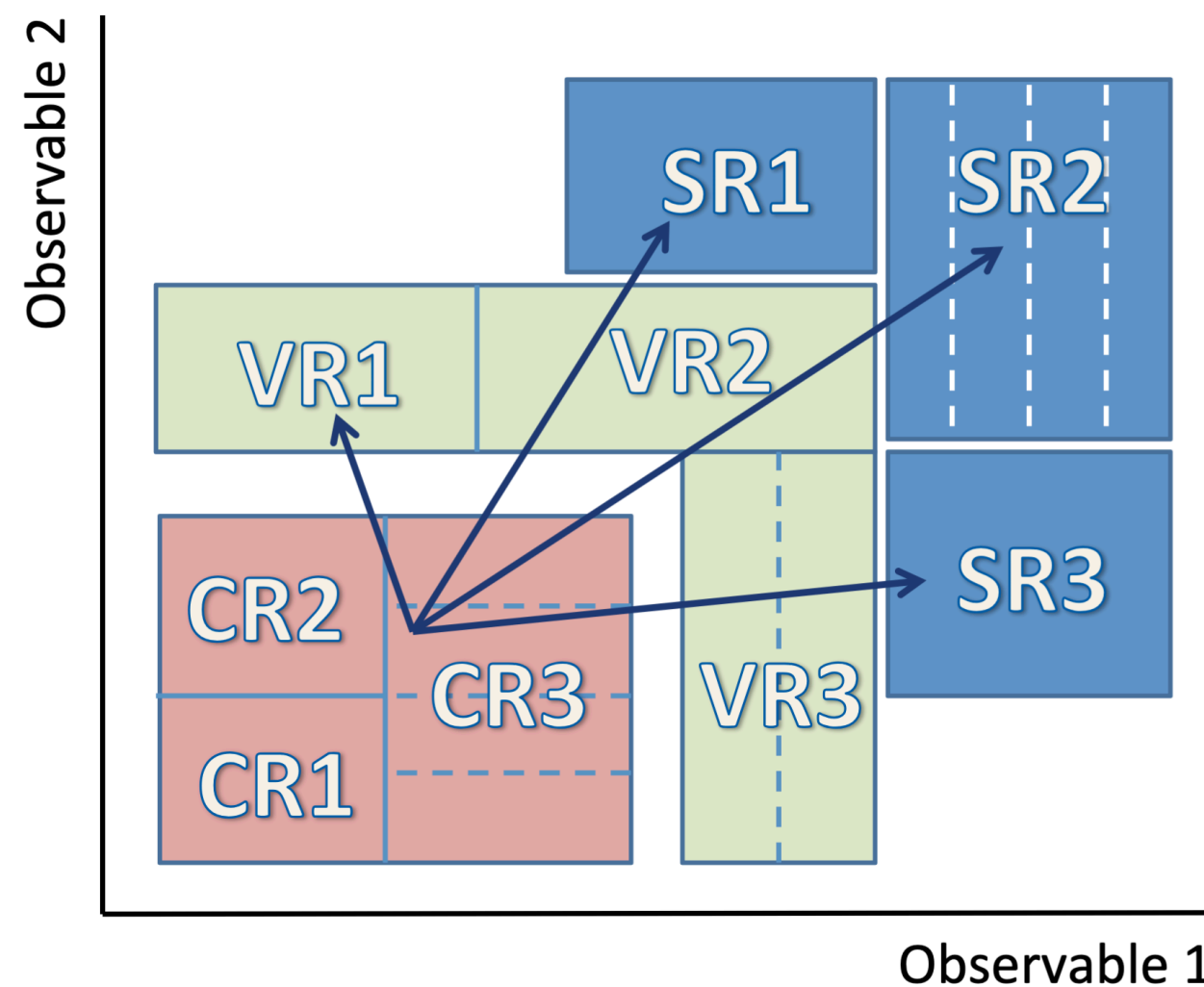
## Strategy

- No favoured **theory** or **mass** of new physics
- Test **any** incarnation of new physics at the LHC
- Impossible to test all models and the parameter space, priority given to feasible searches with best theoretical motivation



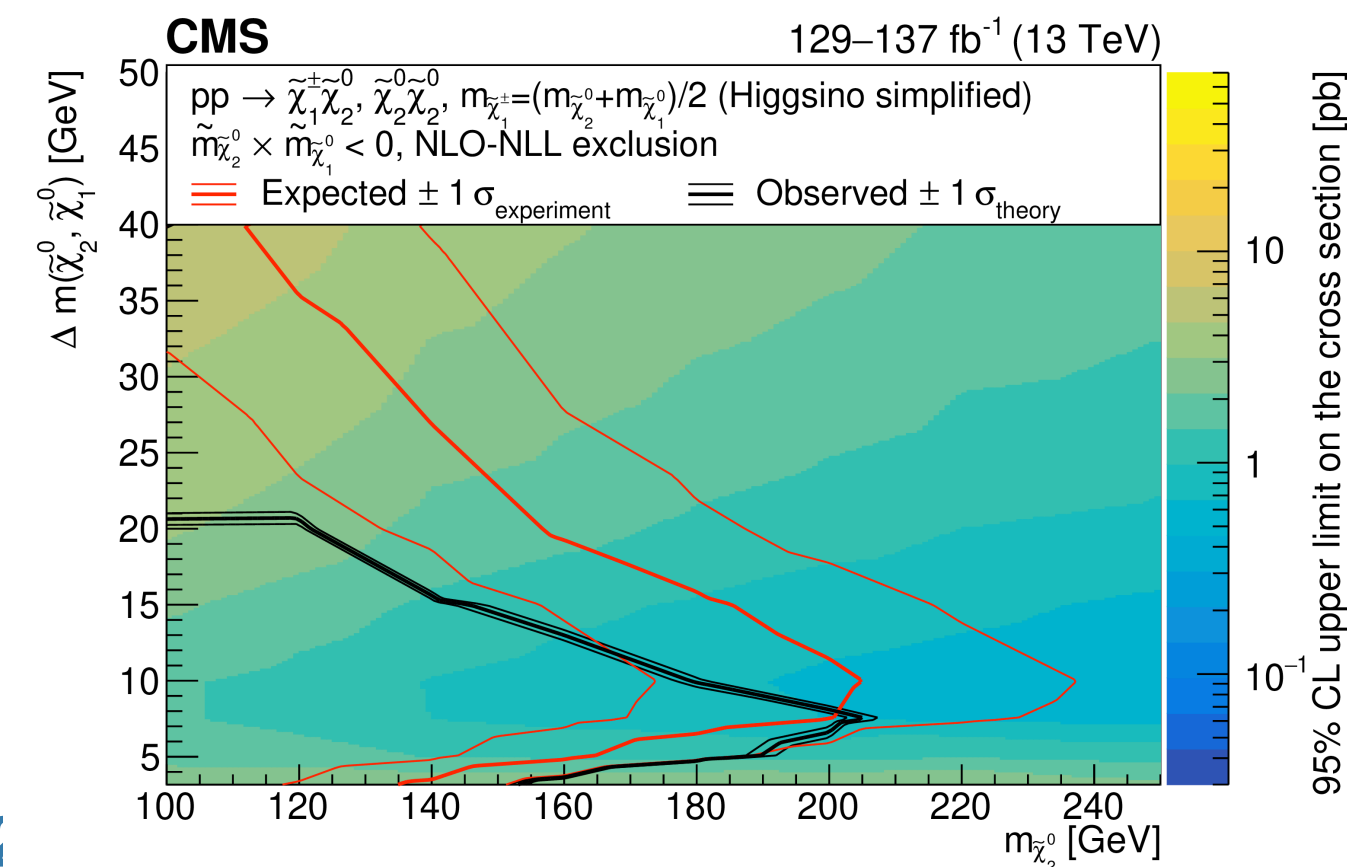
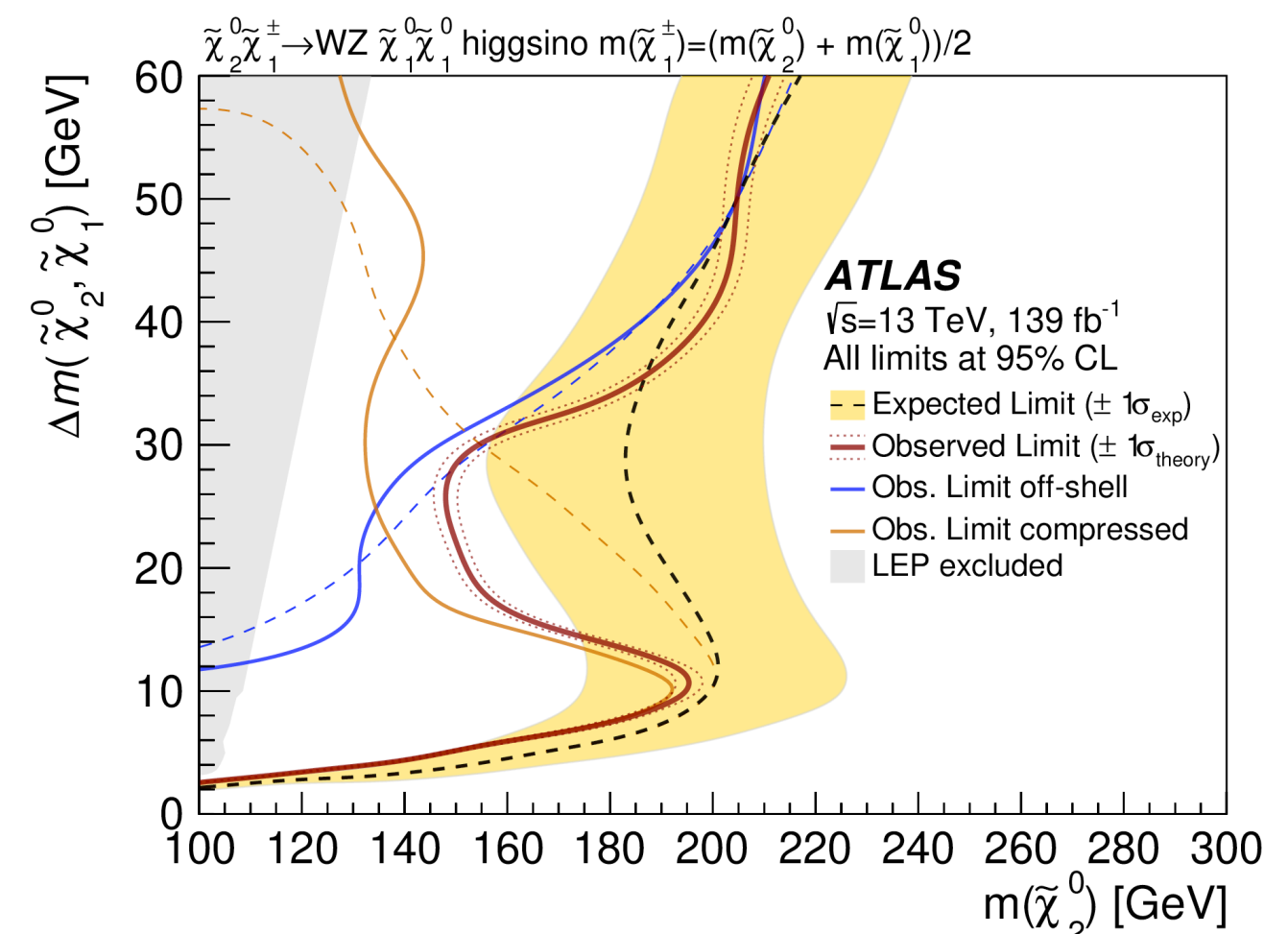
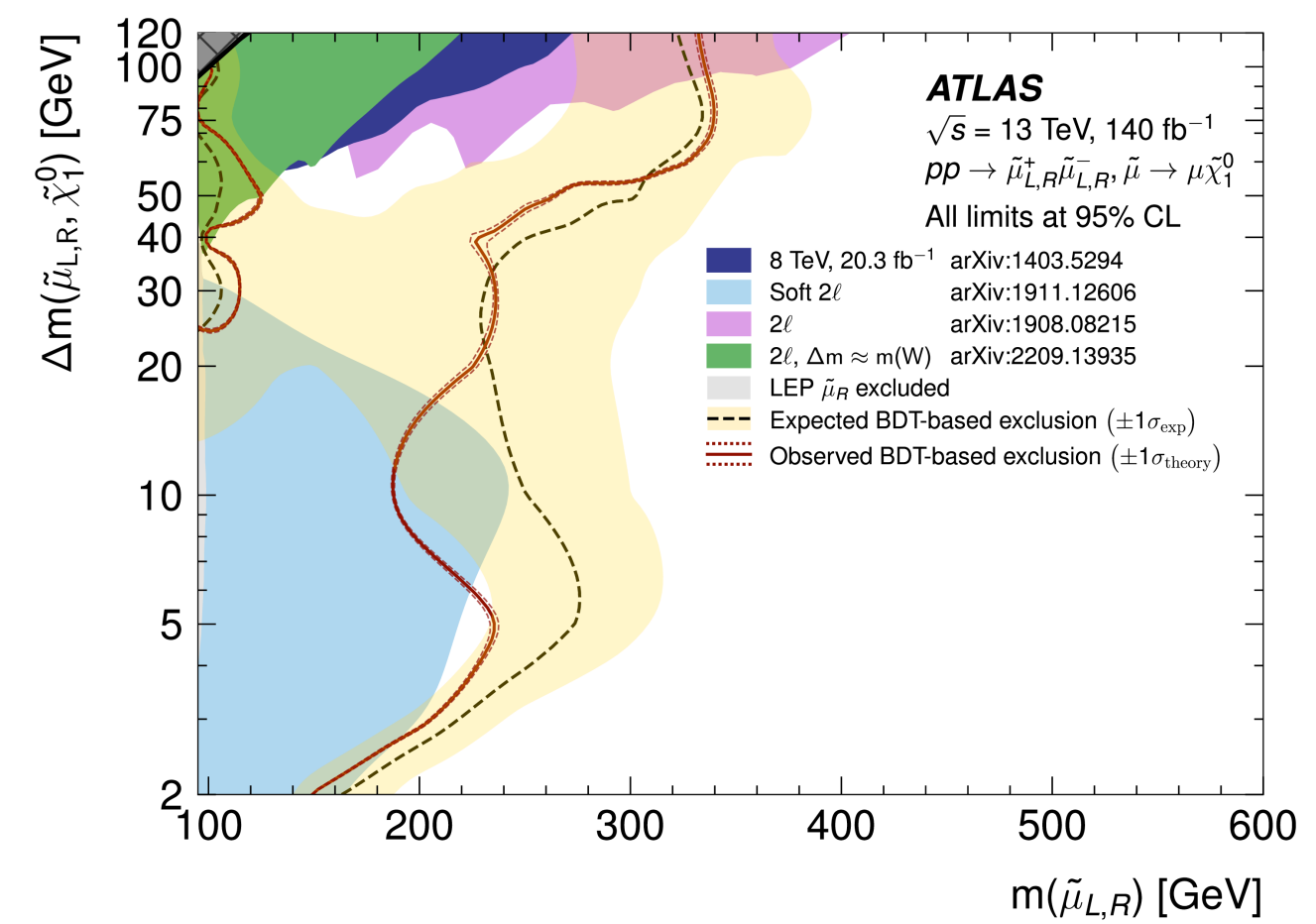
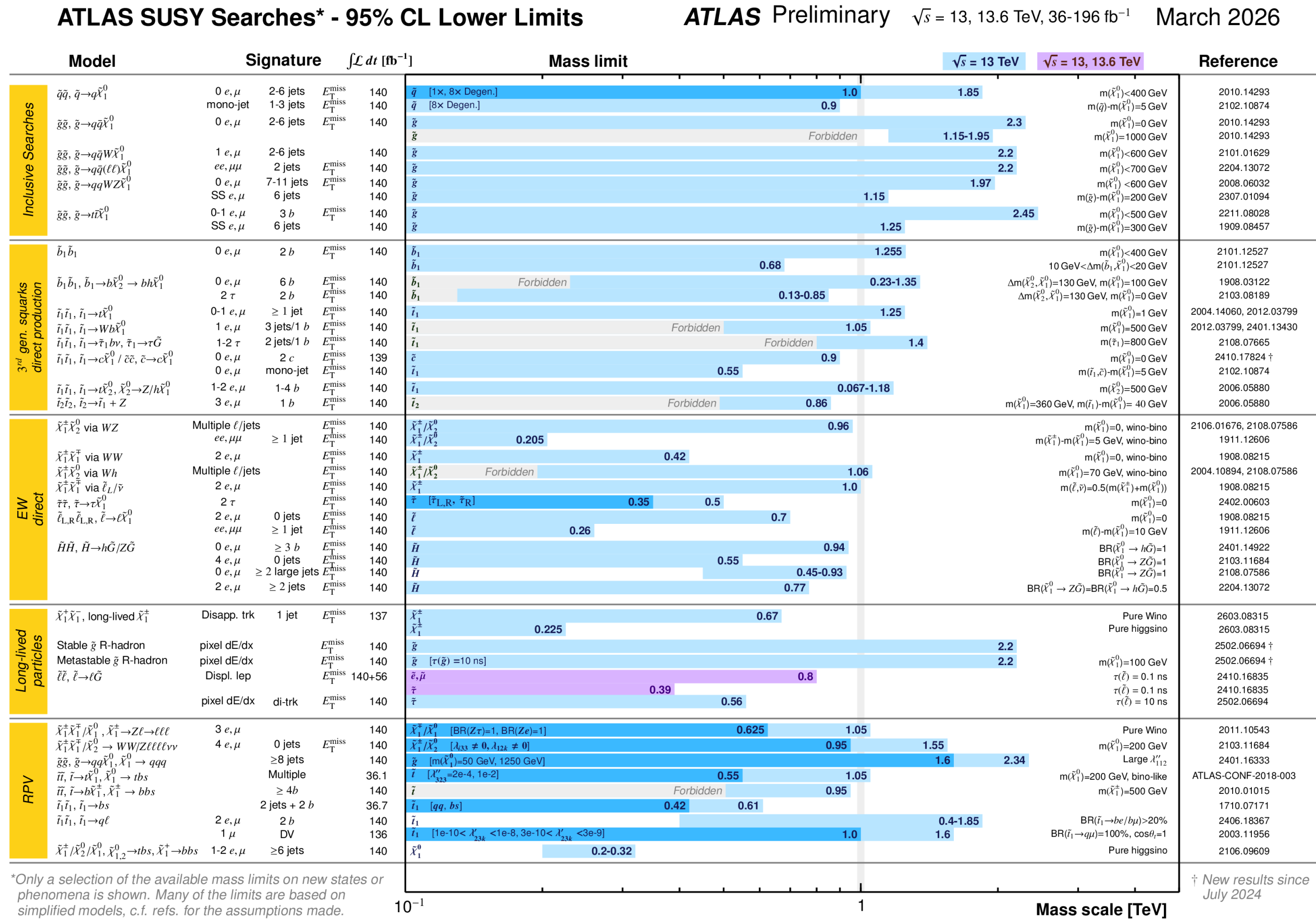
$$\frac{\tilde{\chi}_2^0 / \tilde{\chi}_1^\pm}{\tilde{\chi}_1^0}$$

DOI:10.1088/1742-6596/664/7/072004





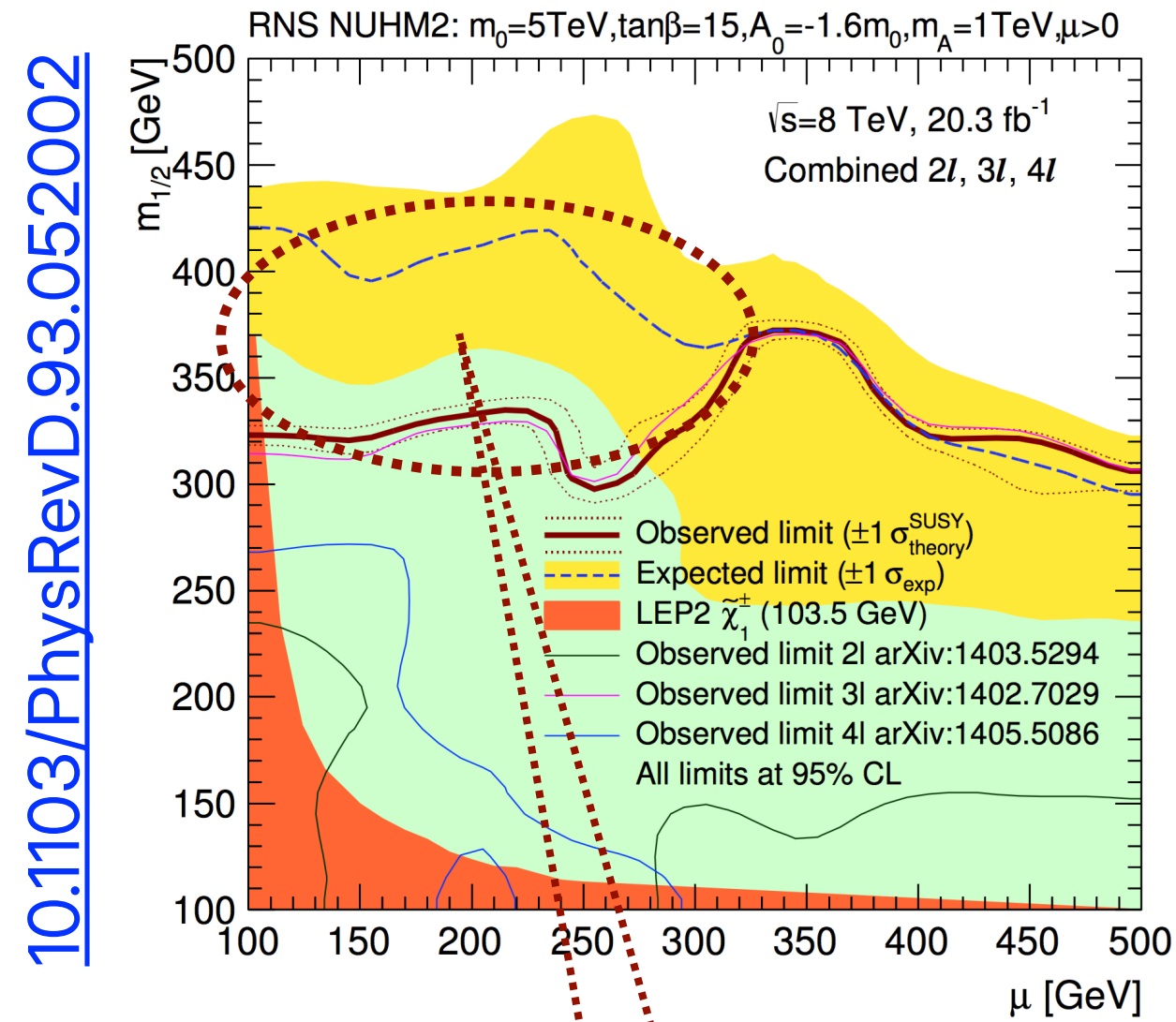
[HMBS-2024-64, JHEP \(2025\)](#)



- In BSM searches in ATLAS or CMS **no significant sign of new physics ( $> 3 \sigma$  SD) was found** to date 😭, exclusion limits set (mapping of excluded parameter space)
- Few analyses have observed small excesses  $\sim 2 \sigma$  SD, [LHC BSM WG](#) 😊
- Model-independent  $2 \sigma$  SD excesses should be considered directly!
- Not all BSM models were tested at  $3 \sigma$  level, discovery at LHC is still viable!



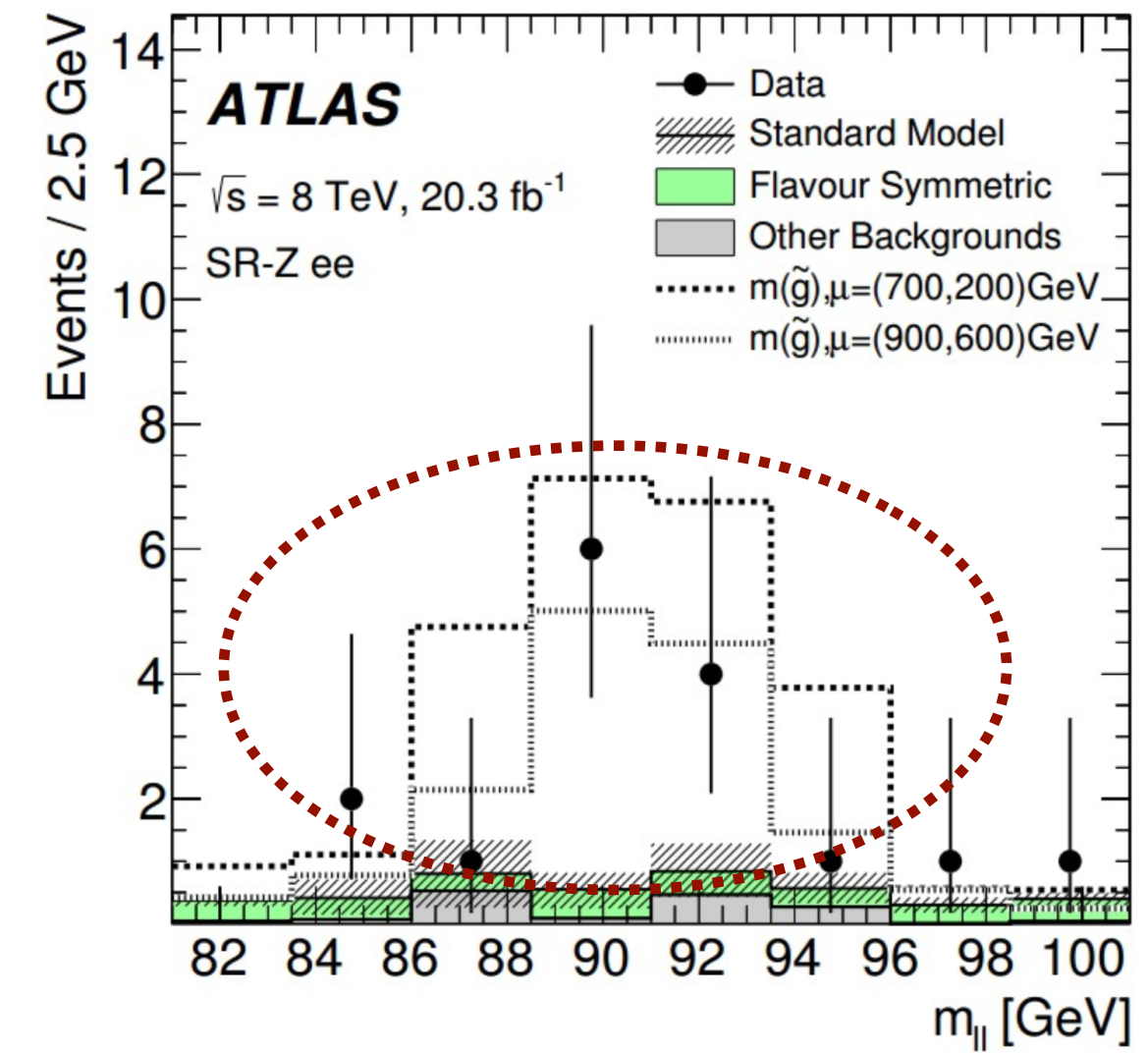
# IDEA: SMALL EXCESSES



Find a **connection** for  
final states and models

Potential new physics:

Good expected sensitivity, but not excluded due to the small excess



- **New type of study:** Look into models that were expected to be excluded, but were not due to small excess (potential new physics).
- Small Excess analysis criteria:  $CL_S^{\text{exp}} \leq 0.05, CL_S^{\text{obs}} > 0.05, (Z > 2\sigma \text{ SD})$
- Study properties of the **models** and **final states** used in the analyses that yield a small excess:
- **Collectively consider small excesses as hints of new physics**
- **Exploit connections of small excesses in BSM searches (systematic study)**

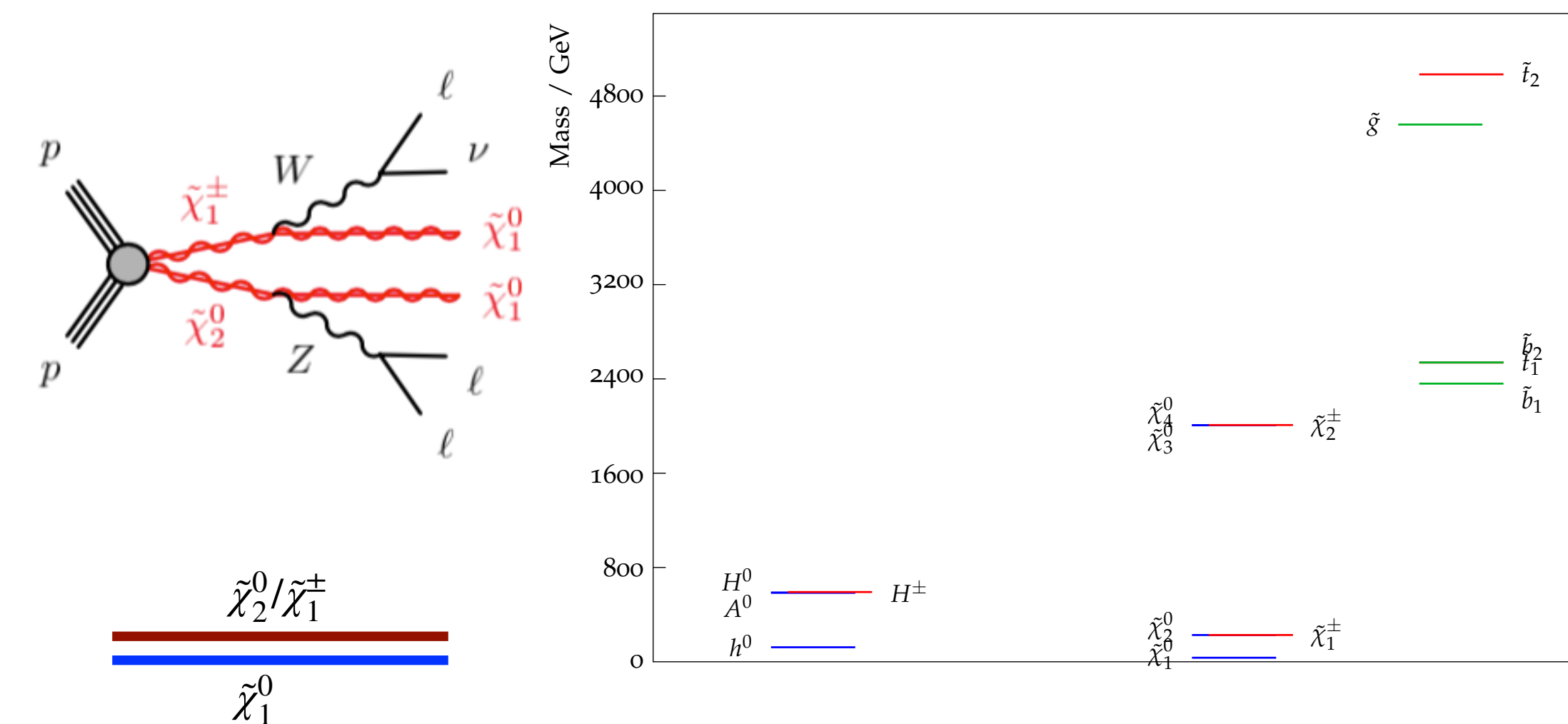


# PAST: LARGE-SCALE REINTERPRETATION

Simplified model

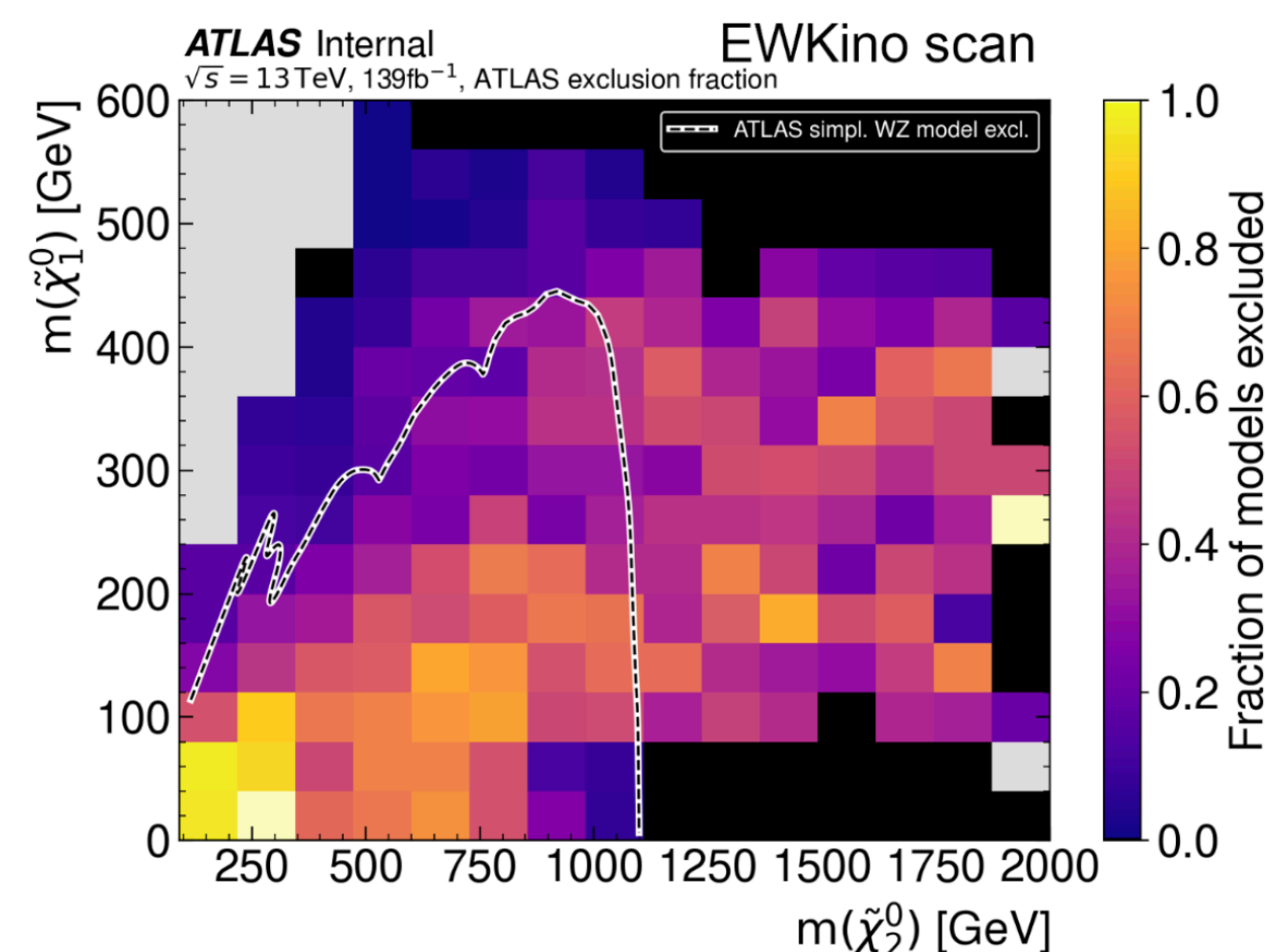
pMSSM model

pMSSM-19 scan



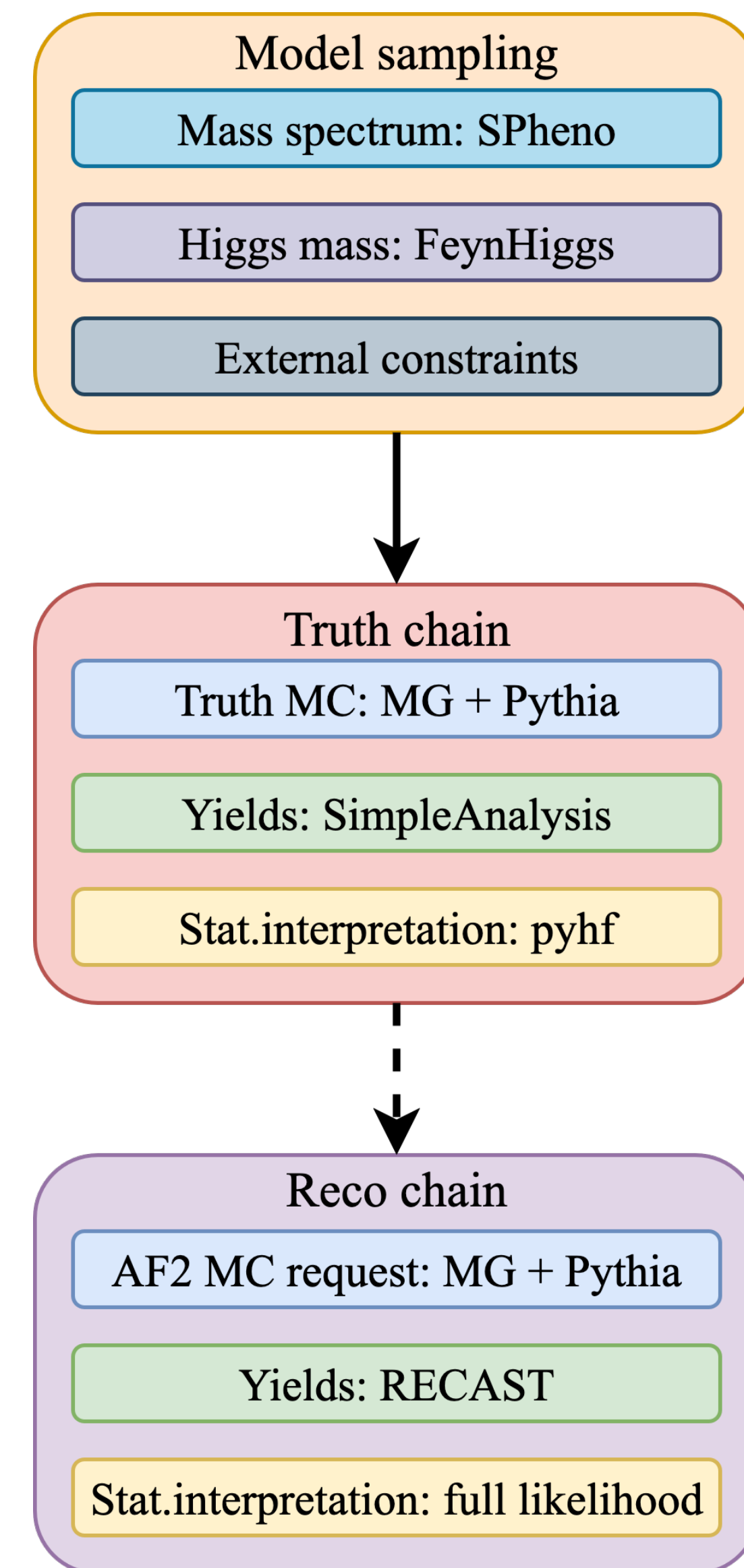
| Parameter                            | min    | max    | Note                                                 |
|--------------------------------------|--------|--------|------------------------------------------------------|
| $M_{\tilde{L}_1} (=M_{\tilde{L}_2})$ | 10 TeV | 10 TeV | Left-handed slepton (first two gens.) mass           |
| $M_{\tilde{e}_1} (=M_{\tilde{e}_2})$ | 10 TeV | 10 TeV | Right-handed slepton (first two gens.) mass          |
| $M_{\tilde{L}_3}$                    | 10 TeV | 10 TeV | Left-handed stau doublet mass                        |
| $M_{\tilde{e}_3}$                    | 10 TeV | 10 TeV | Right-handed stau mass                               |
| $M_{\tilde{Q}_1} (=M_{\tilde{Q}_2})$ | 10 TeV | 10 TeV | Left-handed squark (first two gens.) mass            |
| $M_{\tilde{u}_1} (=M_{\tilde{u}_2})$ | 10 TeV | 10 TeV | Right-handed up-type squark (first two gens.) mass   |
| $M_{\tilde{d}_1} (=M_{\tilde{d}_2})$ | 10 TeV | 10 TeV | Right-handed down-type squark (first two gens.) mass |
| $M_{\tilde{Q}_3}$                    | 2 TeV  | 5 TeV  | Left-handed squark (third gen.) mass                 |
| $M_{\tilde{u}_3}$                    | 2 TeV  | 5 TeV  | Right-handed top squark mass                         |
| $M_{\tilde{d}_3}$                    | 2 TeV  | 5 TeV  | Right-handed bottom squark mass                      |
| $M_1$                                | -2 TeV | 2 TeV  | Bino mass parameter                                  |
| $M_2$                                | -2 TeV | 2 TeV  | Wino mass parameter                                  |
| $\mu$                                | -2 TeV | 2 TeV  | Bilinear Higgs mass parameter                        |
| $M_3$                                | 1 TeV  | 5 TeV  | Gluino mass parameter                                |
| $A_t$                                | -8 TeV | 8 TeV  | Trilinear top coupling                               |
| $A_b$                                | -2 TeV | 2 TeV  | Trilinear bottom coupling                            |
| $A_\tau$                             | -2 TeV | 2 TeV  | Trilinear $\tau$ lepton coupling                     |
| $M_A$                                | 0 TeV  | 5 TeV  | Pseudoscalar Higgs boson mass                        |
| $\tan \beta$                         | 1      | 60     | Ratio of the Higgs vacuum expectation values         |

ATLAS pMSSM EWK,  
20k models



SUSY-2020-15, JHEP 2024

pMSSMFactory



- Re-interpret SUSY analyses using full models.
- Use the Phenomenological Minimal SUSY (pMSSM) [1],[2] with 19 parameters at the high mass scale, imposing theoretical and experimental constraints to reduce the number of parameters.
- Generate large number of models, explore high number of final states using existing analyses.
- Points to uncovered regions to which no sensitivity was obtained, design new round of searches

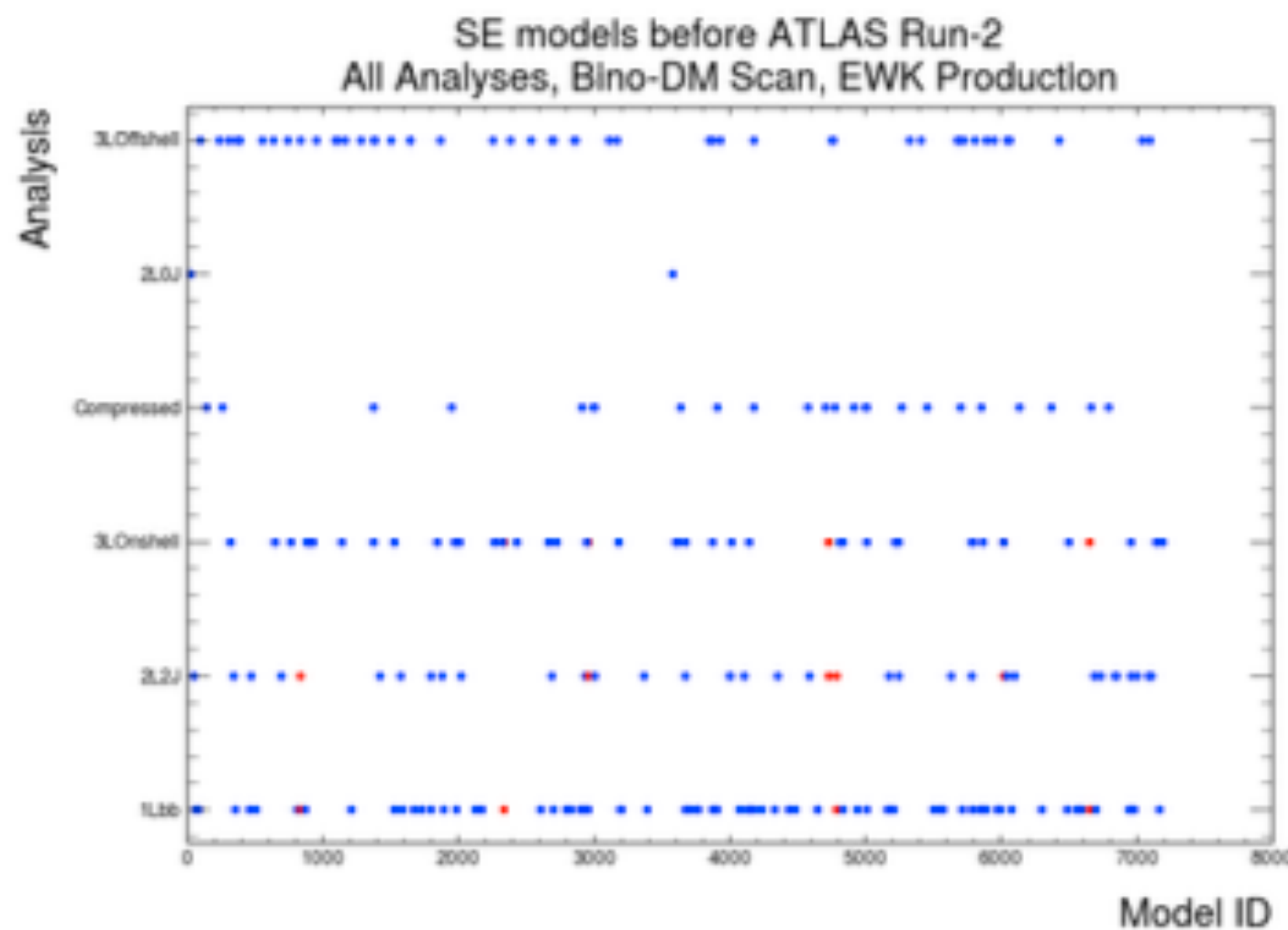
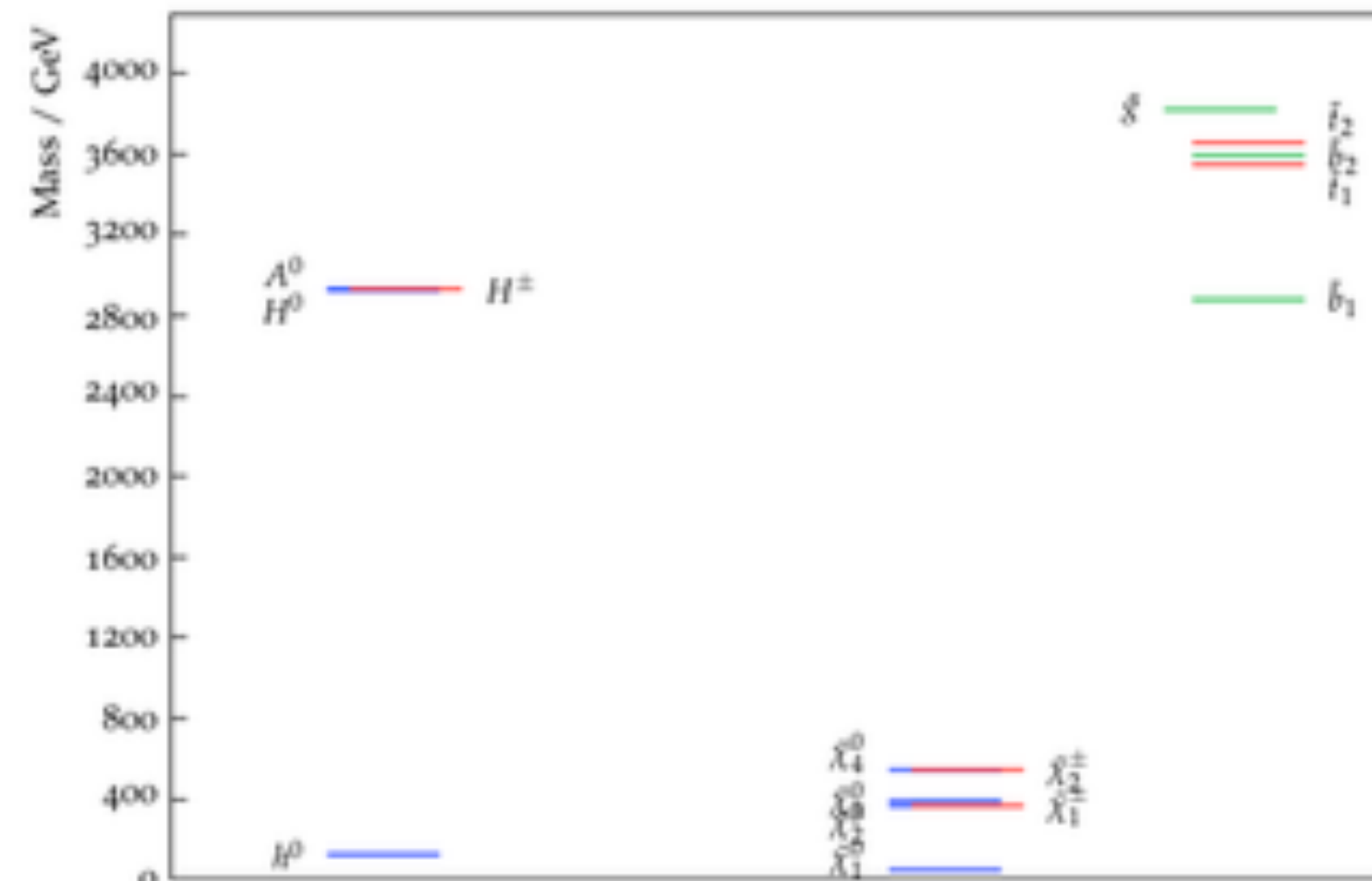


# PAST: ANALYTIC SCAN OF SE

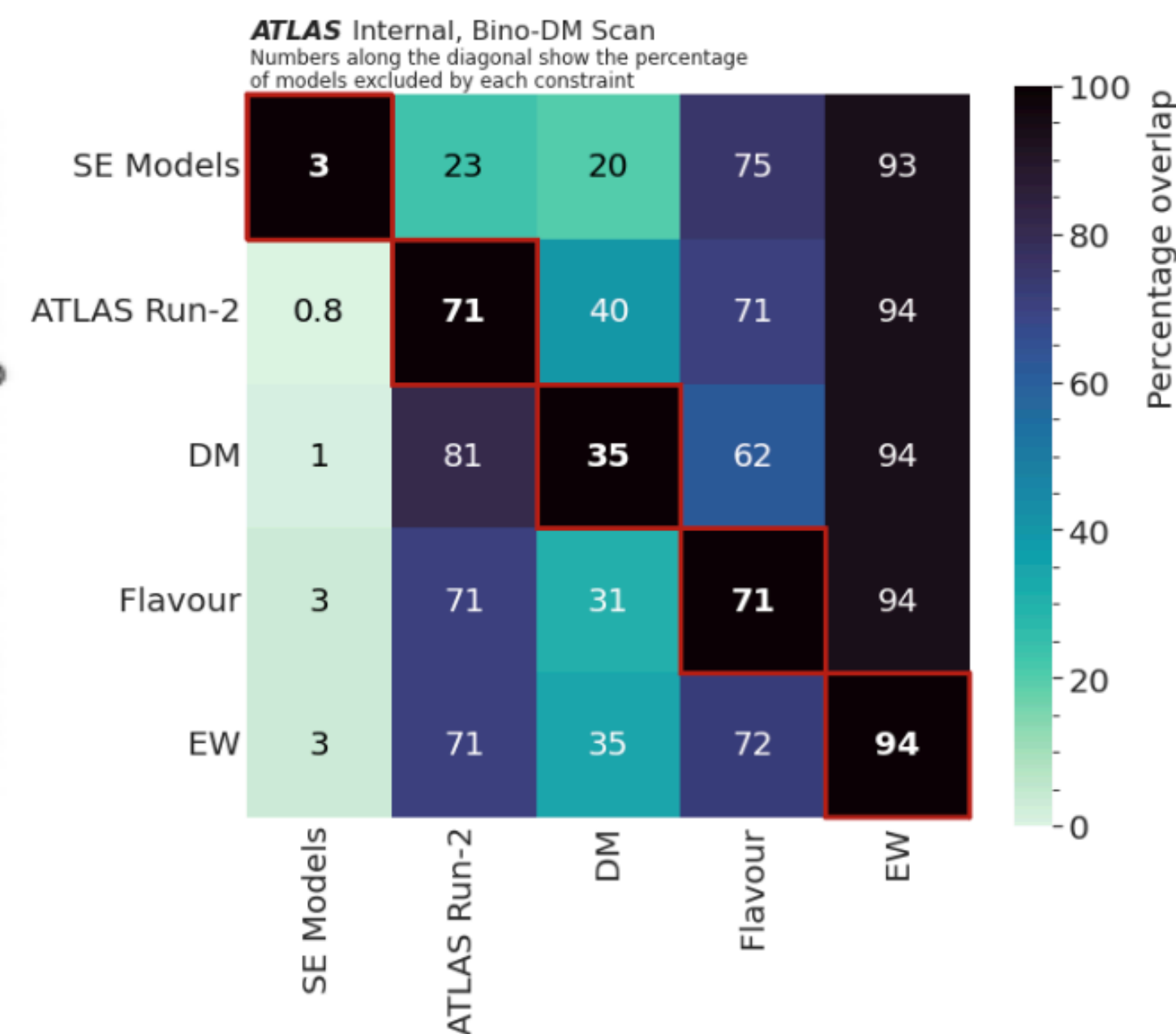
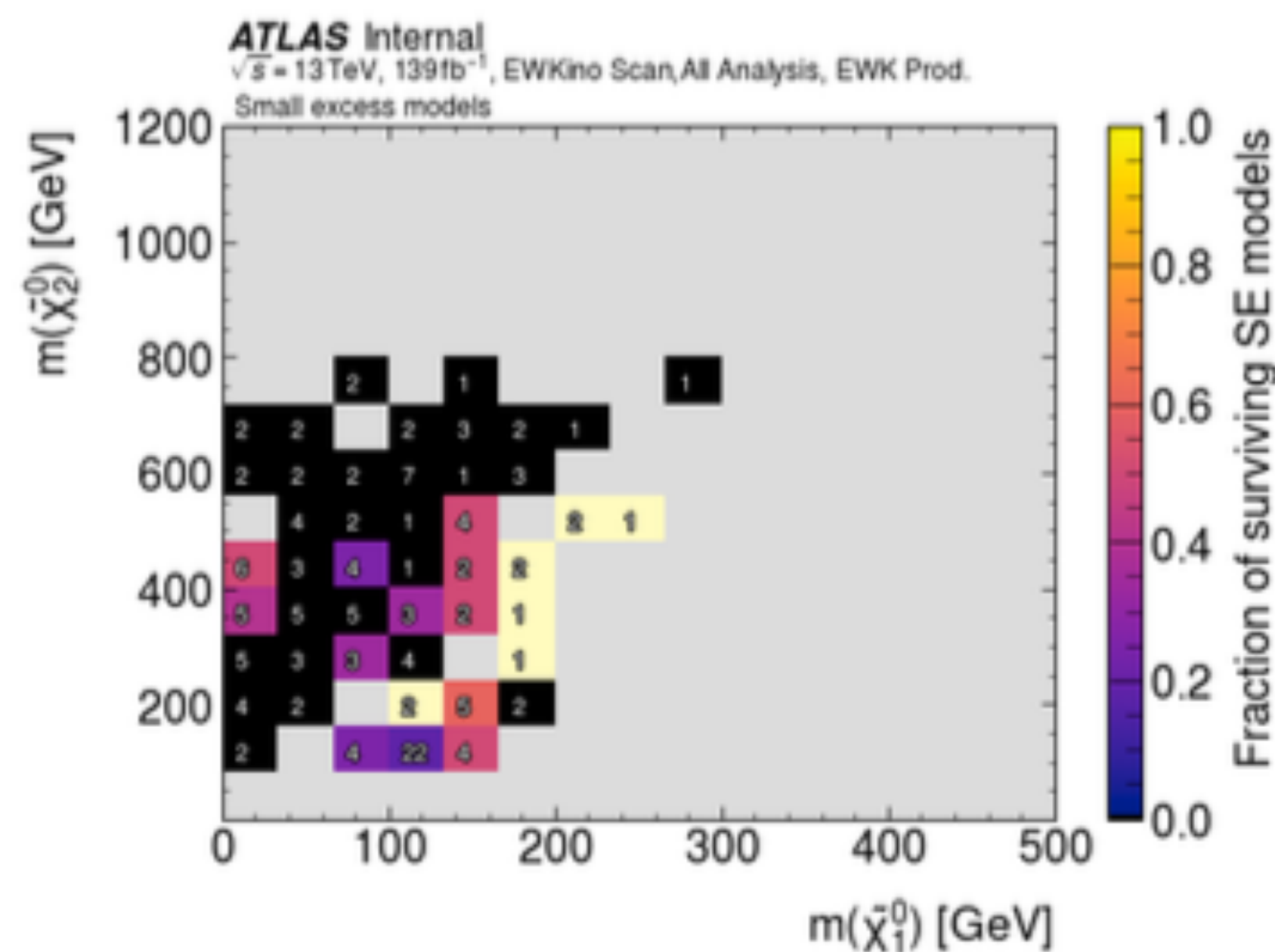
EWK scans: SEM 100-200,  
found twice  $\sim 10$

- ATLAS EWK pMSSM study done in [Nishant Gaurav master thesis](#).
- EWK pMSSM scan (public)
- Single out best motivated models for Run3 searches.

- High mass parameters of SE models
- Focus on models found by multiple analyses
- Single out interesting models for the next round of searches
- Explore correlations among analyses



- Models found once
- Models found twice

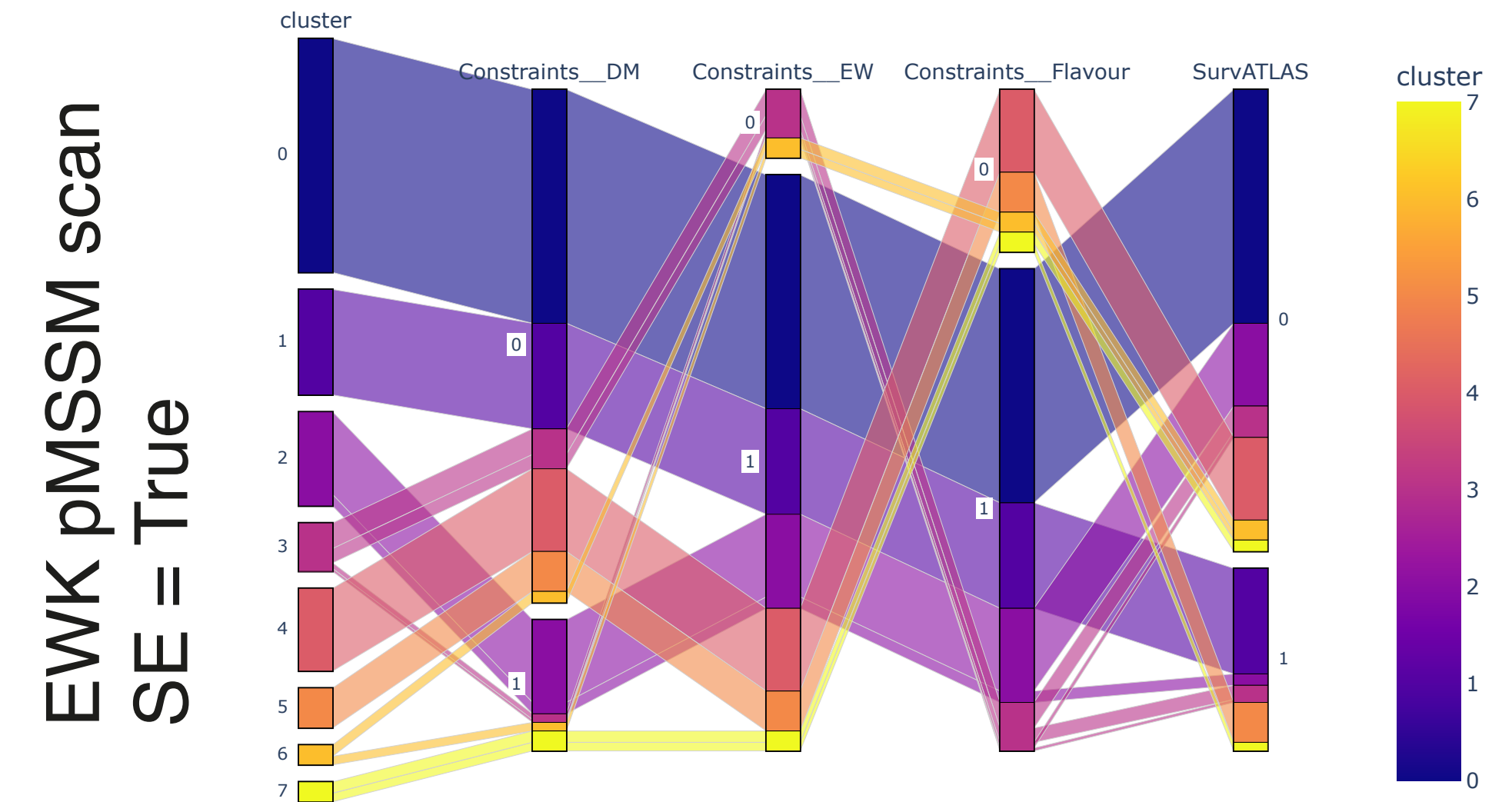




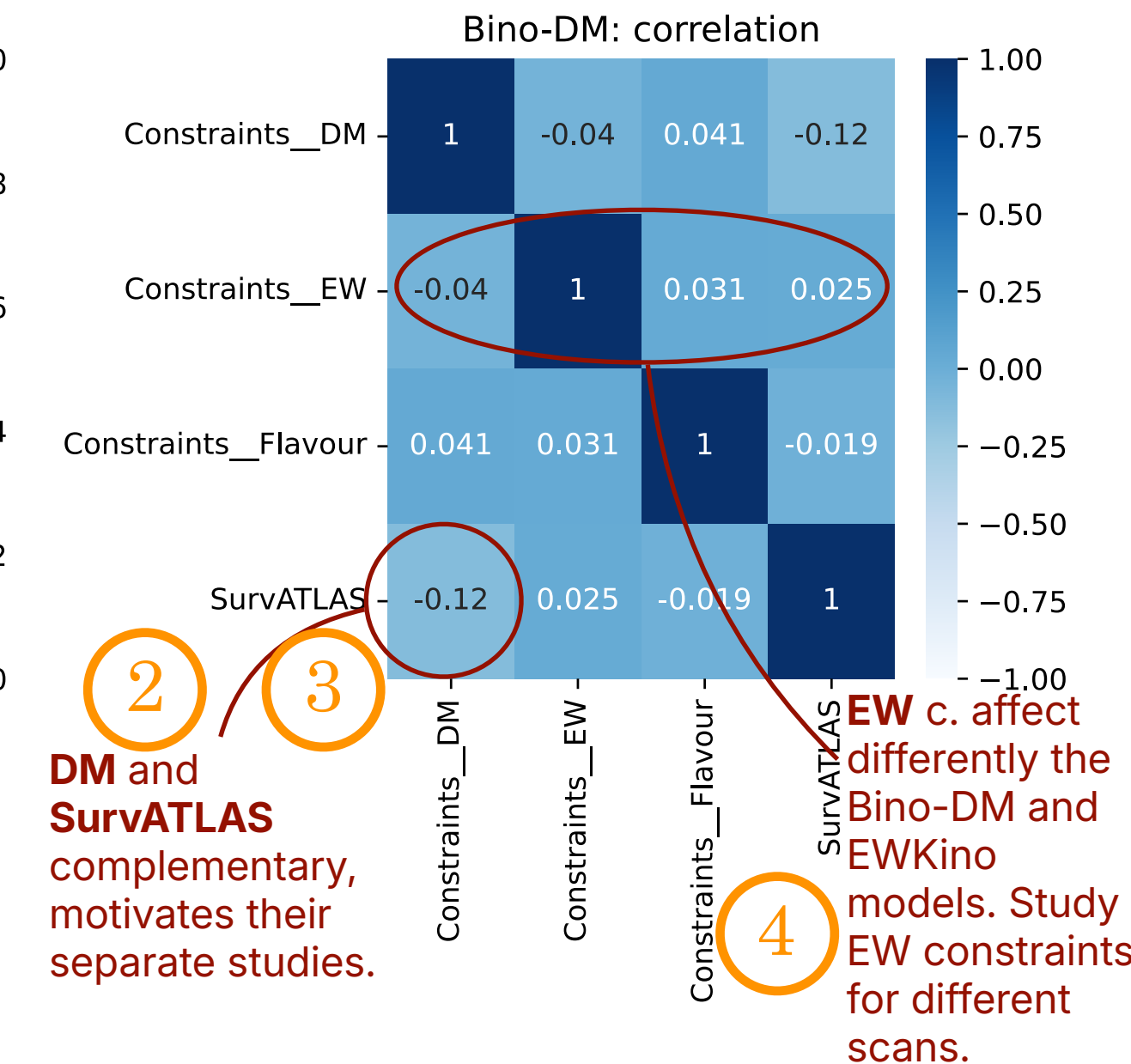
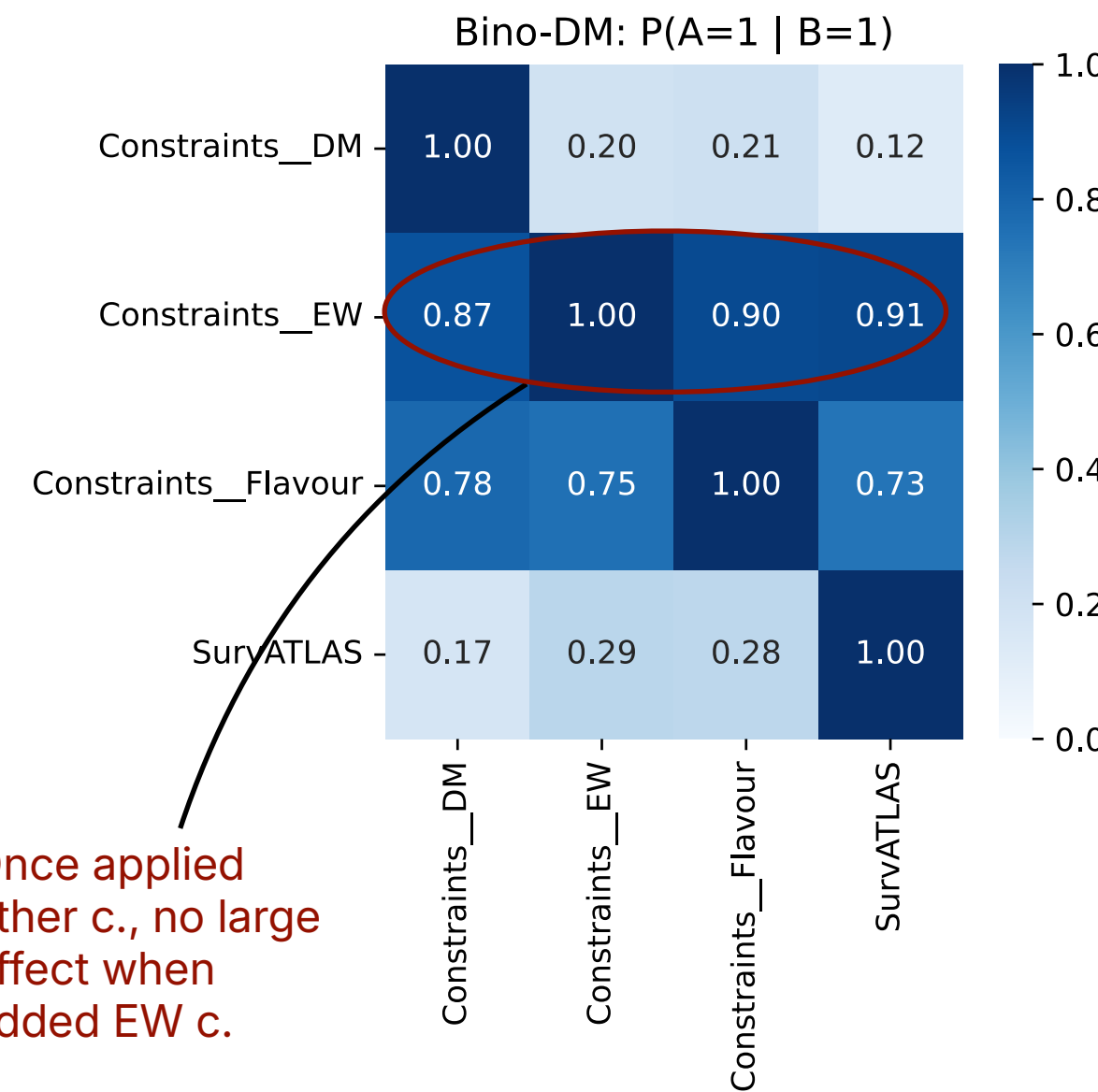
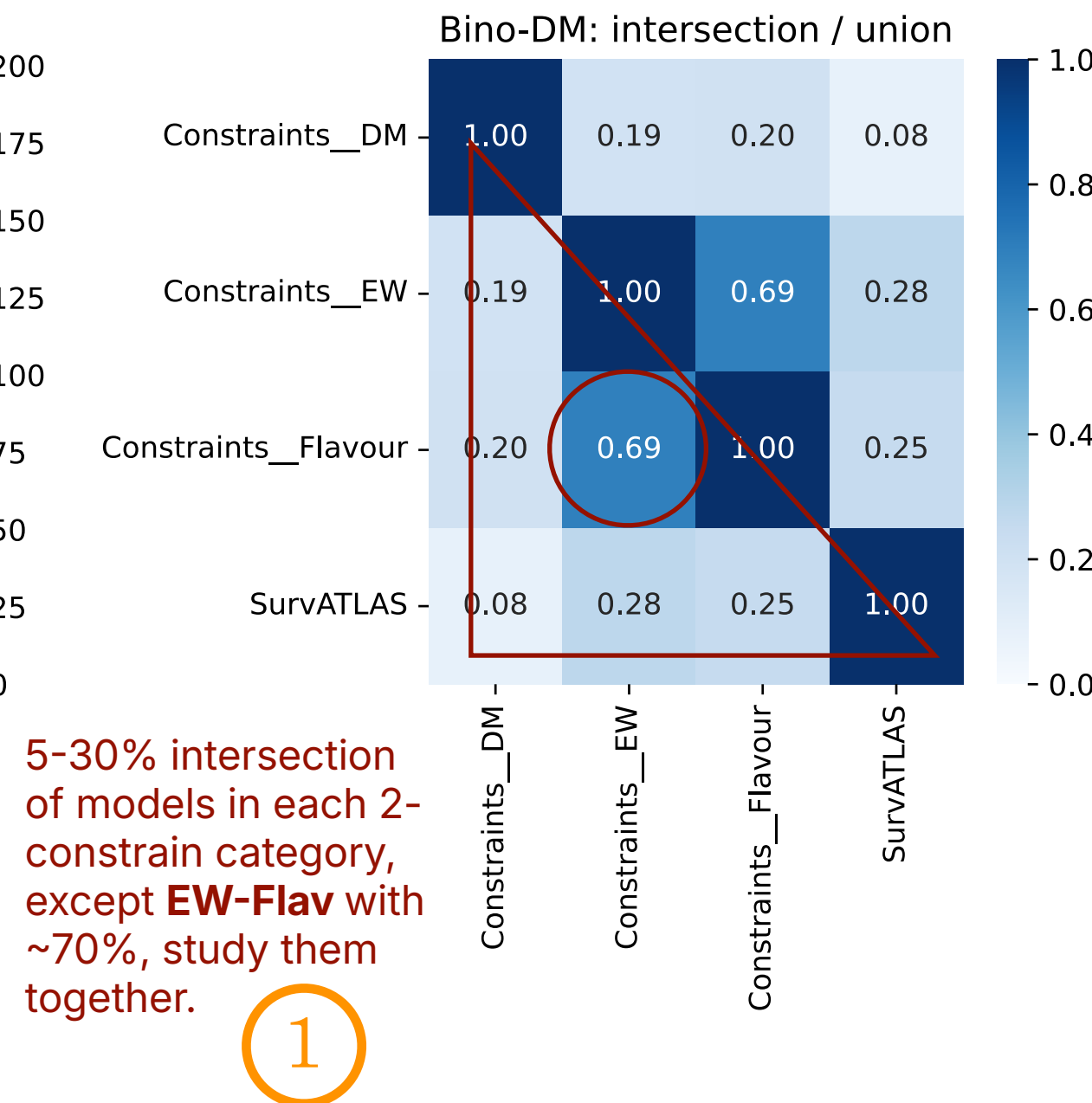
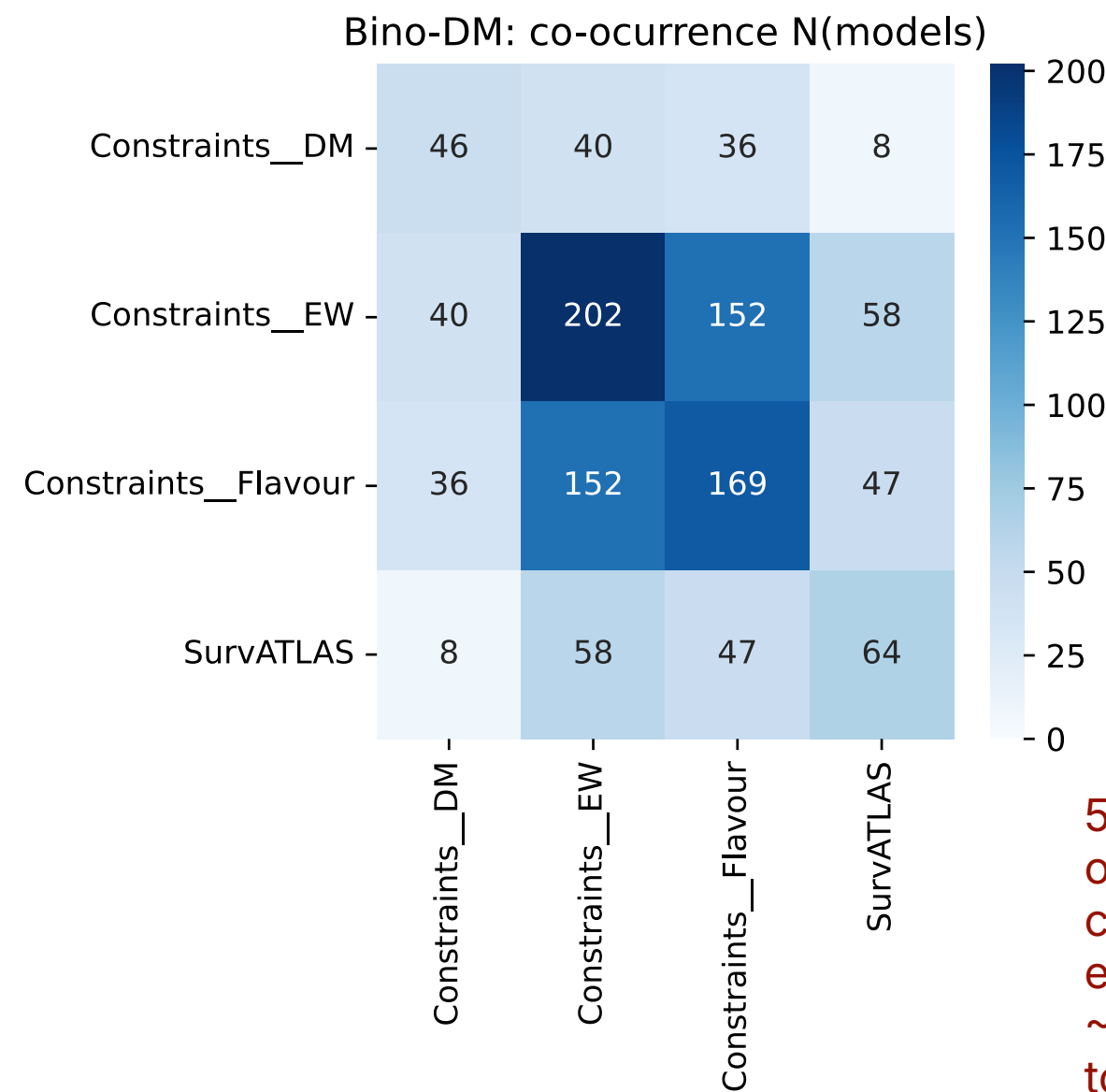
# PRESENT: MOTIVATION FOR SE WITH CLUSTERING

- SUSY should not be considered a theory of everything. Doubt each external constraint and ATLAS measurement.
- Create **categories** for combinations of constraints and model parameter space
- Aim to exploit correlations from different measurements and analyses.
  - Find the most interesting combinations of parameters and constraints
  - Each SE model is interesting, single them out for future study.
- Use clustering algorithms to create SE model categories according to a selection of model parameters and constraints.

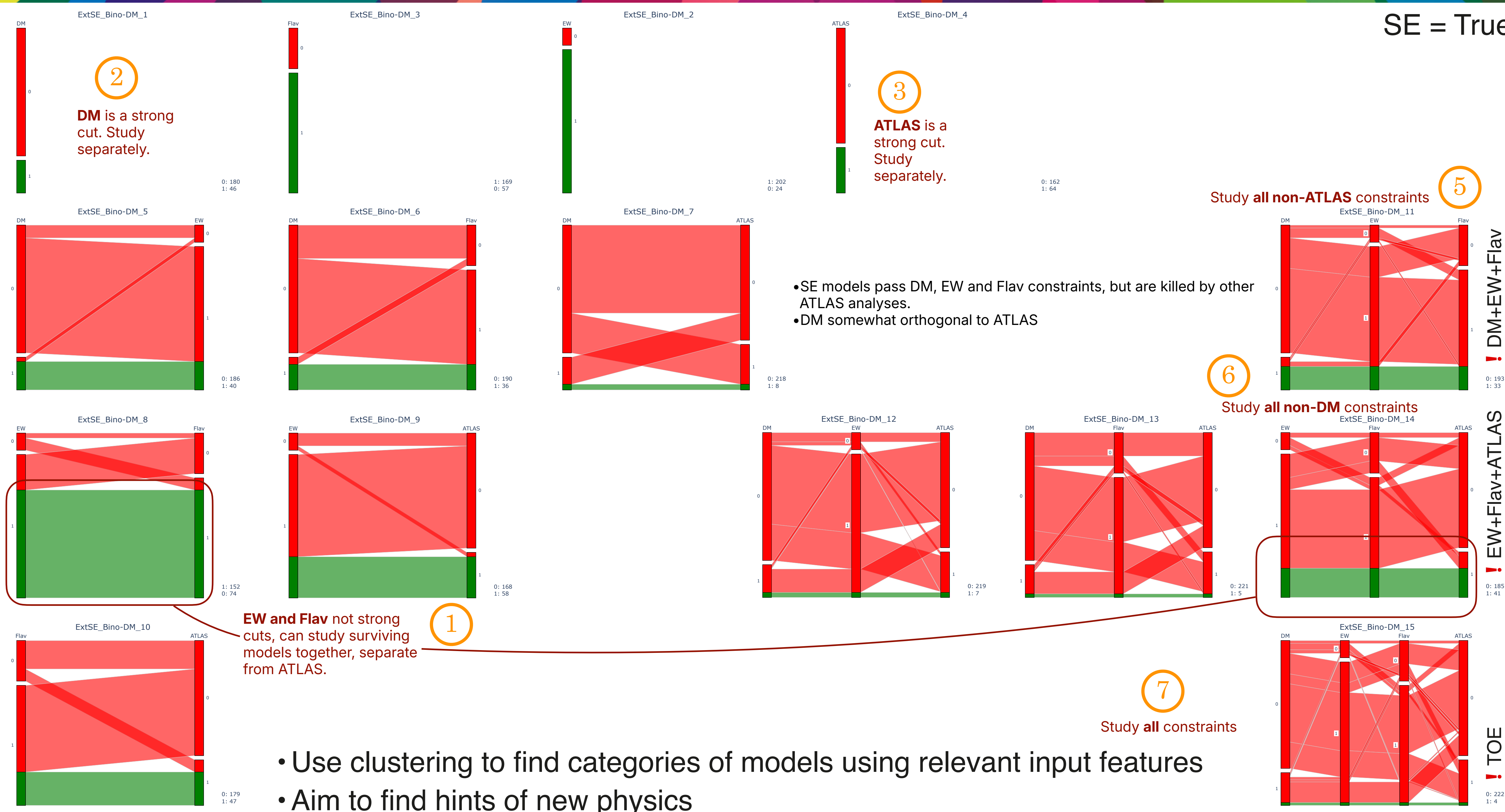
Parallel sets: SE clusters with constraint dimensions (red-green gradient)



Bino-DM, 226 SEM









# PRESENT: CLUSTERING STUDY WORKFLOW

Optimal clustering algorithm chosen per input type!

Judita Mamužić (IJS) in collaboration with  
Andrej Filipčič (IJS), Brian Petersen  
(CERN), Jonas Würzinger (TUM), Lukas  
Alexander Heinrich (TUM), Nishant Gaurav  
(Vanderbilt University), Roberto Ruiz De  
Austri Bazan (IFIC), Aurelio Juste (IFAE),  
Ciril Bohak (UL FRI)

**INPUT FEATURES**  
Binary  
Continuous  
Continuous sparse  
Mixed

**Examine the input features:**

- Binary: external constraints and analyses surviving SE criteria
- Continuous: mass, BF
- Continuous sparse: input parameters, cross-sections
- Mixed: binary+continuous, binary+sparse continuous, continuous+sparse continuous, all

**PREPROCESSING**

Determine if the input benefits from **transformation before any clustering** (for any clustering)

**DOMINANCE**

Determine if input feature requires a **transformation to mitigate dominance** (important for mixed input features, per clustering algorithm)

**k-SWEEP**  
(some algorithms)

**Determine the number of clusters** (per clustering algorithm, some algorithms have a built-in determination of number of clusters)

**CLUSTERING**  
Performance evaluation  
Benchmarking

**Clustering, performance evaluation, benchmarking** according to typical criteria (per input type and clustering algorithm)

**PHYSICS ANALYSIS**

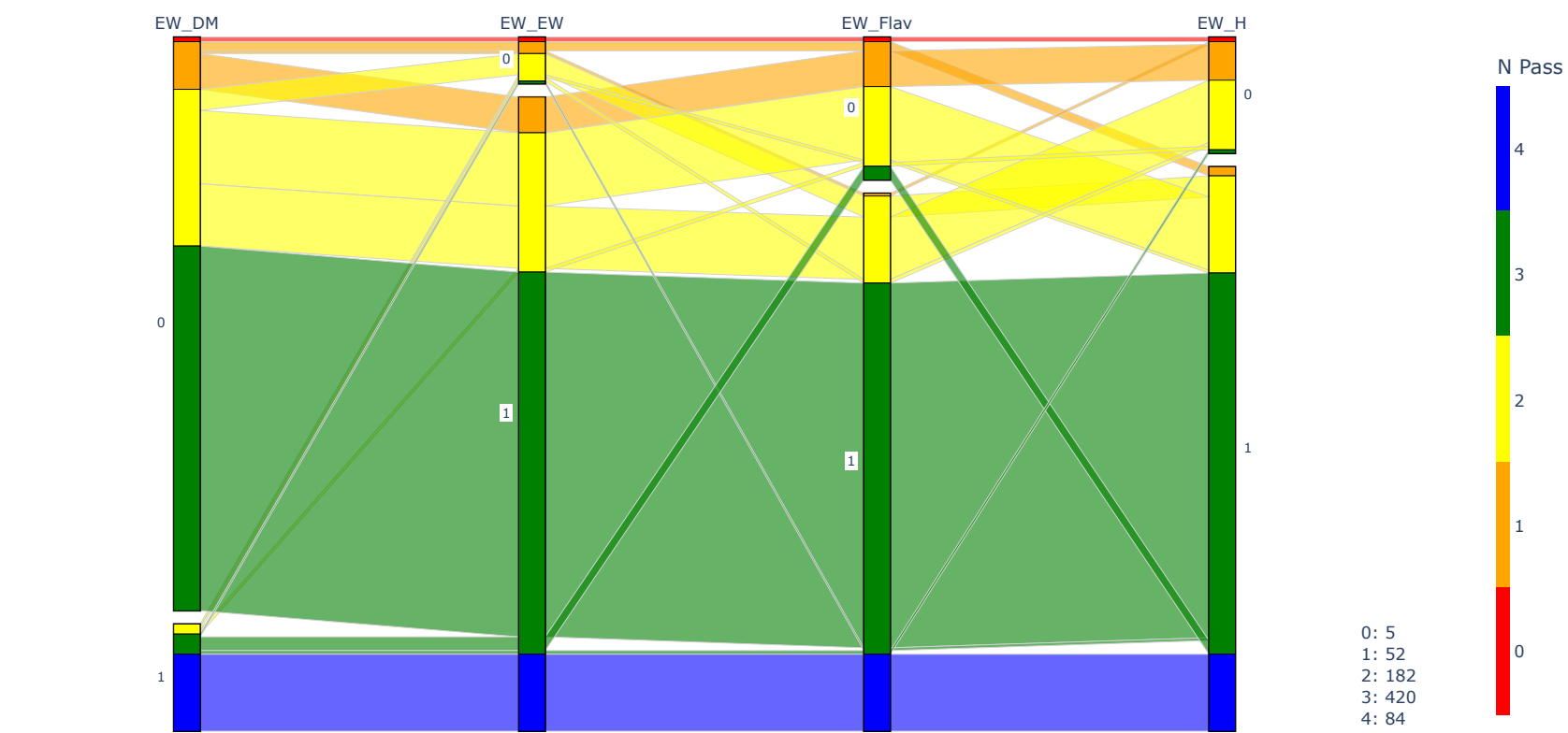
**Physics analysis** with the optimal clustering algorithm(s)



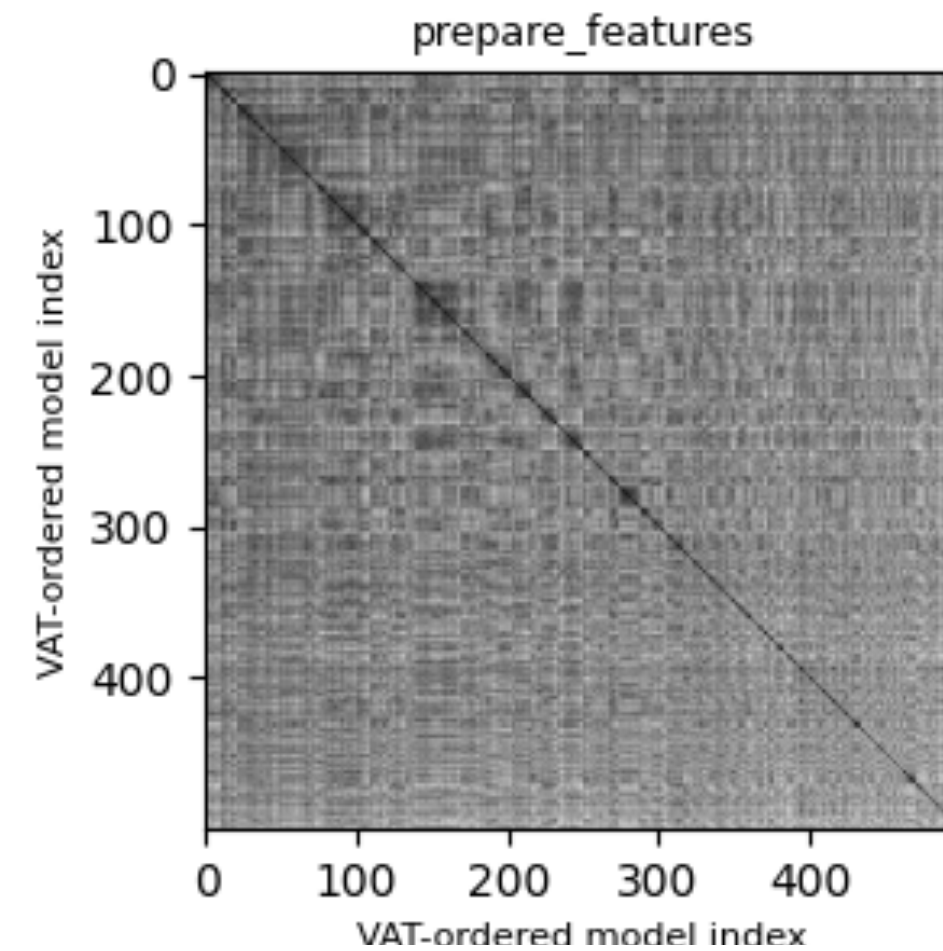
# PRESENT: CLUSTERING FOR SE

General EW pMSSM scan, SE = True

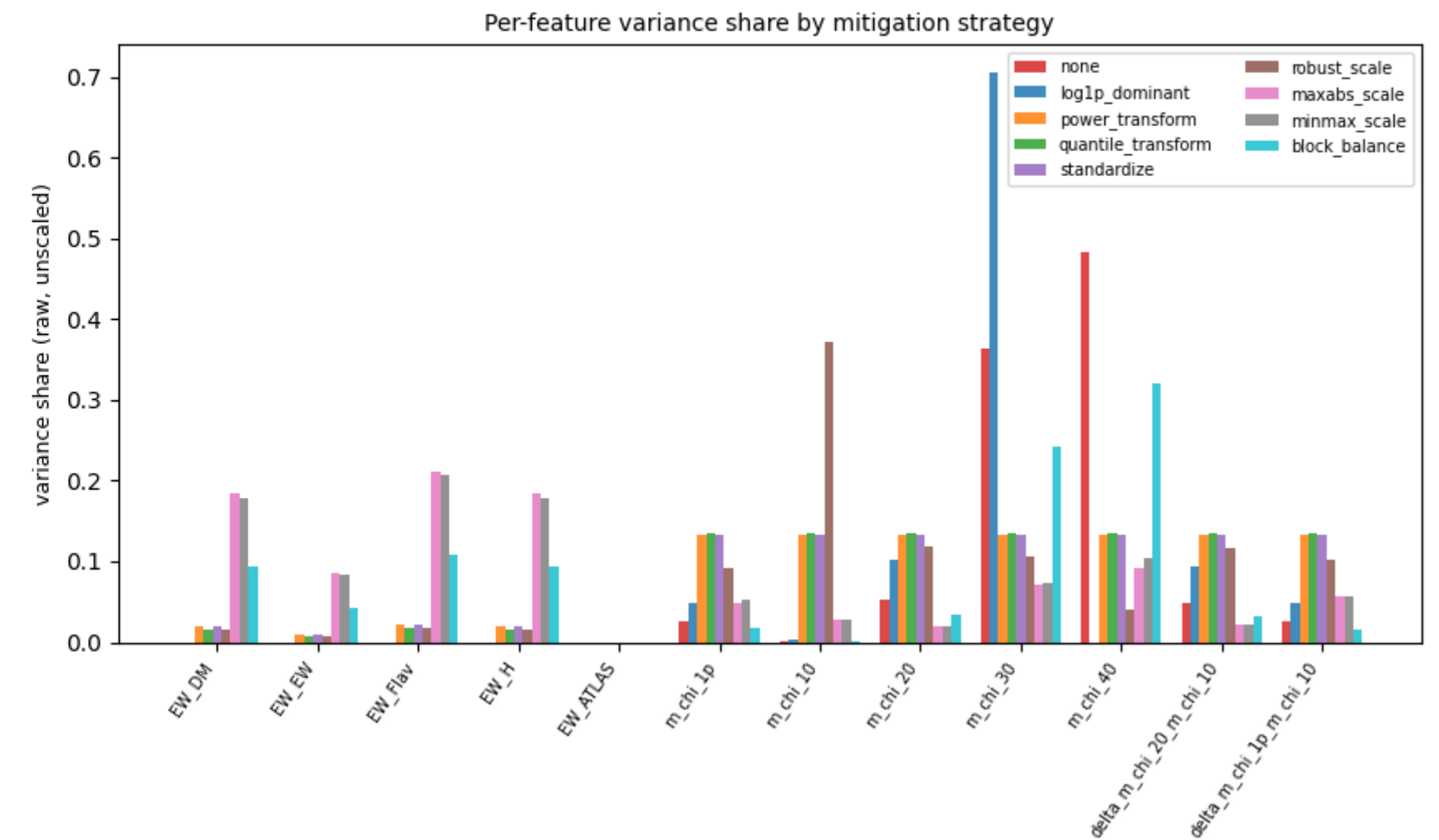
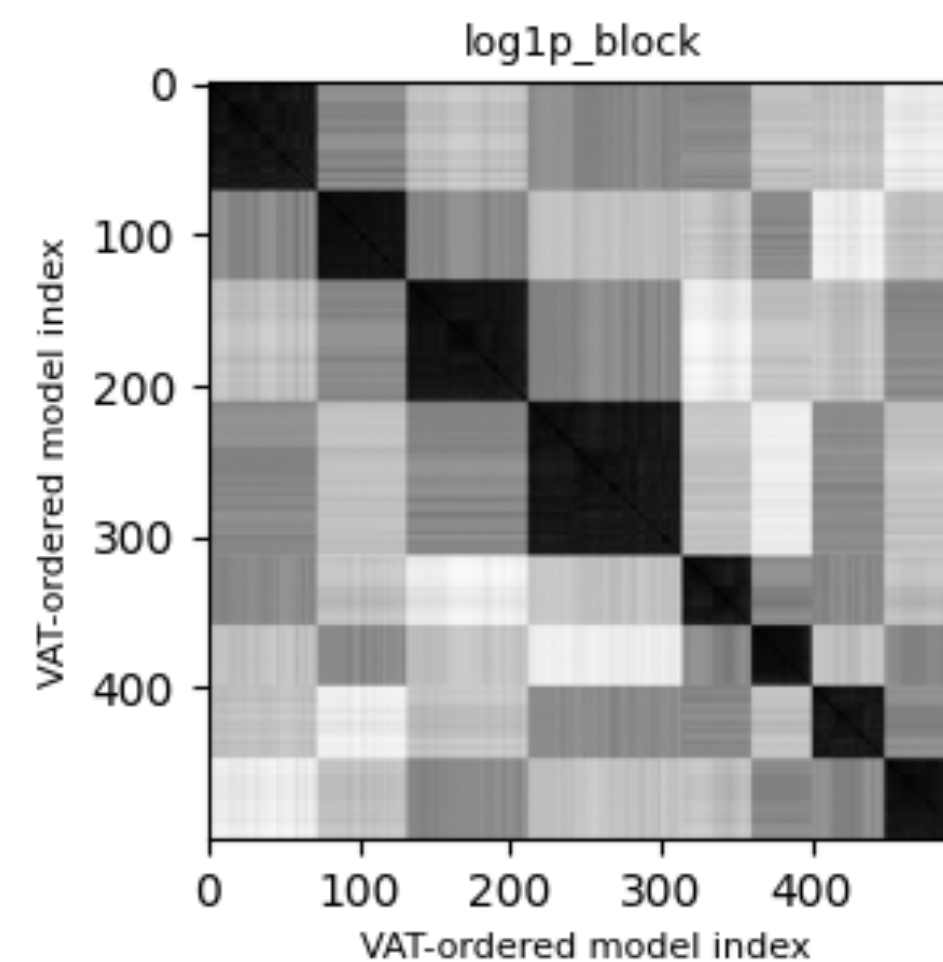
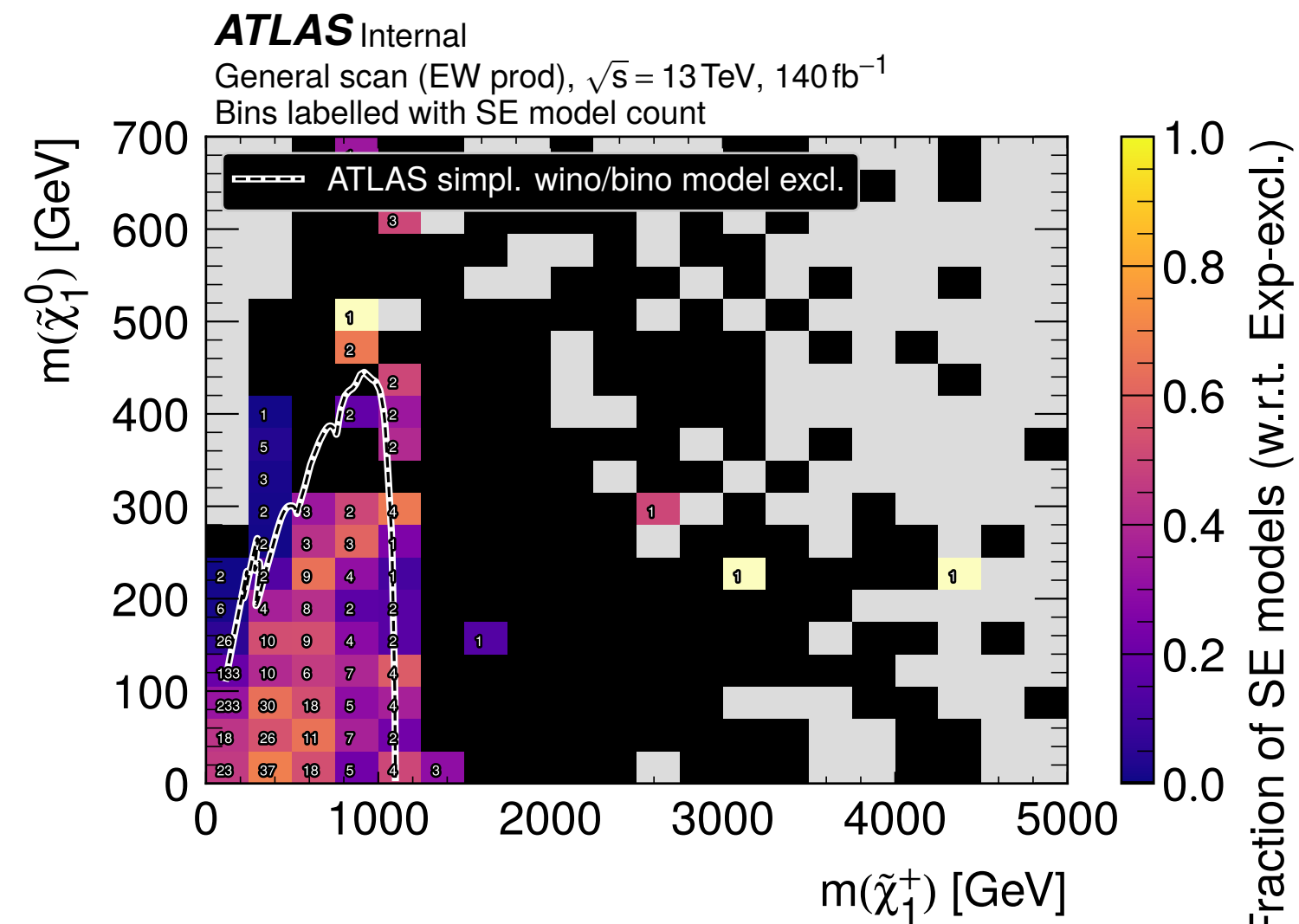
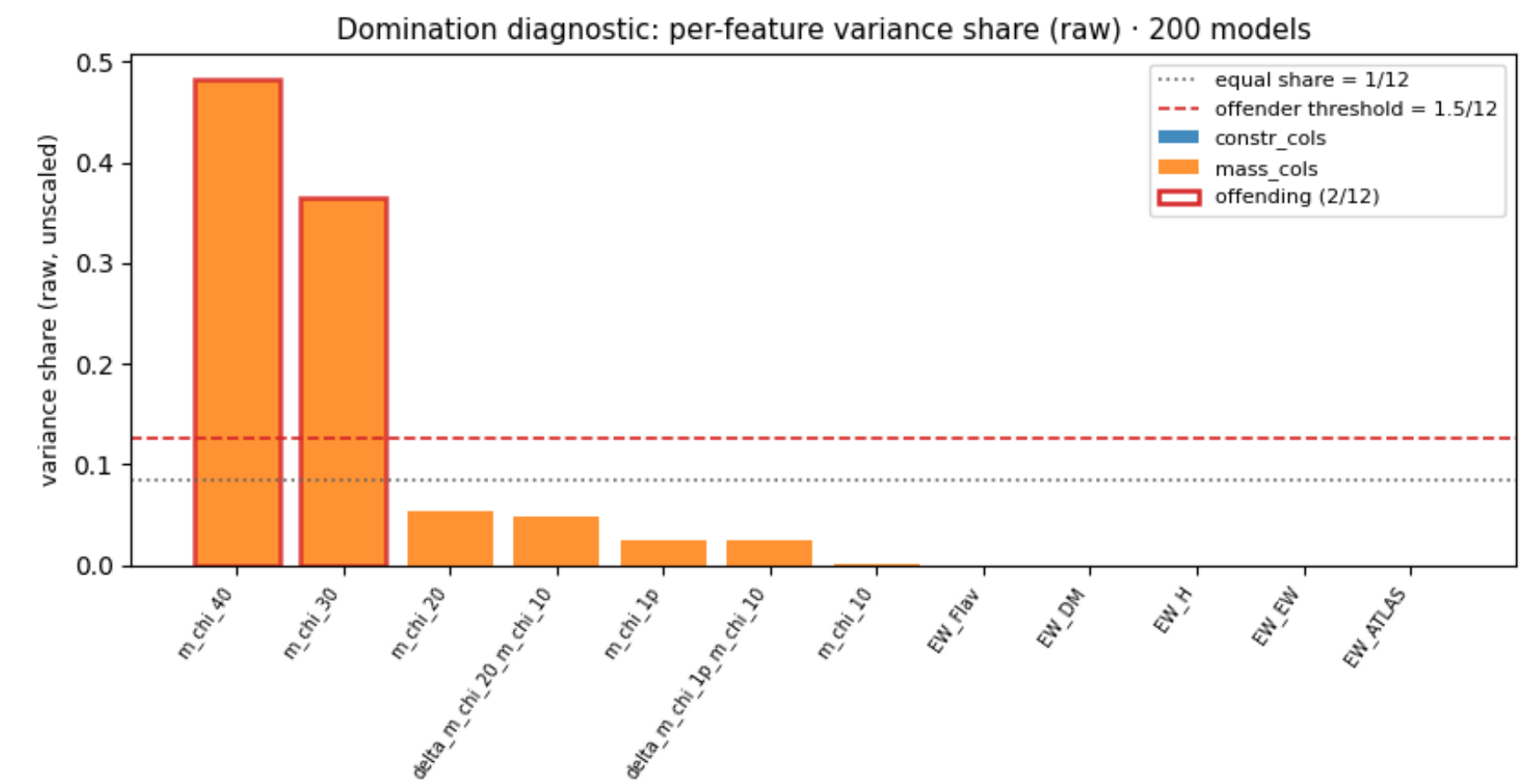
Input



Preprocessing



Domination



• Different **clustering algorithms** correspond to different inputs:

- **Binary**: analyses, external constraints
- **Continuous**: BF, mixing
- **Continuous sparse**: masses, input parameters, cross section
- **Mixed**: mixed binary continuous, mixed binary sparse, mixed sparse and continuous, mixed binary all

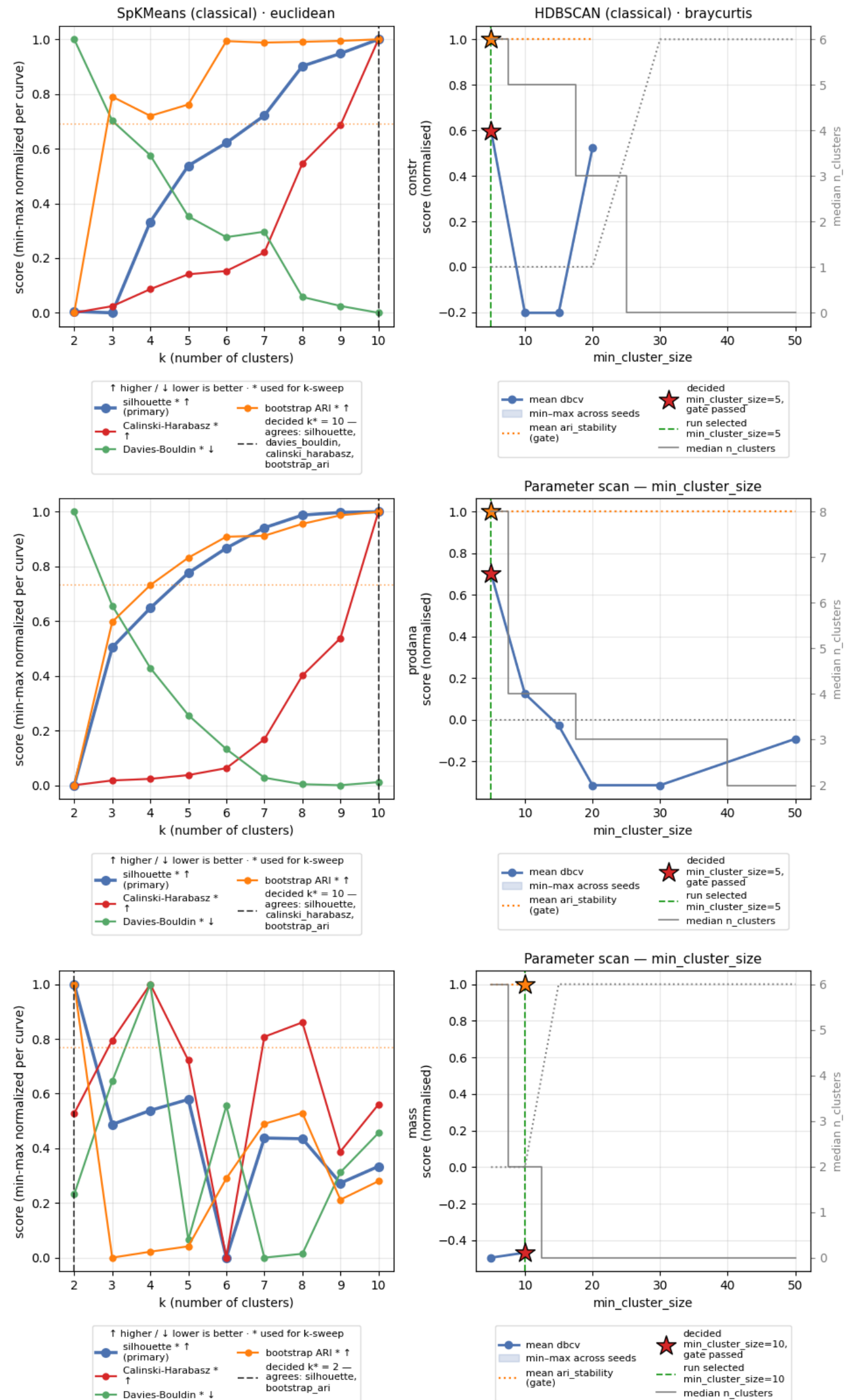
• Using Log transformation for input parameters

• Using per block and per feature within block normalisation



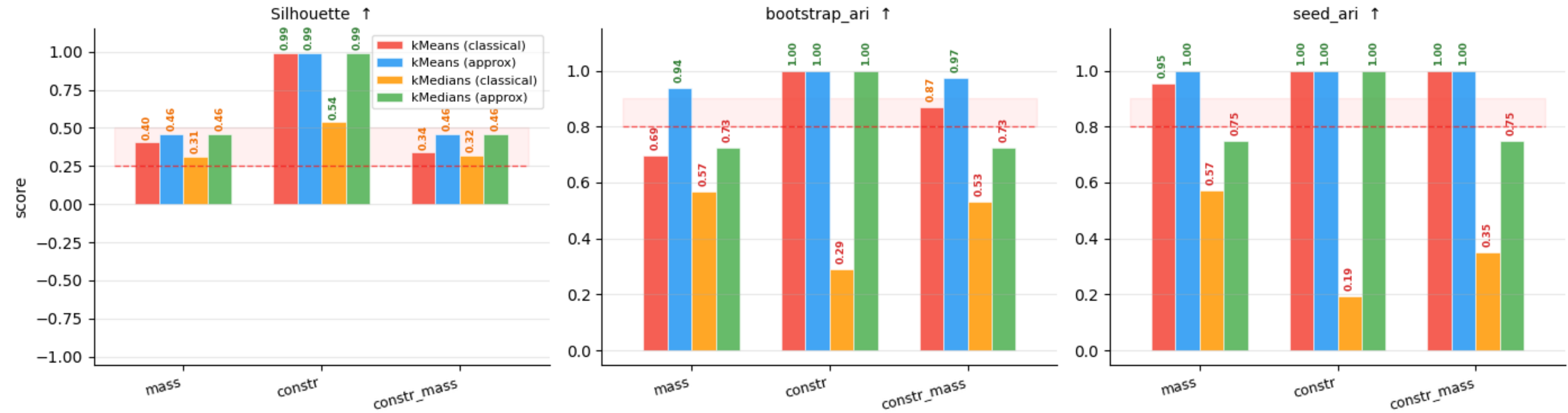
# PRESENT: CLUSTERING FOR SE

## k-Sweep / builtin

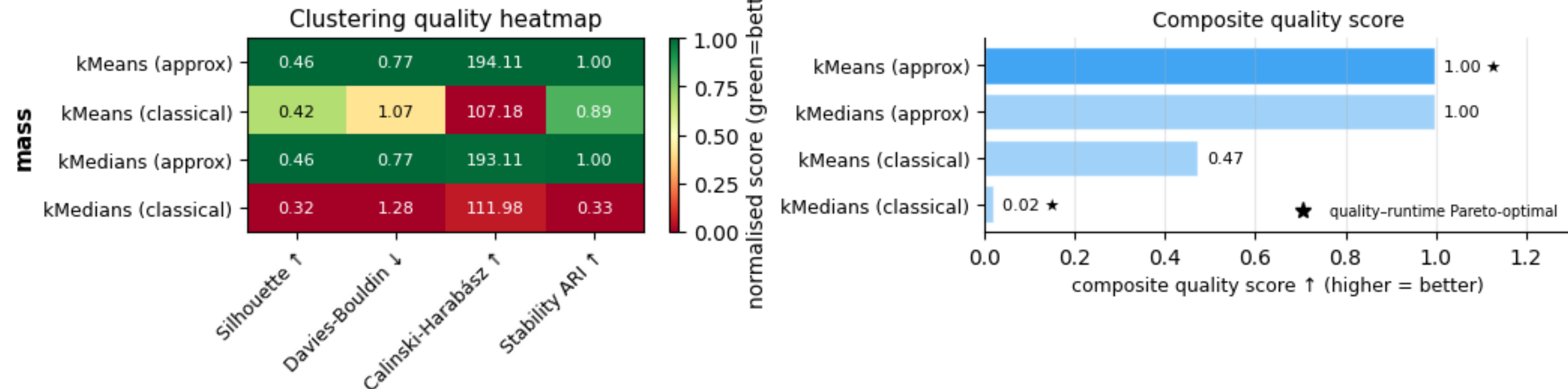


## Clustering

### Algorithm stability



### Algorithm ranking

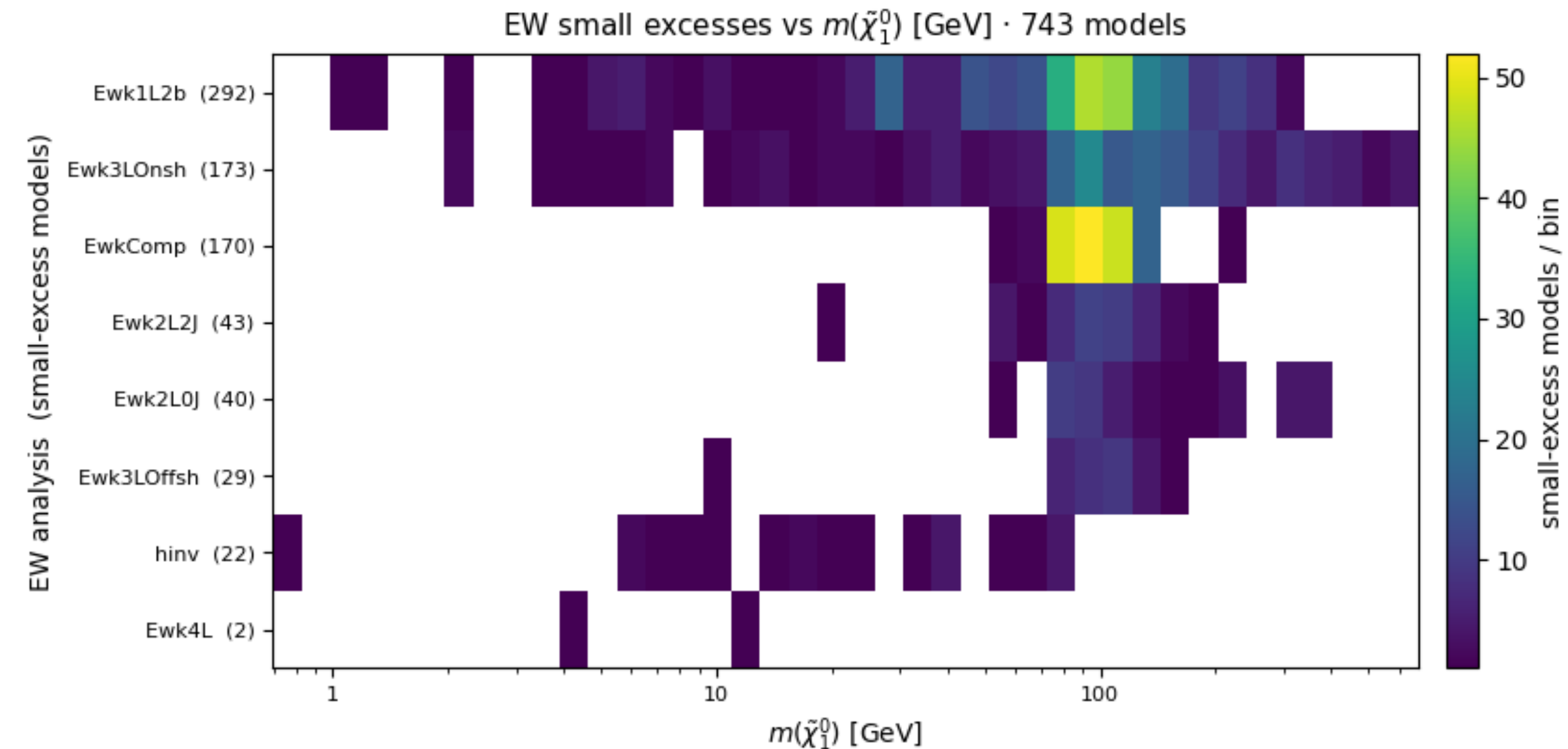




# PRESENT: CLUSTERING FOR SE MODELS

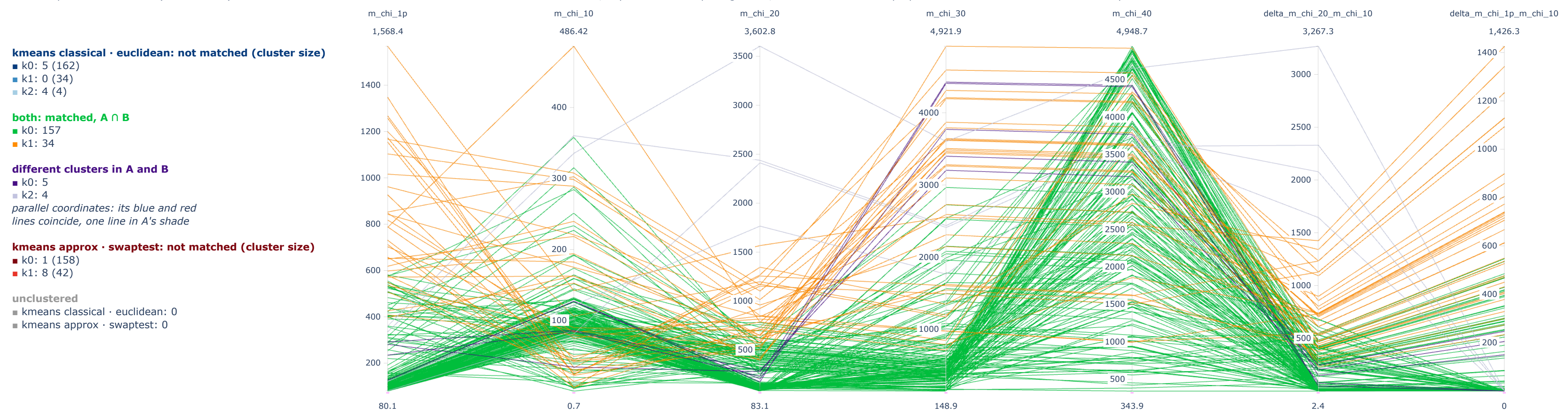
## Physics

- Start with a minimal example, then extend complexity.
- Carefully choose input blocks, extract physics meaning.
  - Interesting to see all SE models point to ~100 GeV neutralino
  - With clustering dominant cluster has ~100 GeV neutralino-1,2 and chargino-1
  - More precise characterisation and considering a higher number of dimensions



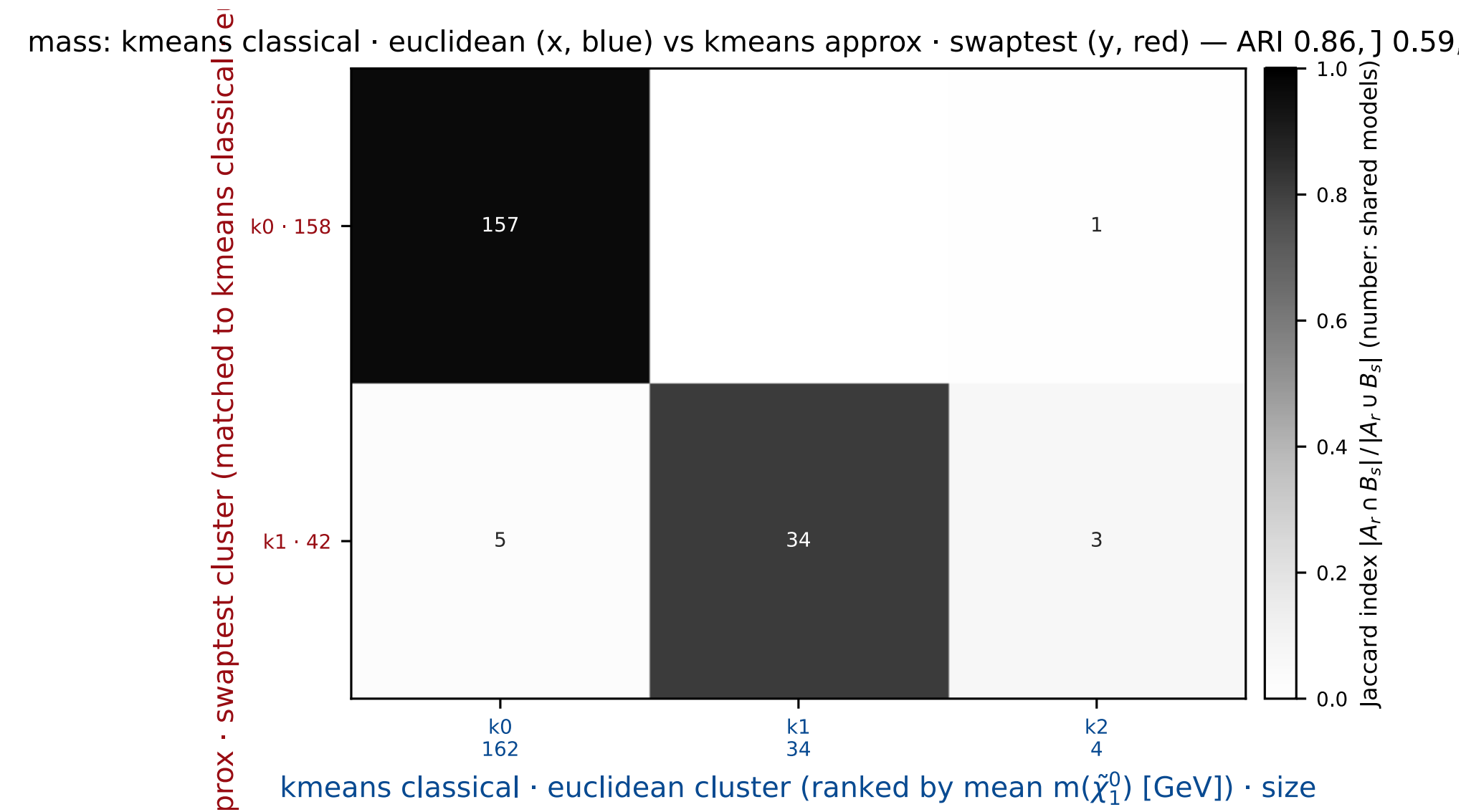
mass: kmeans classical · euclidean (blue) vs kmeans approx · swapttest (red), matched: one colour per rank — ARI 0.86, J 0.59, seed 0

parallel coordinates (7 continuous) · 200 models: a model in matched clusters is one line in its rank's colour, any other one line per algorithm; in the coordinates one purple line where A and B cluster it differently

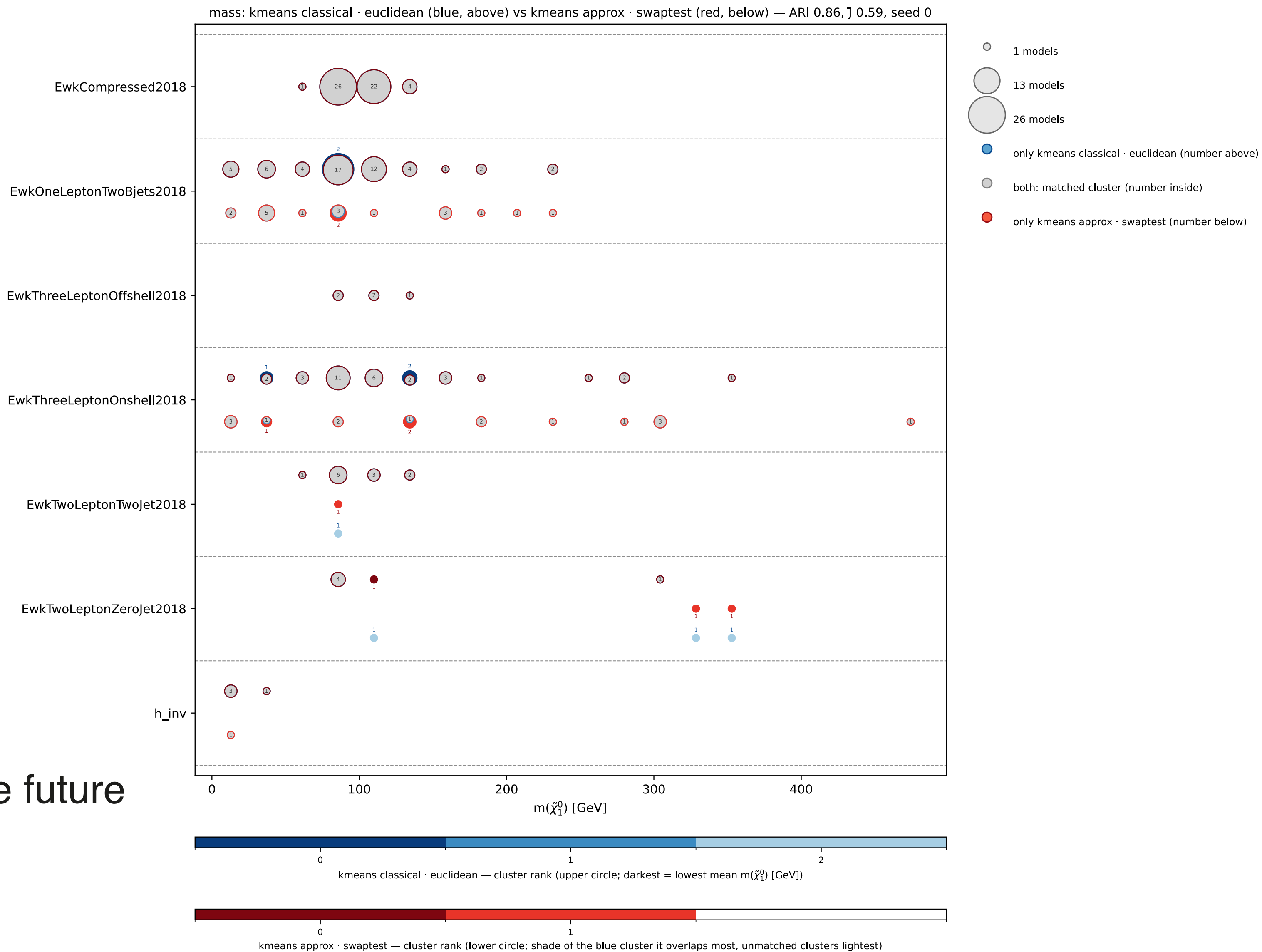




# PRESENT: CLUSTERING FOR SE MODELS



- Look for better agreement of algorithms in the future
- SE clusters around ~100 GeV neutralino

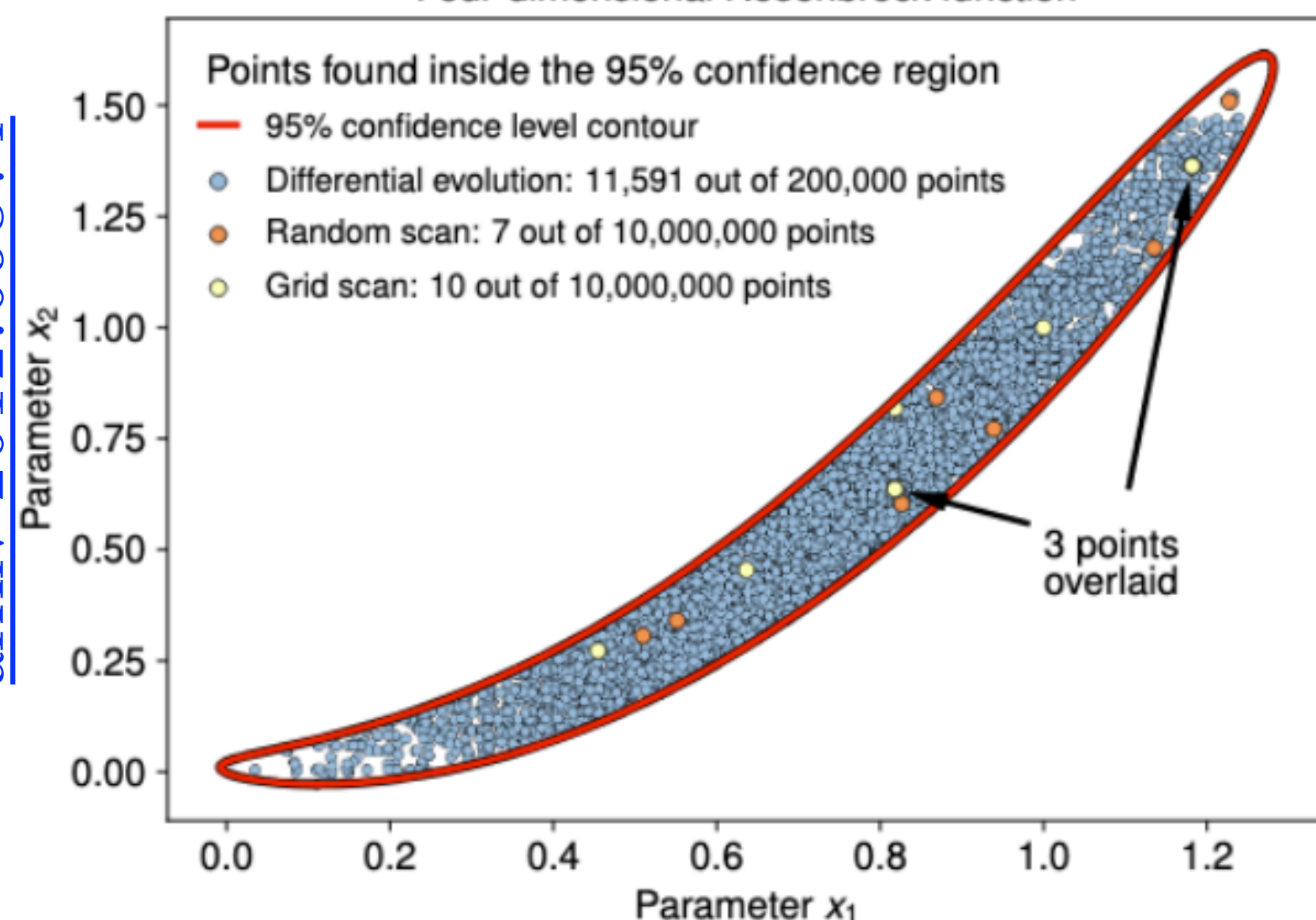




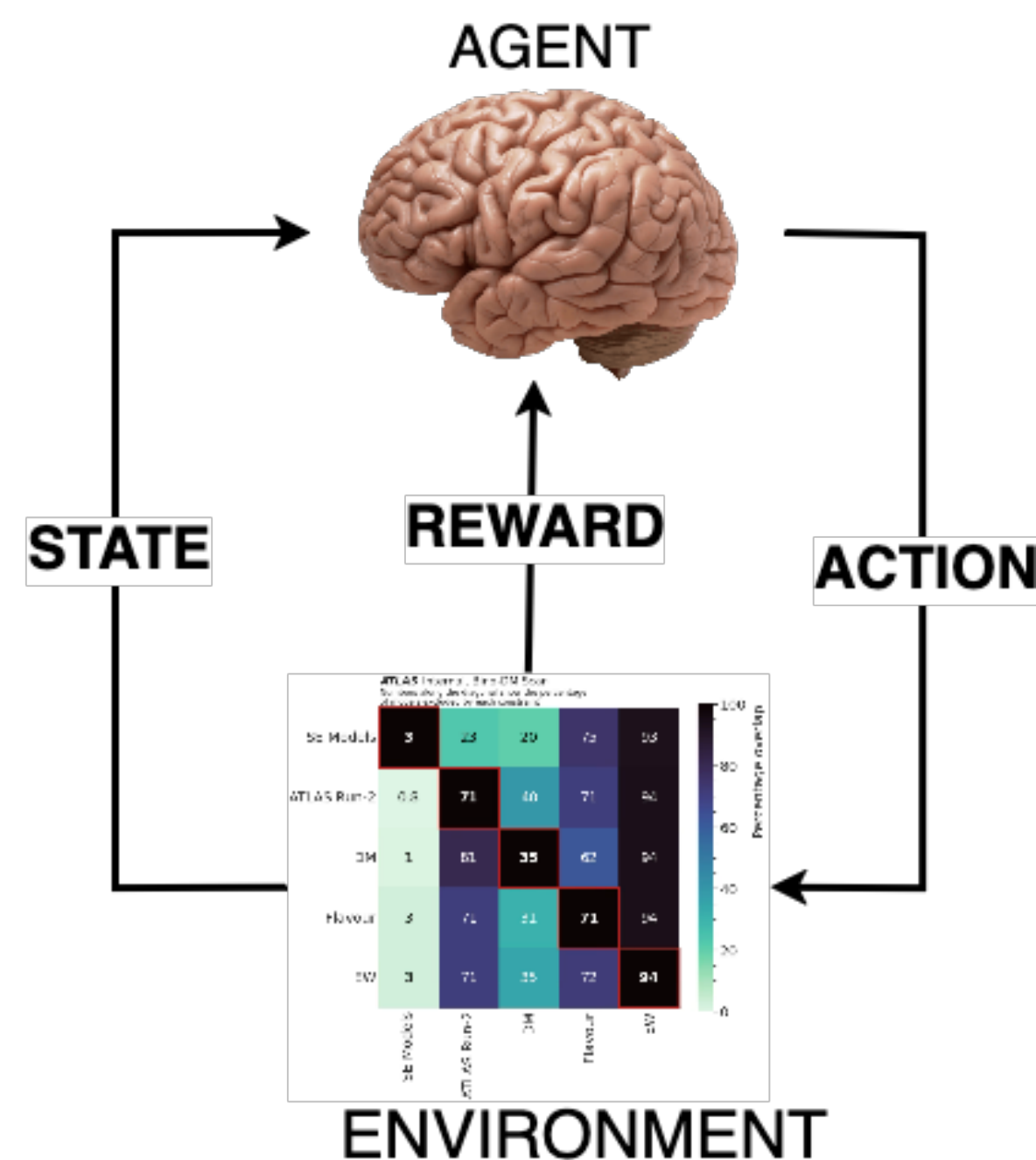
# PRESENT: REINFORCEMENT LEARNING FOR SE MODELS

## Model oversampling

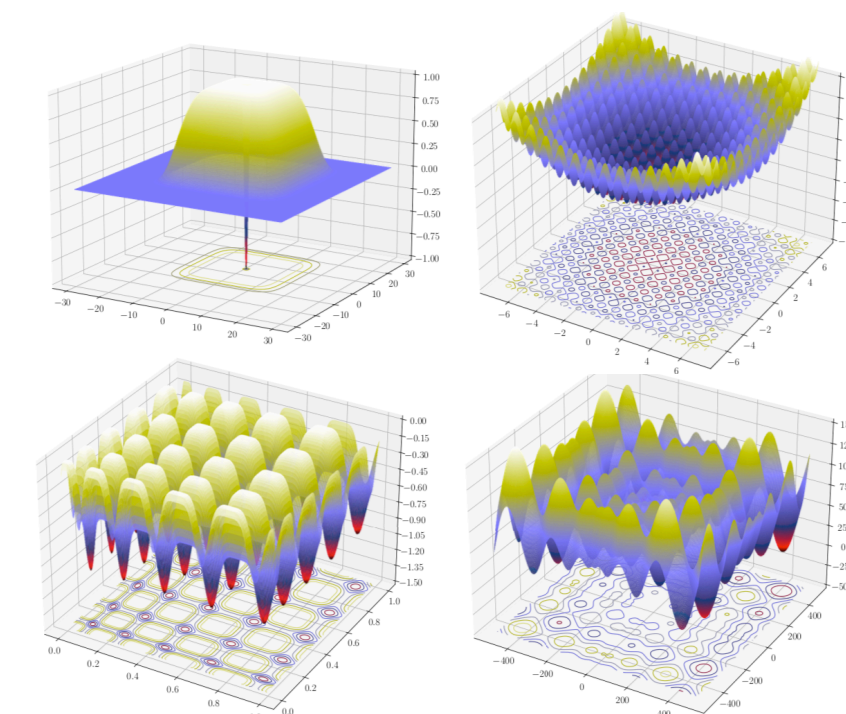
Four-dimensional Rosenbrock function



## Search for SE models



Better/faster  
minimization than  
humans



- Oversampling in the interesting region using Small Excesses criteria to scan better the pMSSM parameter space
- **Reinforcement Learning** (RL) to search through parameter space: Performs minimization, very good at finding a global minimum for high dimensional parameter space (state-of-the-art algorithm): STATE, REWARD, ACTION
  - Agent discovers by trial and error what actions lead to good outcomes.
  - Reinforcement learning algorithm choices strongly affect decisions, make more robust
- Explore interesting high-dimensional parameter space, find better motivated regions.
- Single out models and kinematic parameter space for Run3 searches



# PRESENT: NEXT STEPS

- **Clustering for SE, classical:**

- Go through each block of input and mixed, provide a description of SE models in relation to SUSY parameters and external constraints, provide proto-models for Run3 studies
- Publish a proof-of-concept study using public EWK pMSSM models (*REI WS, November 2026*)
- Implement ATLAS study for the General and 3G scans in ongoing pMSSM scans (*Spring conferences 2027*)

- **Clustering for SE, quantum:**

- Quantum approximation algorithms rank high for many input blocks, summarise findings and estimate quantum computation requirements
- Publish proof-of-concept study using public EWK pMSSM models (*QTI conference, November 2026*)

- **Reinforcement learning for SE, classical:**

- Search for SE models and uncovered parameter space motivated by SE
- Proof-of-concept study (*End-of-year 2026 time scale*)

- **Clustering for SE in Data, classical/approx:**

- Extend clustering study on input features using kinematic distributions (or mean/median) on data events of SR with an excess
- Proof-of-concept study (*March 2026 time scale*)



# FUTURE: RESEARCH PROPOSAL (IN PREPARATION)

- **Strategic BSM searches** to enhance **discovery** potential at the LHC
- **Current paradigm:** model is either found or excluded, use physics motivation for where to look next (**linear**)
- **Change of paradigm:** anomalies are hints of new physics, strategic iterative decision for the next block of searches (**minimisation problem**)

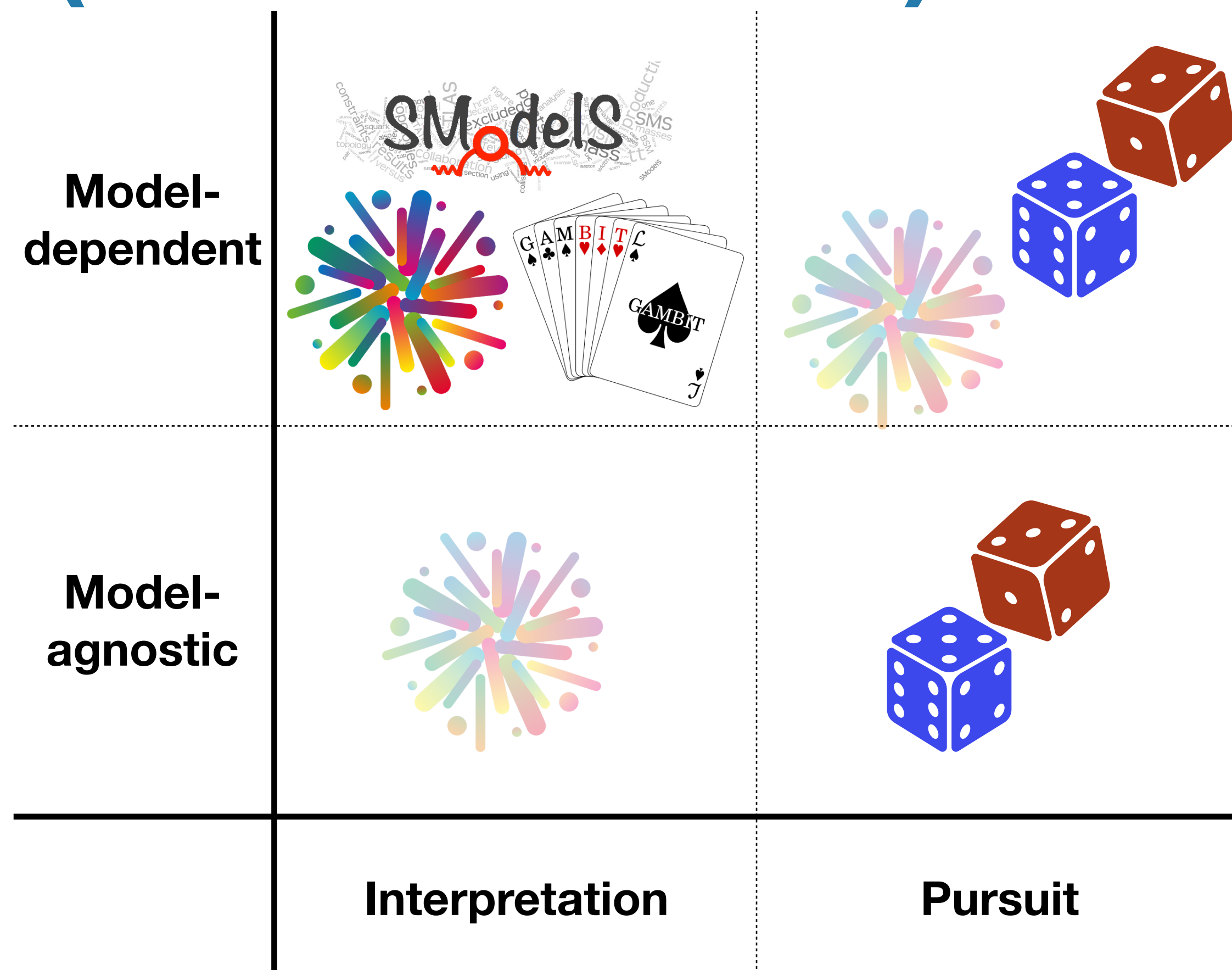
- Instead of the next single optimal search, perform **strategic iterative searches (consider search as a decision-making problem, a game)**
- **Exploit correlations of small excesses** (systematic approach)
- **Exploit the tested parameter space** for the next round of searches (consider **anomalies** and **history of searches**)

- **Previous activity:**

- **Physics:** BSM searches, large-scale reinterpretation, automated cut-flow analysis optimisation, SE, global p-val calculation (EB, analysis contact, MC contact, LHC SUSY XSEC WG contact, currently statistics HMBS liaison, anomaly detection benchmarking, high-dimensional sampling benchmarking)
- **Operations:** Trigger (TRP first main developer, trigger expert), DAQ, Tile calorimeter (RC)

- **Interests:**

- Developing techniques for BSM discovery in physics and detector/trigger

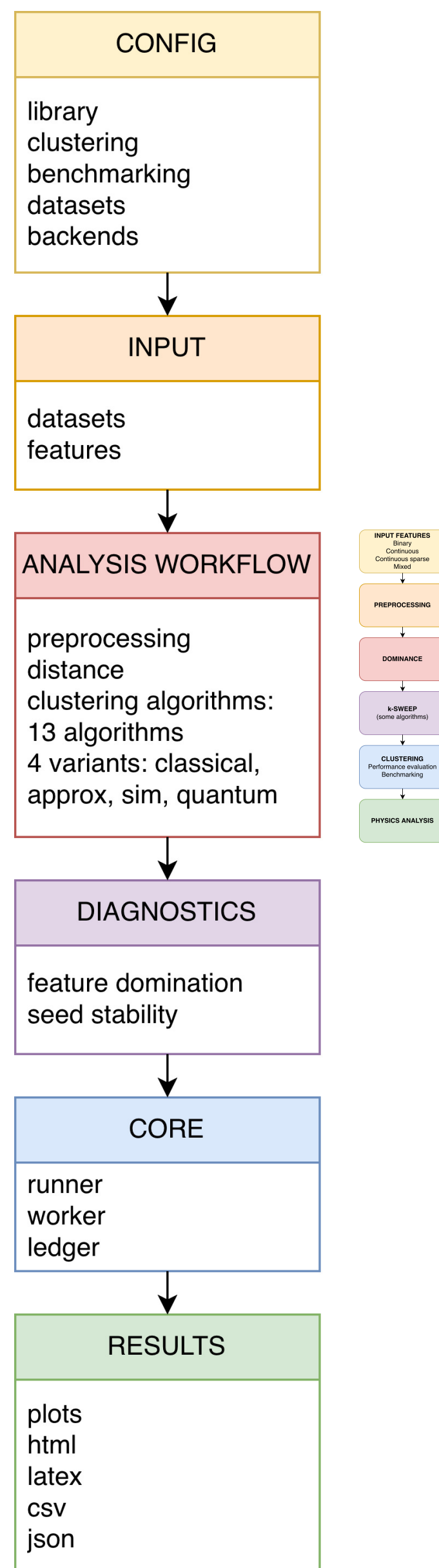




# BACKUP



# QCLUST CODE



## QClust code structure

src/core/

Handles the core orchestration of the system, including registration, plugins, execution, workers, logging, batching, shared helpers, and the CLI.

src/algorithms/

Contains the common algorithm interface plus family-specific adapters and implementation variants across 13 algorithm families.

src/datasets/  
src/features/

Provides dataset-loading utilities and builds the feature groups used by the algorithms.

src/preprocessing.py  
distances.py  
domination.py

Implements the main dataframe-level preprocessing stages, including distance calculations and domination analysis.

src/diagnostics/  
src/output/

Computes diagnostic quality and stability metrics, then stores results and generates reports.

config/

Stores configuration for datasets, execution backends, clustering, and benchmarking.

notebooks/ ↔ run/

Keeps Jupyter notebooks paired one-to-one with executable run scripts, with CI ensuring both versions stay synchronized.

scripts/  
scripts/templates/

Contains automation and support scripts for notebook synchronization, CI, flowcharts, shell utilities, batch execution, and benchmark templates.

tests/

Contains the unit test suite that is run by CI on every push.



# INPUT

Using General pMSSM scan from  
Hammad - before last round of fixes



## CONSTR\_COLS:

- EWSummary\_ObsCLs\_DM\_constraints
- EWSummary\_ObsCLs\_EW\_precision\_constraints
- EWSummary\_ObsCLs\_Flavour\_constraints
- EWSummary\_ObsCLs\_ATLAS\_Higgs\_constraints
- EW\_SurvATLAS

## PRODANA\_COLS:

- Ewk1L2b
- Ewk3LOnsh
- EwkComp
- Ewk2L2J
- Ewk2L0J
- Ewk3LOffsh
- hinv
- EWK4L

## XSEC\_COLS:

- xsec\_20\_1p
- xsec\_20\_1m
- xsec\_1p\_1m
- xsec\_10\_1p
- xsec\_10\_1m

## MIX\_COLS:

- LSP\_Bino\_frac
- LSP\_Wino\_frac
- LSP\_Higgsino\_frac
- chi\_20\_Bino\_frac
- chi\_20\_Wino\_frac
- chi\_20\_Higgsino\_frac

## INPUT\_PARAM\_COLS:

- M\_1
- M\_2
- mu
- tanb
- mA
- meL
- meR
- mmuL
- mmuR
- mtauL
- mtauR

## MASS\_COLS:

- m\_chi\_1p
- m\_chi\_10
- m\_chi\_20
- m\_chi\_30
- m\_chi\_40
- delta\_m\_chi\_20\_m\_chi\_10
- delta\_m\_chi\_1p\_m\_chi\_10

## BF\_COLS:

- BF\_chi\_1p\_to\_chi\_10
- BF\_chi\_1p\_to\_chi\_20
- BF\_chi\_20\_to\_chi\_10\_ele\_ele
- BF\_chi\_20\_to\_chi\_10\_mu\_mu
- BF\_chi\_20\_to\_chi\_10\_Z
- BF\_chi\_20\_to\_chi\_10\_h

# INPUT FEATURES

- Input: SE criteria survival by at least one ATLAS SUSY+ analysis
-  No models survive all ATLAS analyses
- External constraints survival, ATLAS SUSY analyses survival, ATLAS Higgs analyses survival
- pMSSM models: Input paramters, masses, BF, mixing, cross-section

- Different **clustering algorithms** correspond to different inputs:
  - **Binary**: analyses, external constraints
  - **Continuous**: BF, mixing
  - **Continuous sparse**: masses, input parameters, cross section
  - **Mixed**: mixed binary continuous, mixed binary sparse, mixed sparse and continuous, mixed binary all
- From the range of distributions, assign *sparse continuous* to some features (human)

## Binary

|                                           |        |           |               |
|-------------------------------------------|--------|-----------|---------------|
| EWSummary_ObsCLs_DM_constraints           | binary | ✔ kept    | (117/743 = 1) |
| EWSummary_ObsCLs_EW_precision_constraints | binary | ✔ kept    | (692/743 = 1) |
| EWSummary_ObsCLs_Flavour_constraints      | binary | ✔ kept    | (587/743 = 1) |
| EWSummary_ObsCLs_ATLAS_Higgs_constraints  | binary | ✔ kept    | (616/743 = 1) |
| EW_SurvATLAS                              | binary | ❌ DROPPED | (all 743)     |
| EW_DirectStauRun2BDTs2020_se              | binary | ❌ DROPPED | (all 743)     |
| EW_DisappearingTrack_se                   | binary | ❌ DROPPED | (all 743)     |
| EW_DitauMET_OS2020_se                     | binary | ❌ DROPPED | (all 743)     |
| EW_EwkCompressed2018_se                   | binary | ✔ kept    | (170/743 = 1) |
| EW_EwkFullHad2018_se                      | binary | ❌ DROPPED | (all 743)     |
| EW_EwkOneLeptonTwoBjets2018_se            | binary | ✔ kept    | (292/743 = 1) |
| EW_EwkThreeLeptonOffshell2018_se          | binary | ✔ kept    | (29/743 = 1)  |
| EW_EwkThreeLeptonOnshell2018_se           | binary | ✔ kept    | (173/743 = 1) |
| EW_EwkTwoLeptonTwoJet2018_se              | binary | ✔ kept    | (43/743 = 1)  |
| EW_EwkTwoLeptonZeroJet2018_se             | binary | ✔ kept    | (40/743 = 1)  |
| EW_FourLepton2018_se                      | binary | ✔ kept    | (2/743 = 1)   |
| EW_h_inv_se                               | binary | ✔ kept    | (22/743 = 1)  |
| EW_m_A_se                                 | binary | ❌ DROPPED | (all 743)     |
| EW_StrongTwoLeptonTwoJet2018_se           | binary | ❌ DROPPED | (all 743)     |
| EW_OneLepton2018_se                       | binary | ❌ DROPPED | (all 743)     |
| EW_StopTwoLepton2018_se                   | binary | ❌ DROPPED | (all 743)     |

## General EWK pMSSM scan (Sparse) continuous

|                             |            |        |                            |
|-----------------------------|------------|--------|----------------------------|
| m_chi_1p                    | continuous | ✔ kept | (range [80.1, 4.38e+03])   |
| m_chi_10                    | continuous | ✔ kept | (range [0.698, 677])       |
| m_chi_20                    | continuous | ✔ kept | (range [82, 4.38e+03])     |
| m_chi_30                    | continuous | ✔ kept | (range [103, 4.95e+03])    |
| m_chi_40                    | continuous | ✔ kept | (range [217, 5.02e+03])    |
| delta_m_chi_20_m_chi_10     | continuous | ✔ kept | (range [0.264, 4.17e+03])  |
| delta_m_chi_1p_m_chi_10     | continuous | ✔ kept | (range [0.0135, 4.17e+03]) |
| M_1                         | continuous | ✔ kept | (range [-4.97e+03,         |
| M_2                         | continuous | ✔ kept | (range [-4.95e+03,         |
| mu                          | continuous | ✔ kept | (range [-4.94e+03,         |
| tanb                        | continuous | ✔ kept | (range [2.62, 60.8])       |
| mA                          | continuous | ✔ kept | (range [148, 4.99e+03])    |
| meL                         | continuous | ✔ kept | (range [66.5, 5e+03])      |
| meR                         | continuous | ✔ kept | (range [15.3, 5e+03])      |
| mmuL                        | continuous | ✔ kept | (range [66.5, 5e+03])      |
| mmuR                        | continuous | ✔ kept | (range [15.3, 5e+03])      |
| mtauL                       | continuous | ✔ kept | (range [68.2, 5e+03])      |
| mtauR                       | continuous | ✔ kept | (range [109, 5e+03])       |
| BF_chi_1p_to_chi_10         | continuous | ✔ kept | (range [0, 1])             |
| BF_chi_1p_to_chi_20         | continuous | ✔ kept | (range [0, 0.357])         |
| BF_chi_20_to_chi_10_ele_ele | continuous | ✔ kept | (range [0, 0.232])         |
| BF_chi_20_to_chi_10_mu_mu   | continuous | ✔ kept | (range [0, 0.232])         |
| BF_chi_20_to_chi_10_Z       | continuous | ✔ kept | (range [0, 1])             |
| BF_chi_20_to_chi_10_h       | continuous | ✔ kept | (range [0, 0.996])         |
| xsec_20_1p                  | continuous | ✔ kept | (range [2.3e-14, 13.5])    |
| xsec_20_1m                  | continuous | ✔ kept | (range [3.57e-15, 8.42])   |
| xsec_1p_1m                  | continuous | ✔ kept | (range [1.6e-13, 24.1])    |
| xsec_10_1p                  | continuous | ✔ kept | (range [9.42e-10, 28.7])   |
| xsec_10_1m                  | continuous | ✔ kept | (range [2.34e-10, 18.5])   |
| LSP_Bino_frac               | continuous | ✔ kept | (range [5e-13, 1])         |
| LSP_Wino_frac               | continuous | ✔ kept | (range [4.86e-12, 1])      |
| LSP_Higgsino_frac           | continuous | ✔ kept | (range [7.55e-05, 1])      |
| chi_20_Bino_frac            | continuous | ✔ kept | (range [2.46e-11, 1])      |
| chi_20_Wino_frac            | continuous | ✔ kept | (range [7e-08, 1])         |
| chi_20_Higgsino_frac        | continuous | ✔ kept | (range [9.78e-05, 1])      |

INPUT FEATURES

Binary

Continuous

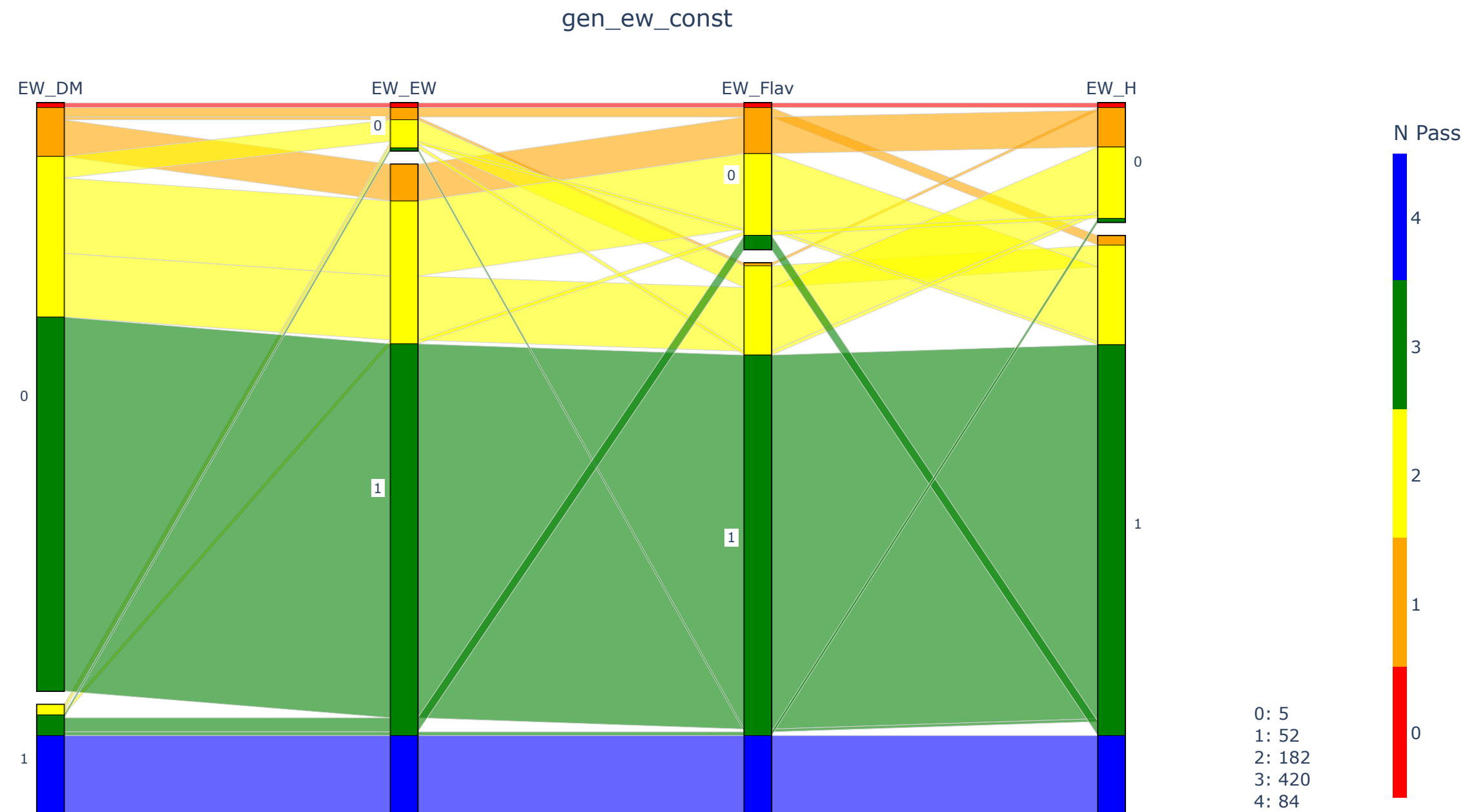
Continuous sparse

Mixed

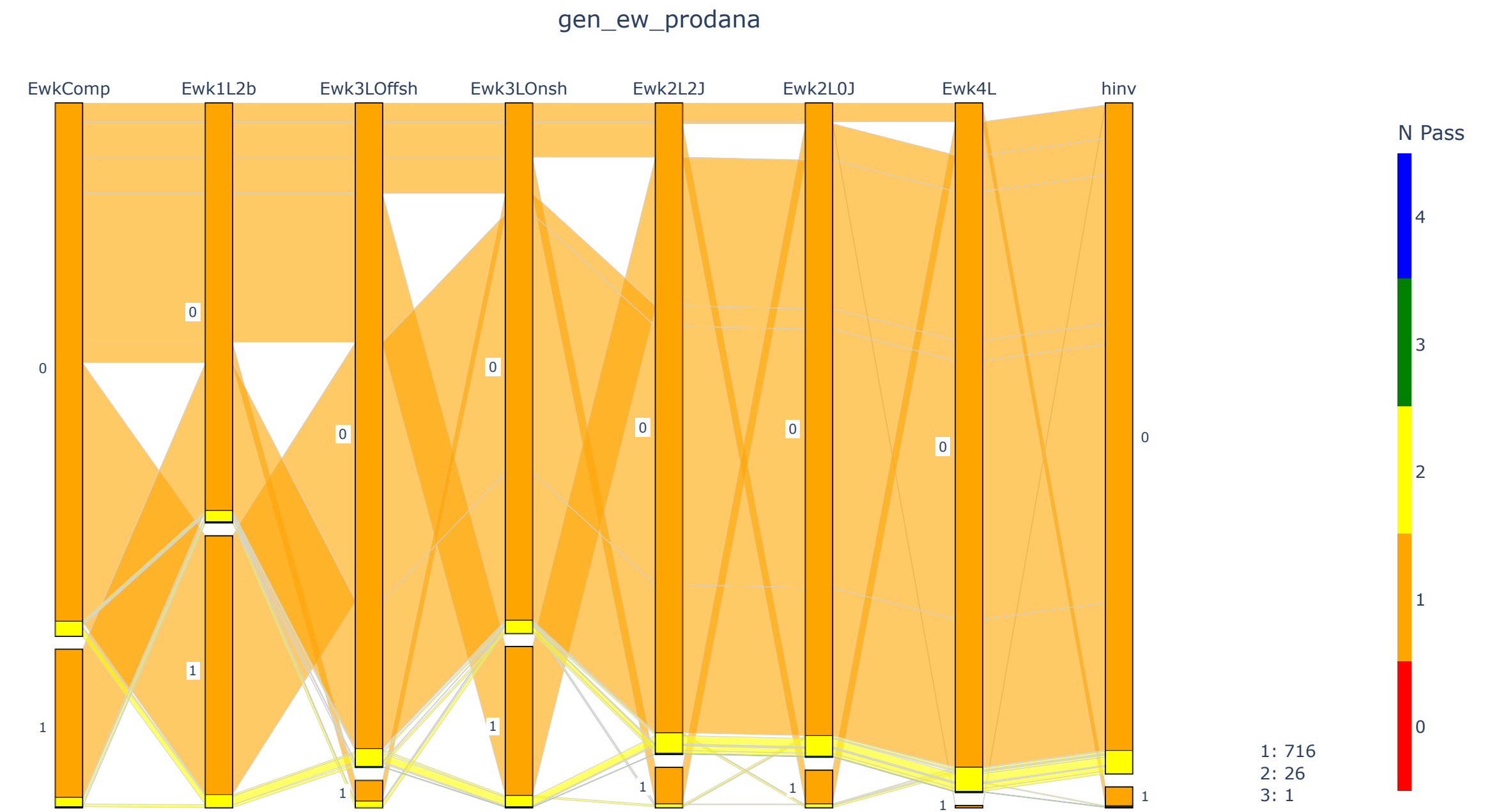


# INPUT FEATURES: BINARY

Number of surviving SE models



External constraints

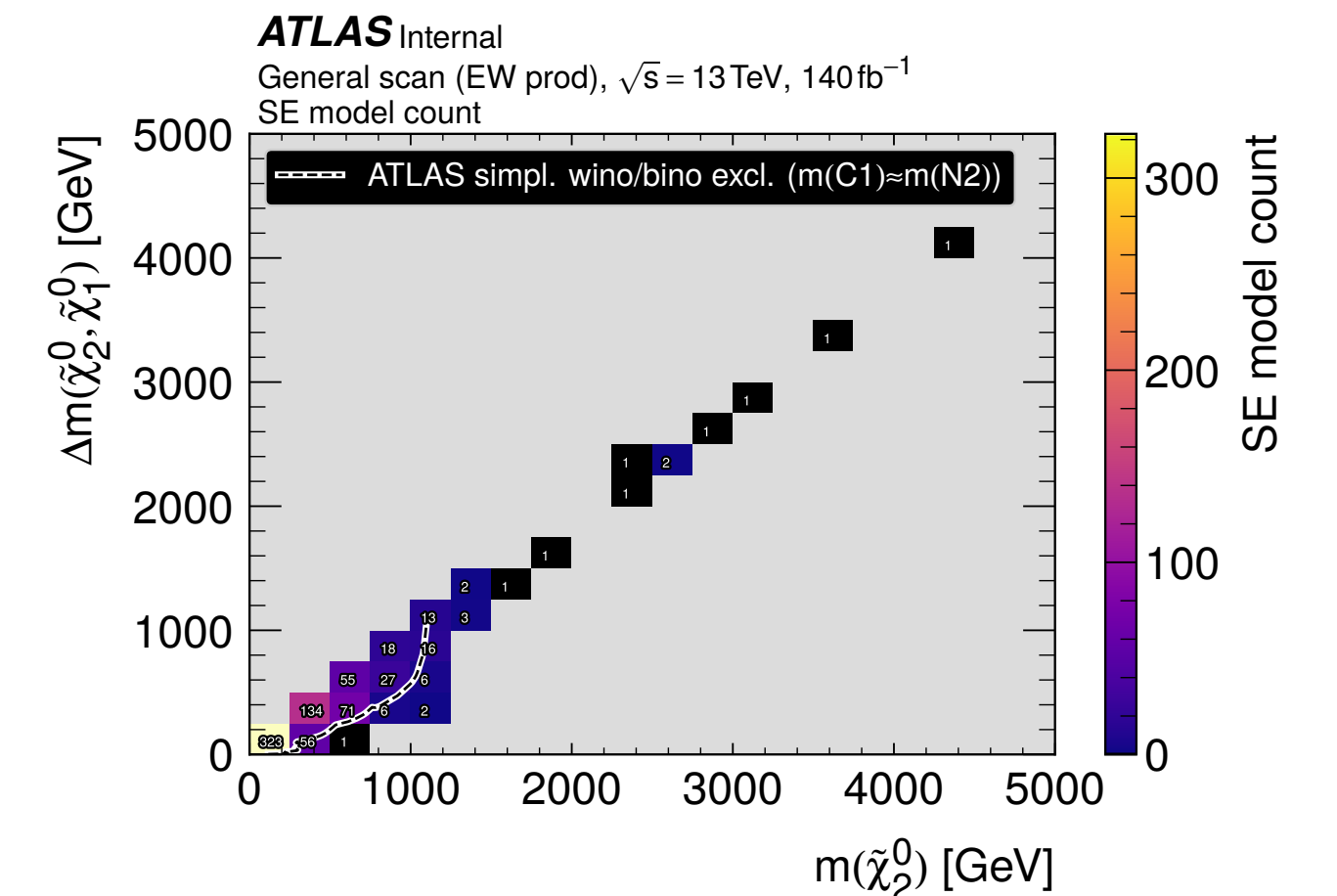
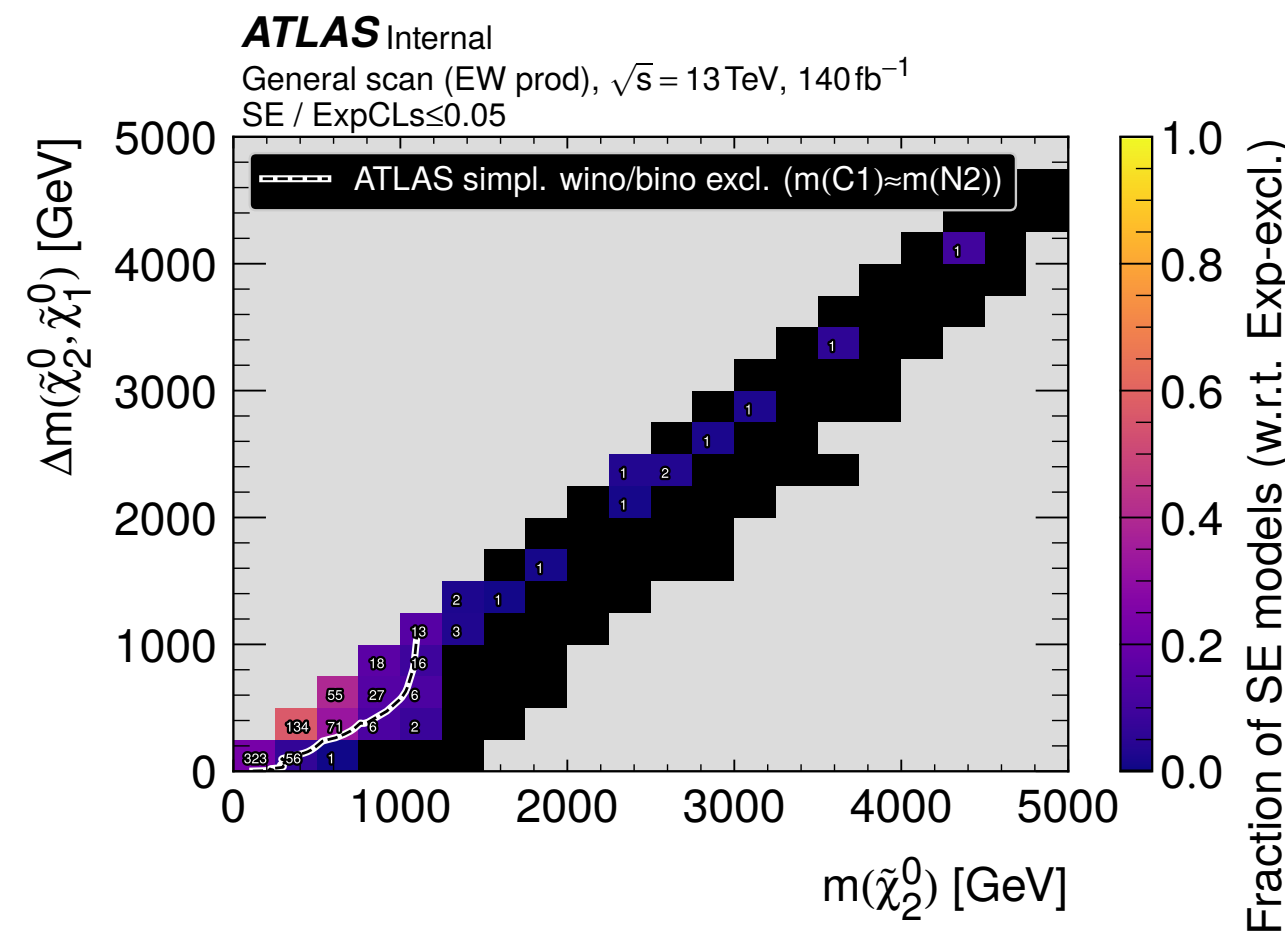
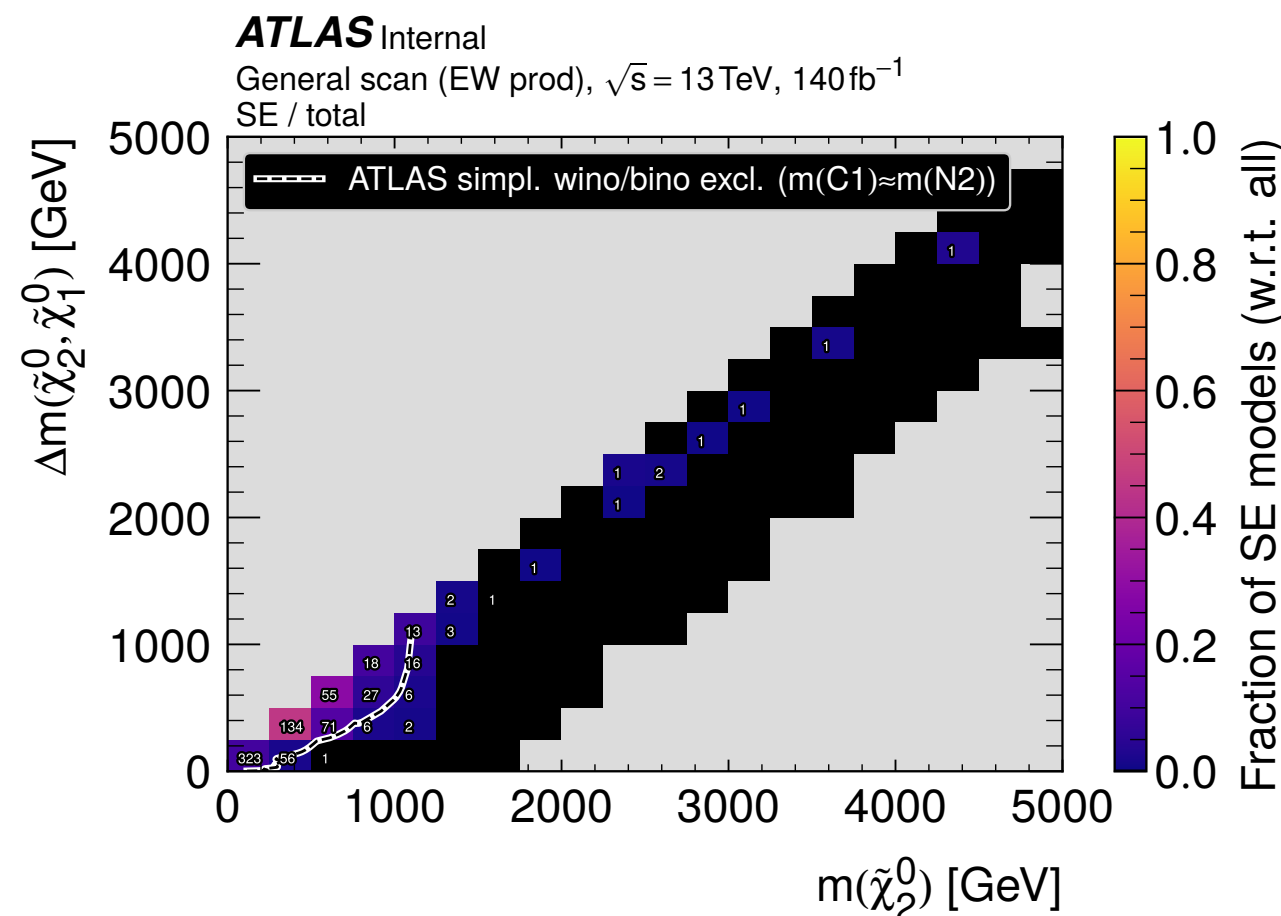
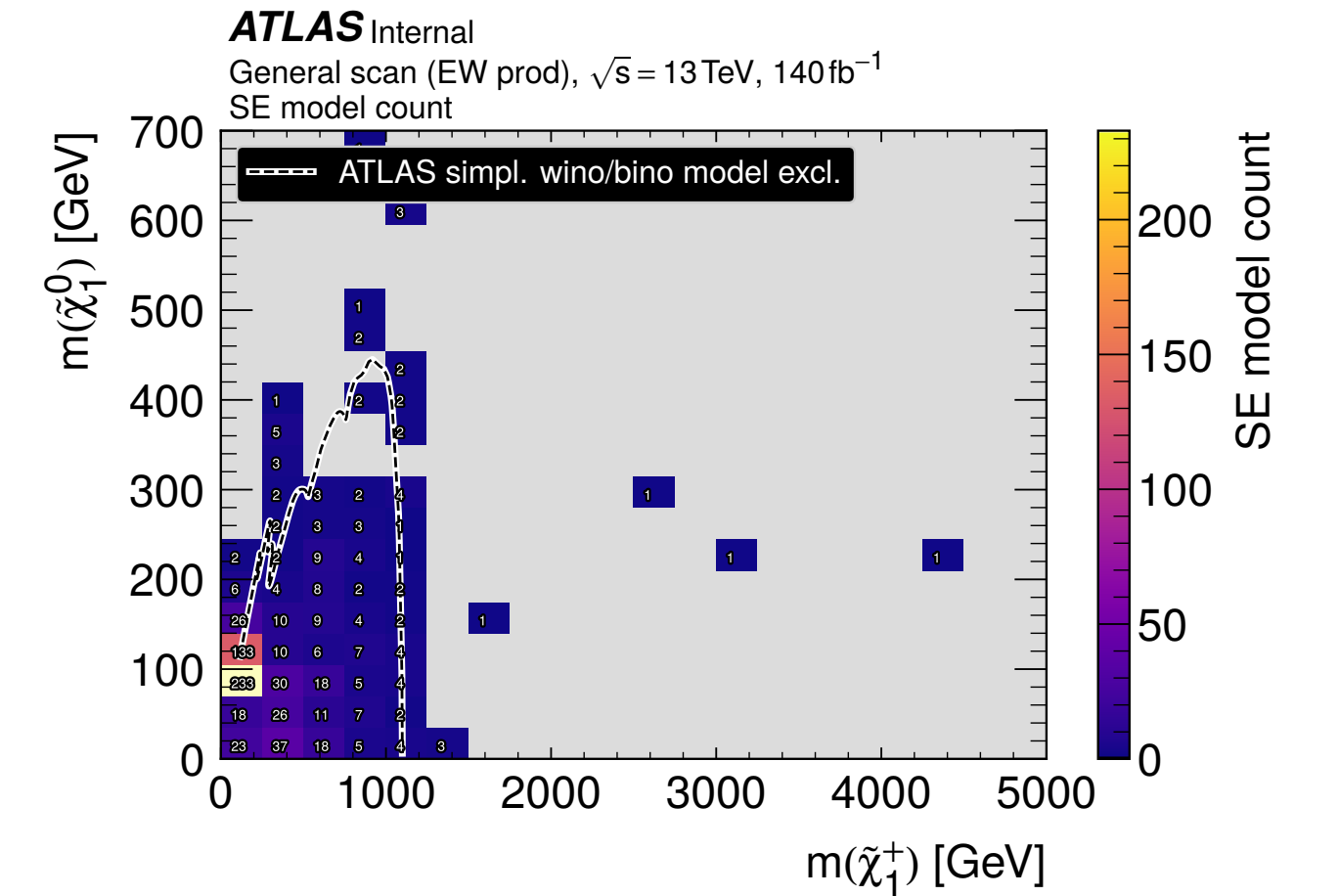
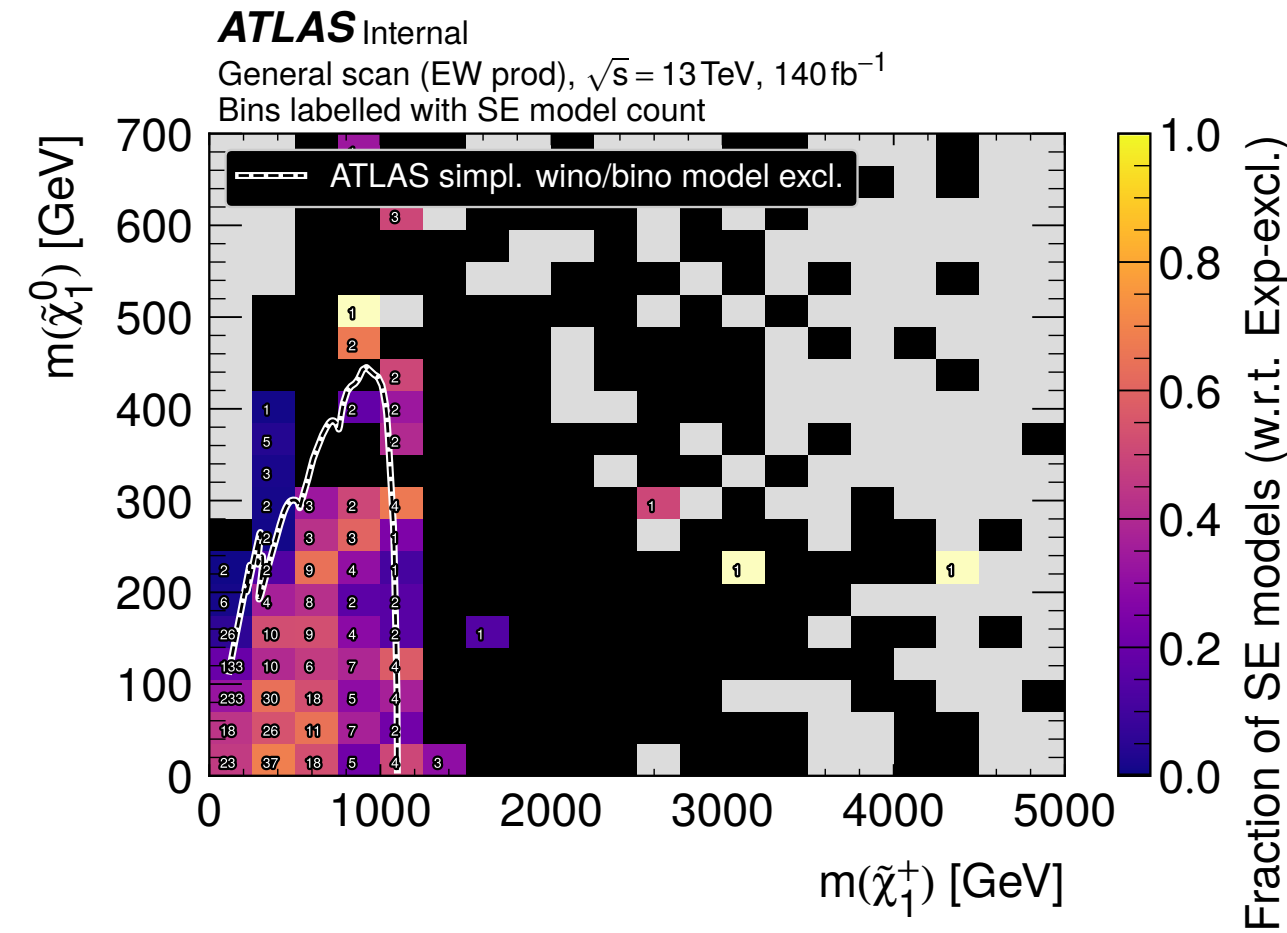
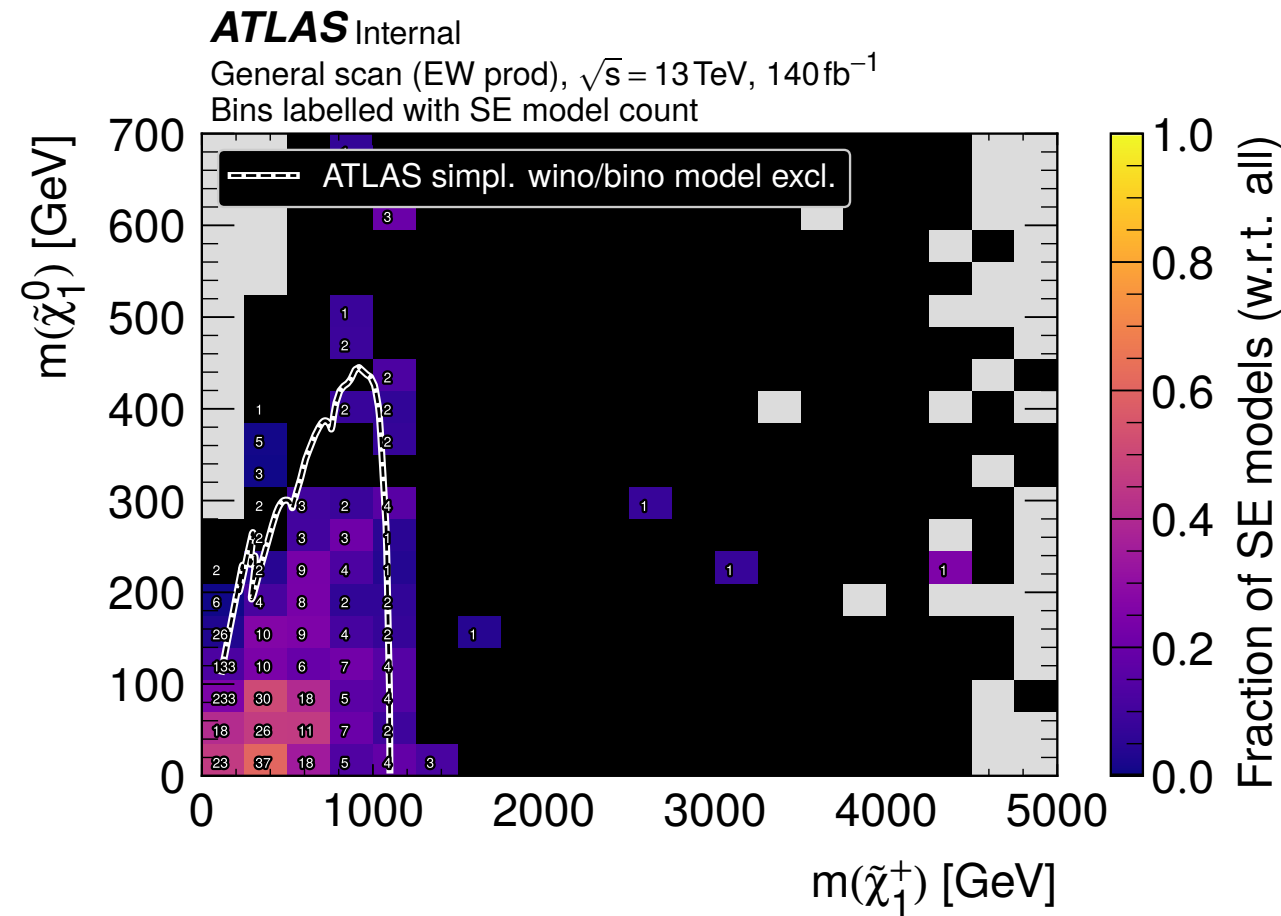


Production × analyses

# INPUT FEATURES: SPARSE CONTINUOUS

SPARSE

## 1.1 Masses – $m(\tilde{\chi}_1^\pm)$ vs $m(\tilde{\chi}_1^0)$ and $m(\tilde{\chi}_2^0)$ vs $\Delta m$

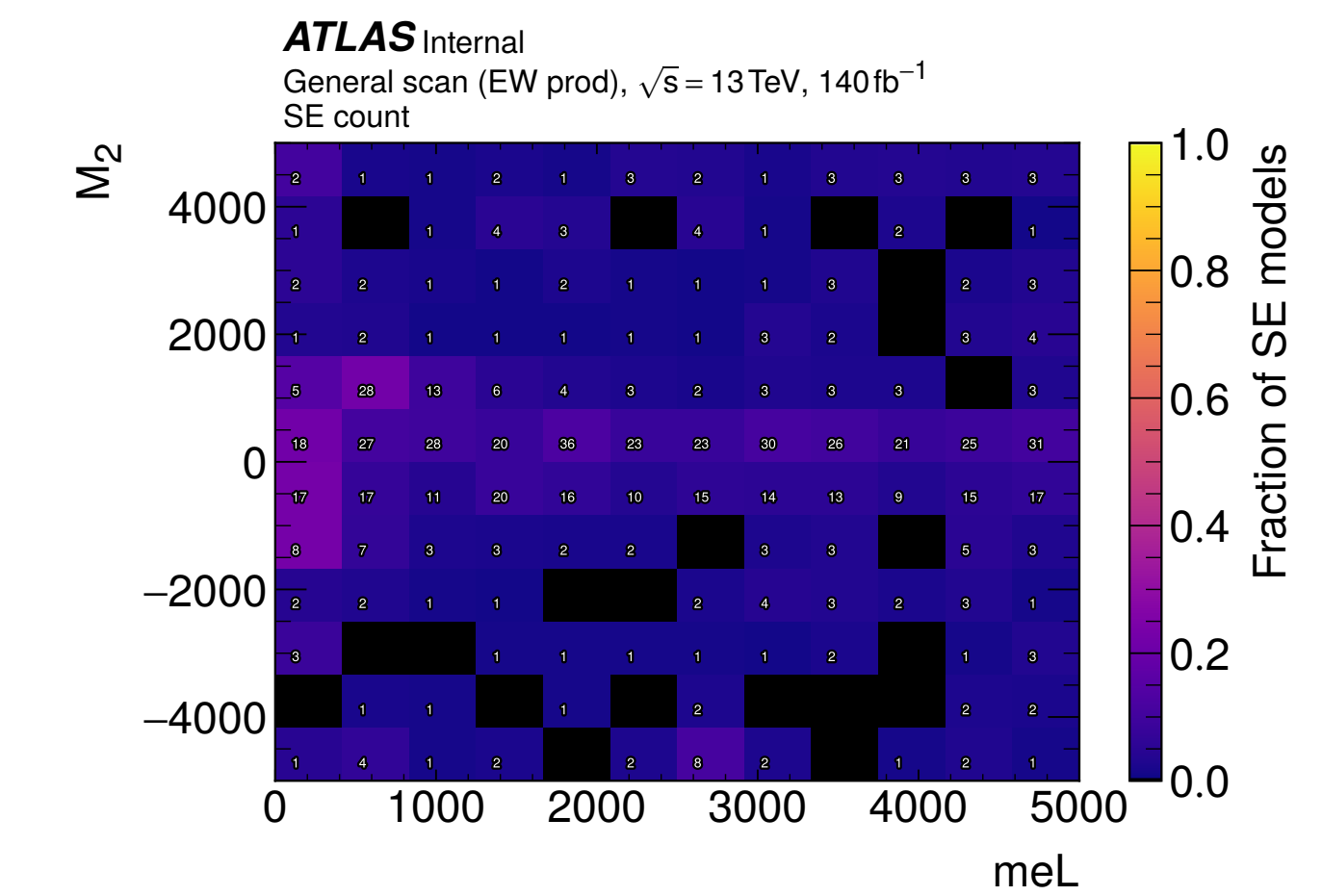
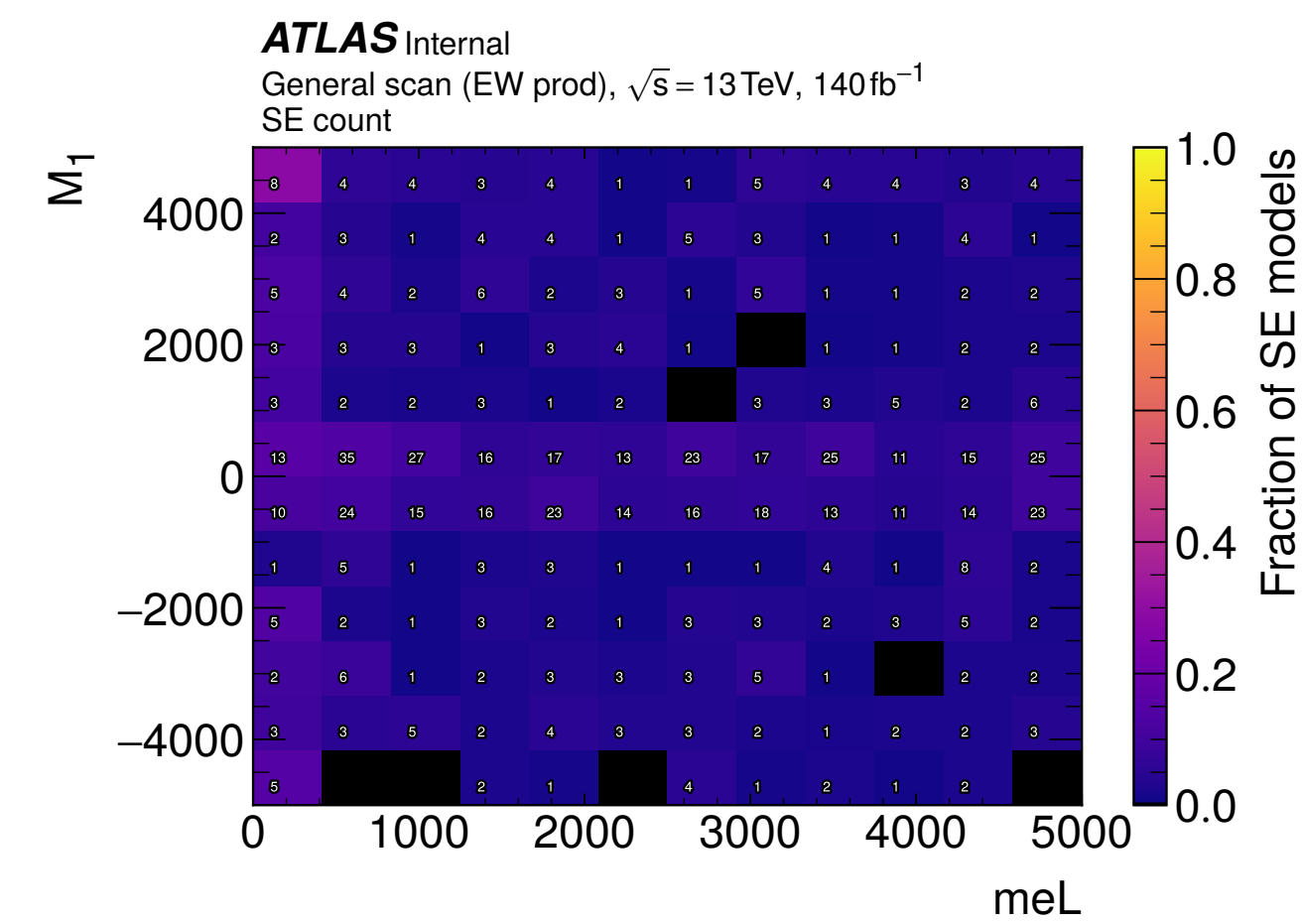
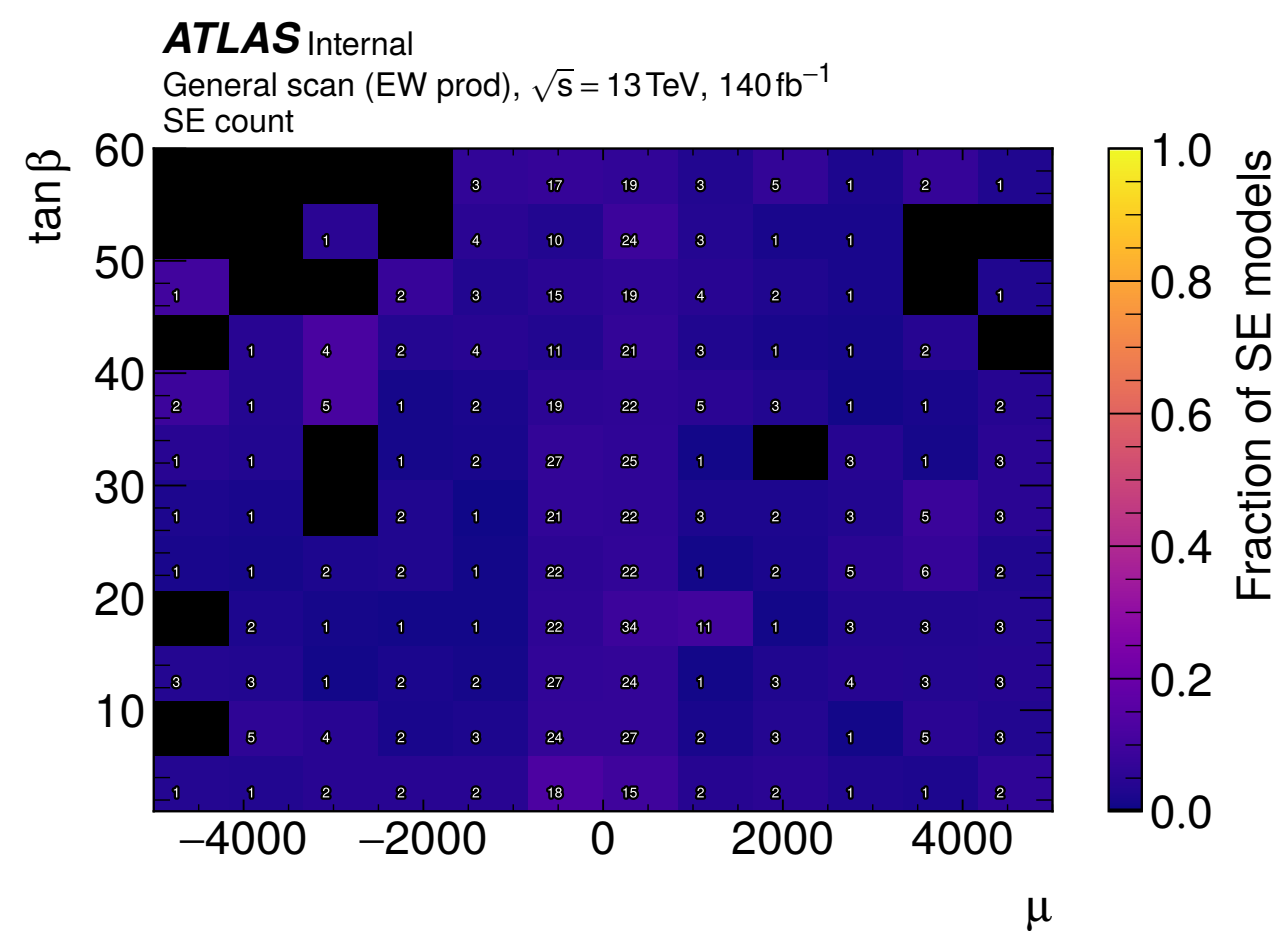
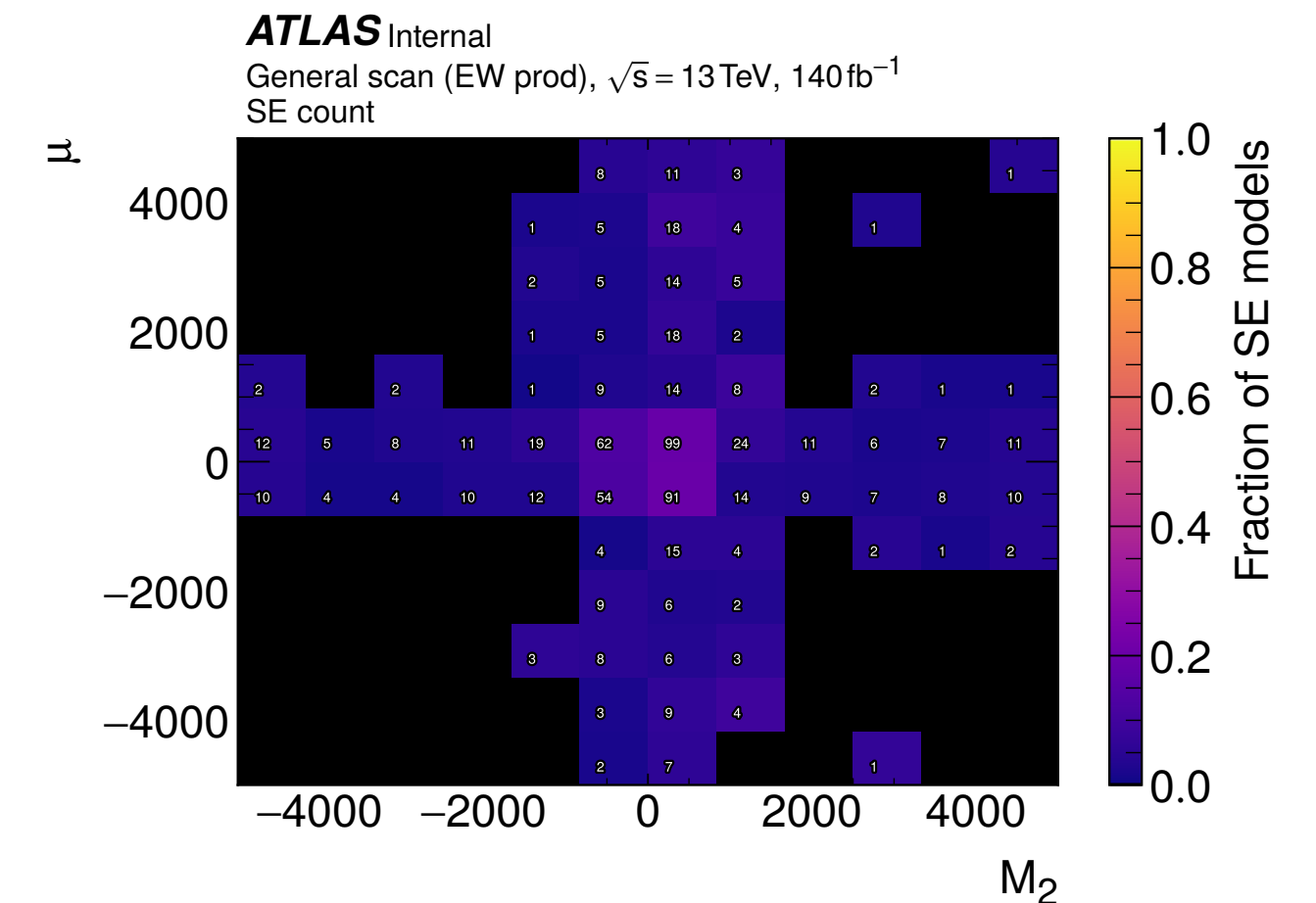
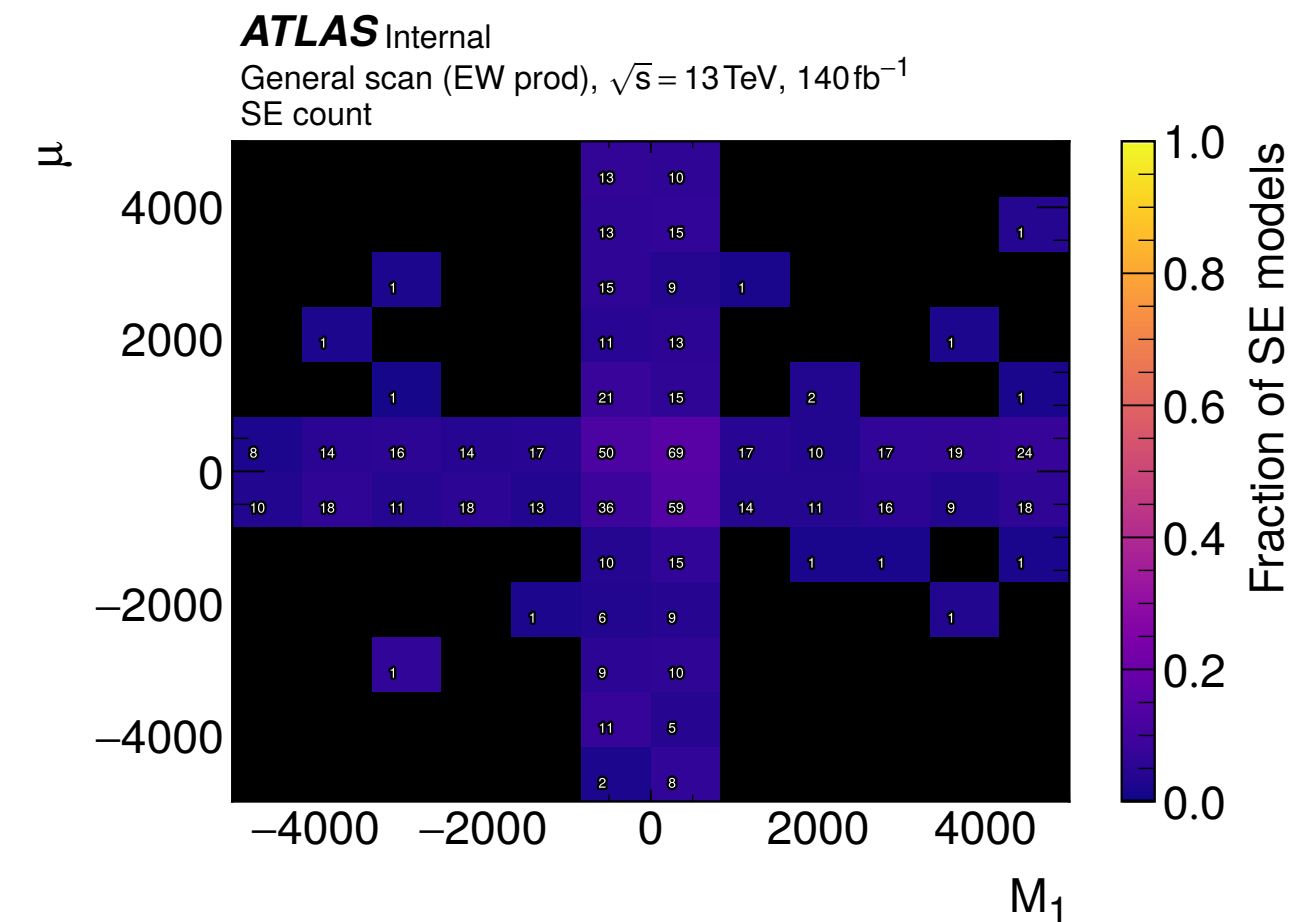
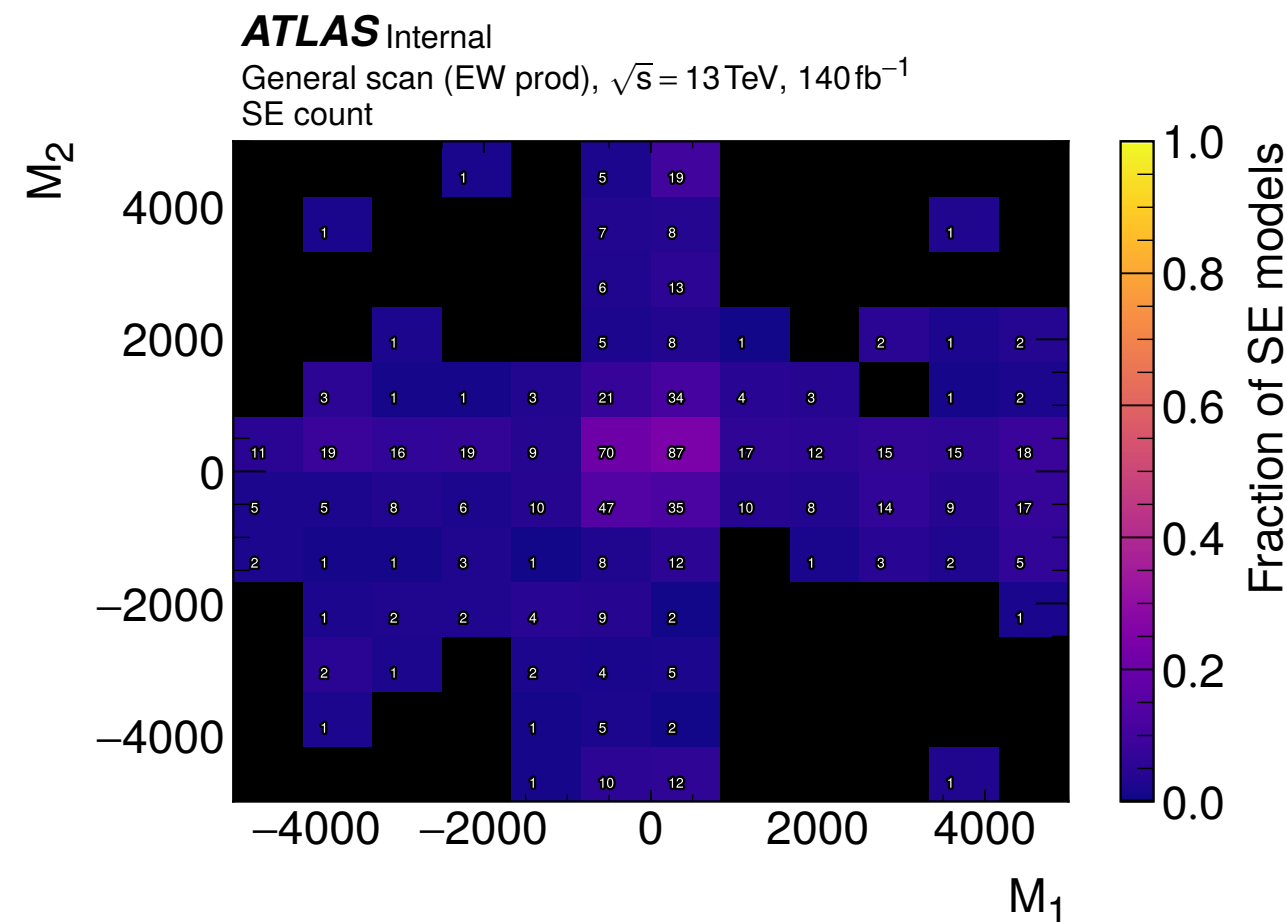




# INPUT FEATURES: SPARSE CONTINUOUS

SPARSE

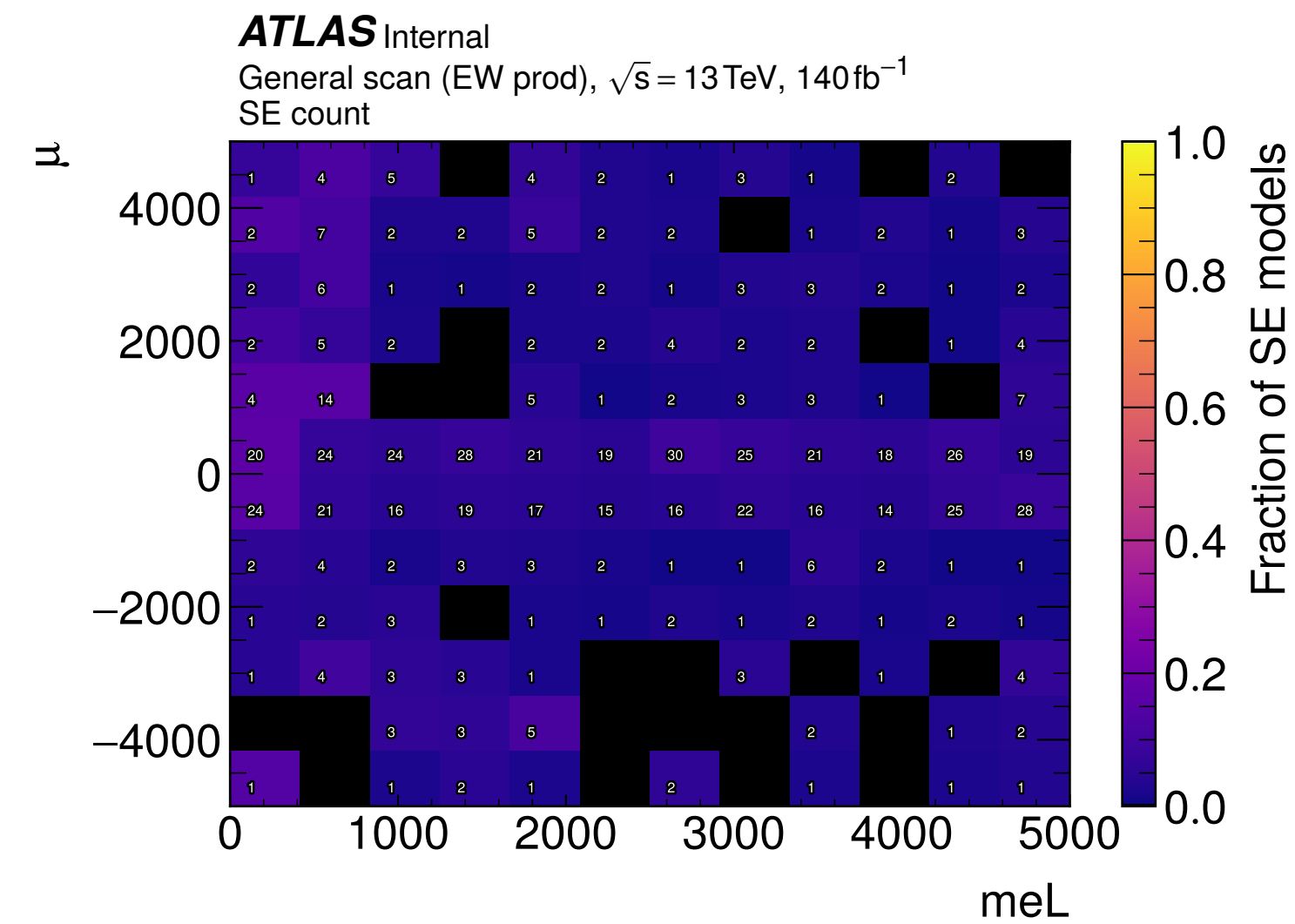
## 1.2 Input params – gaugino mass parameters



# INPUT FEATURES: SPARSE CONTINUOUS

SPARSE

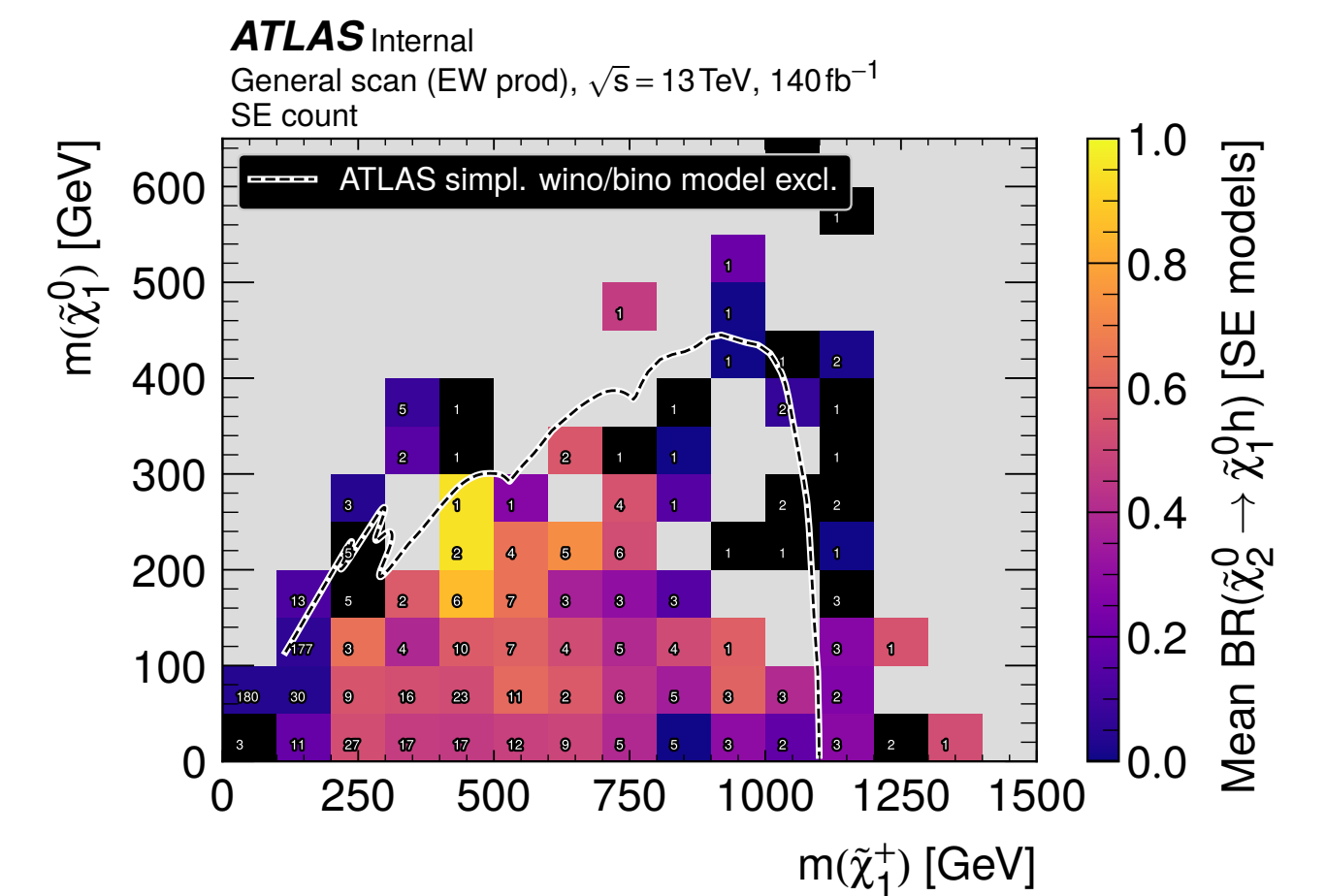
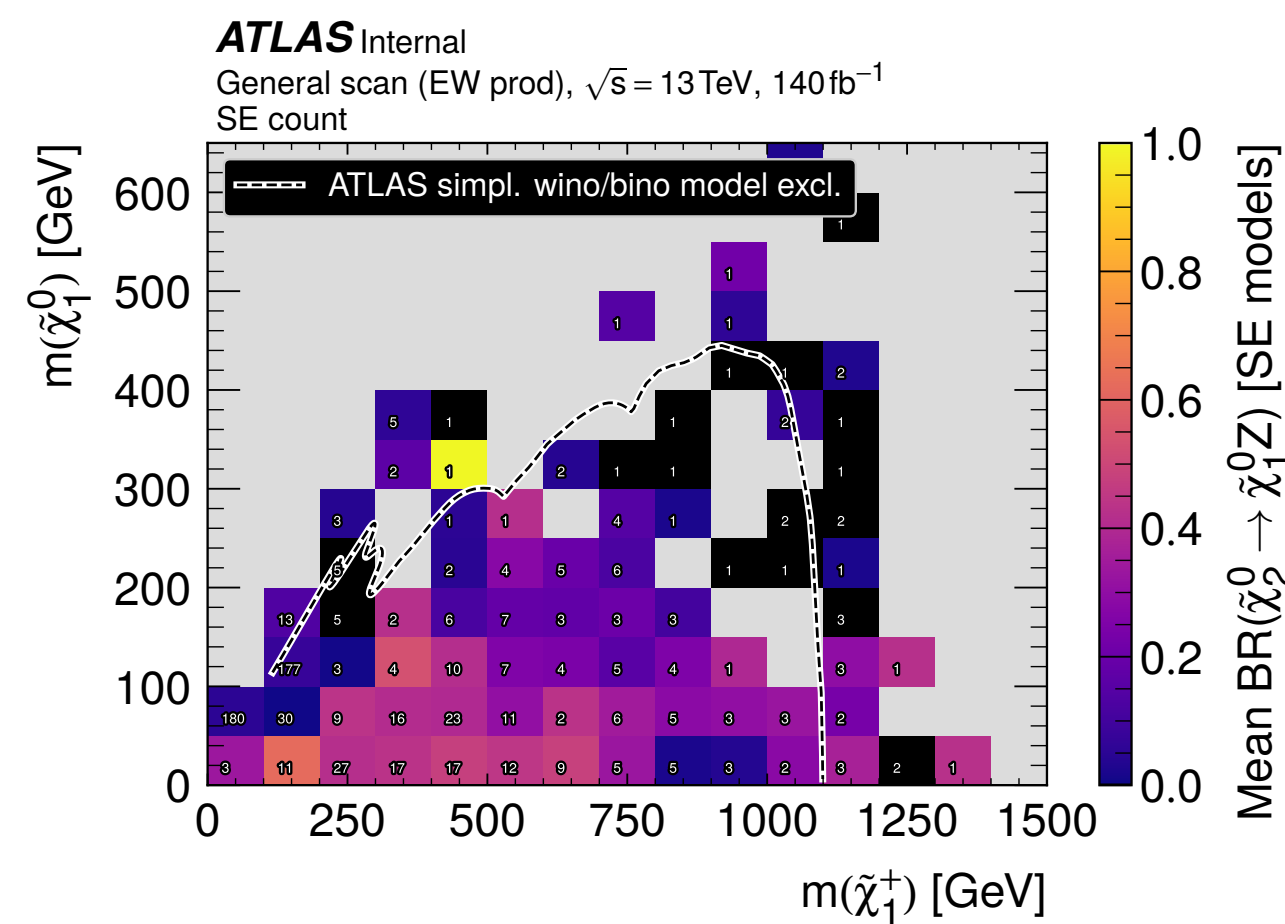
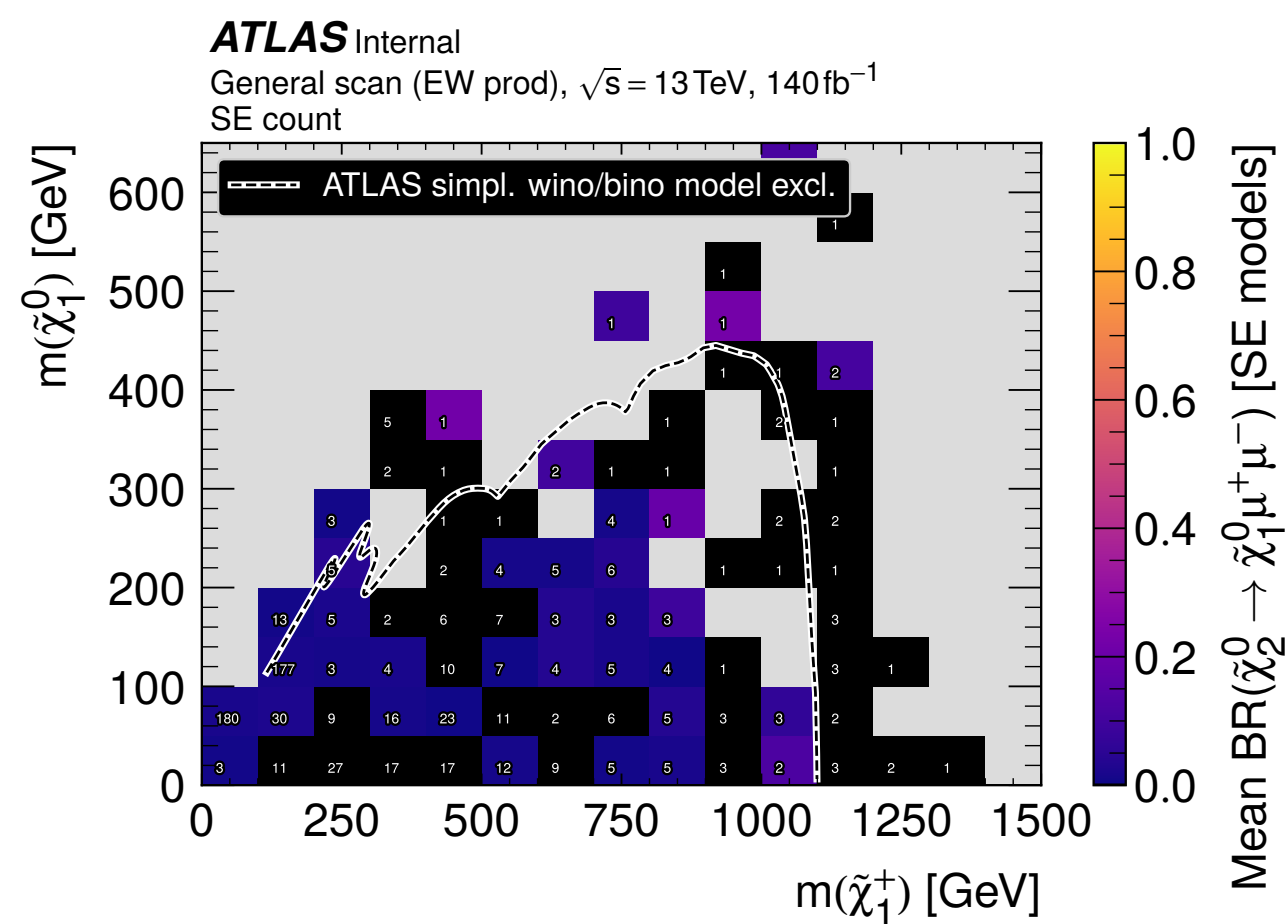
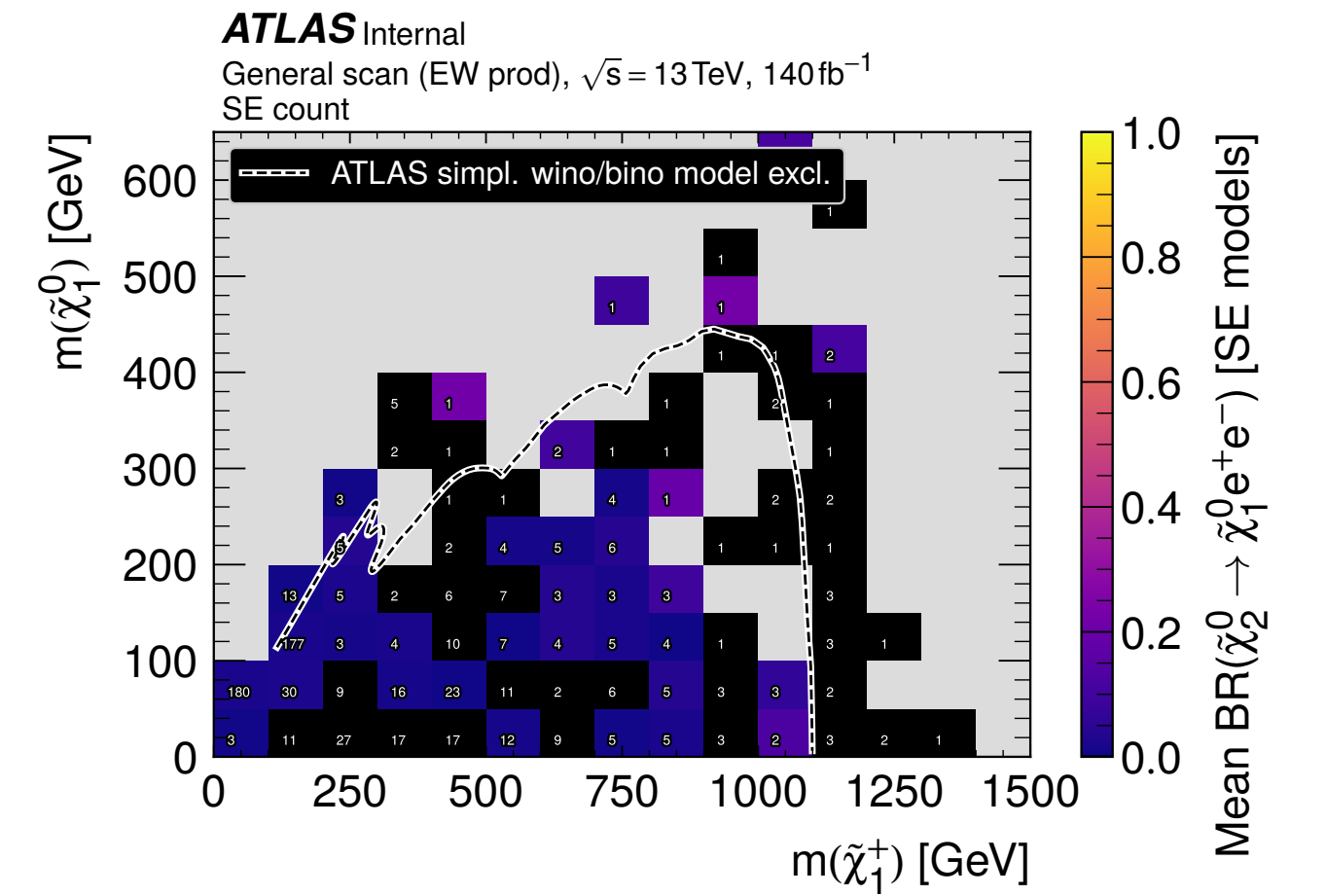
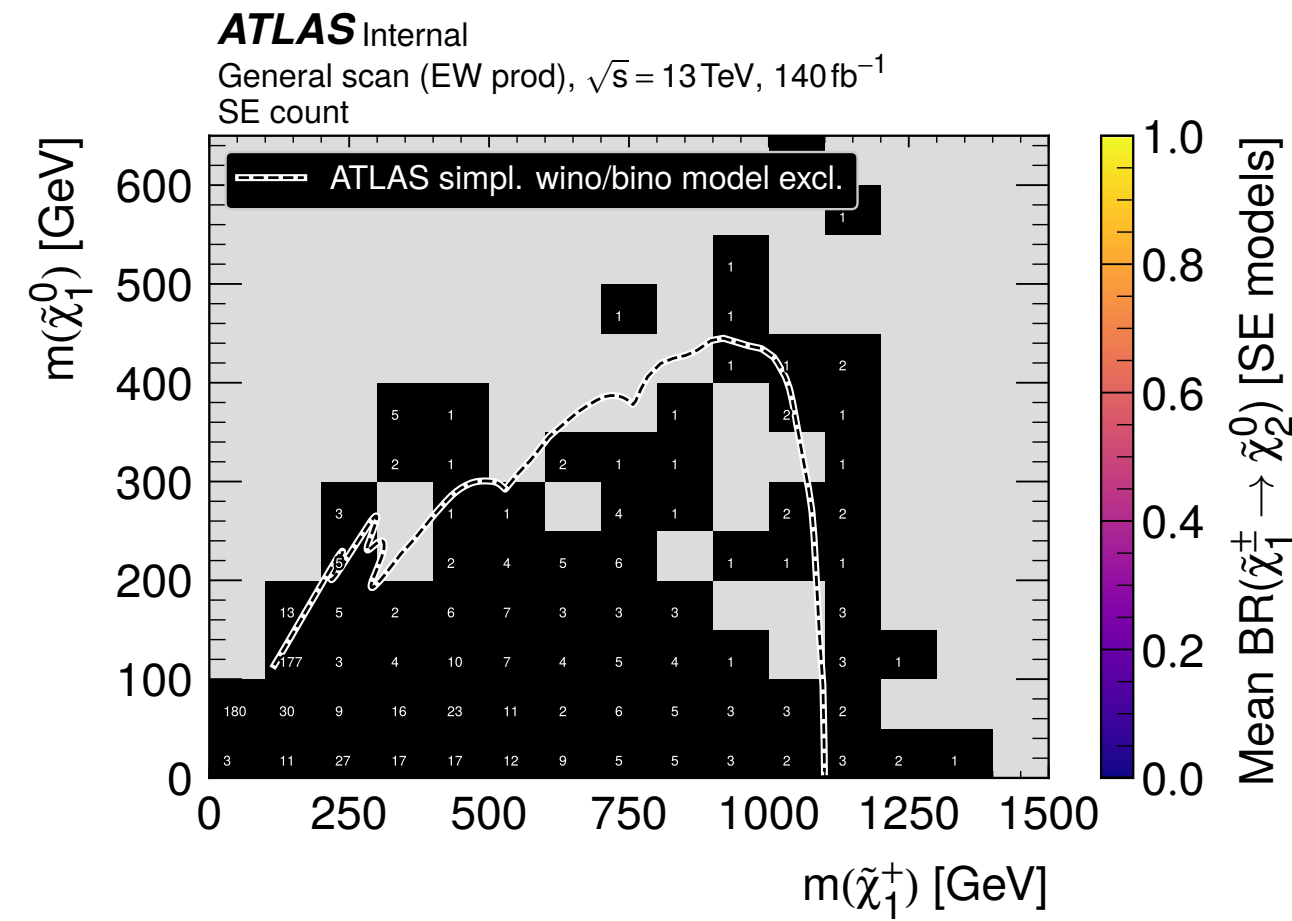
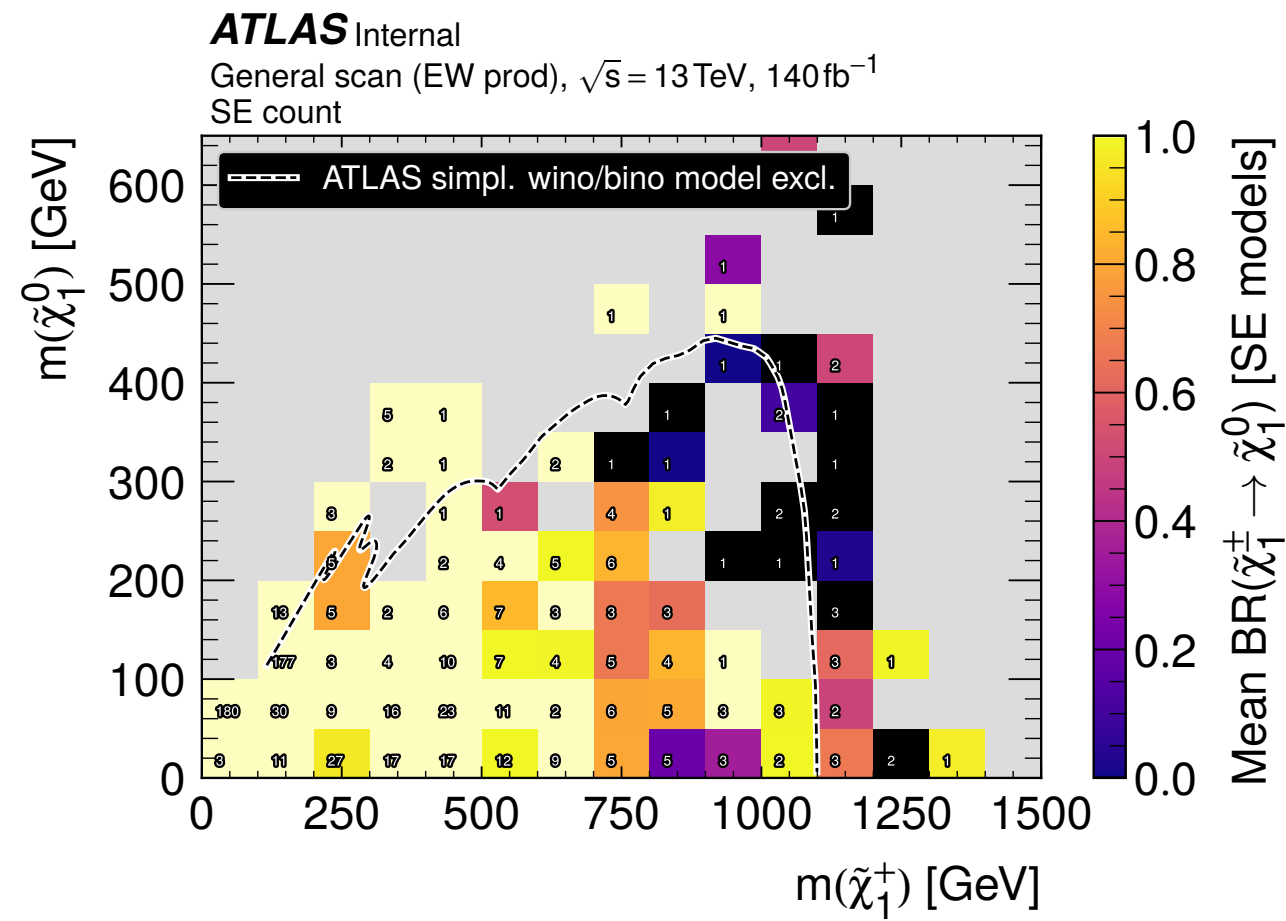
## 1.2 Input params – slepton vs higgsino mass





# INPUT FEATURES: CONTINUOUS

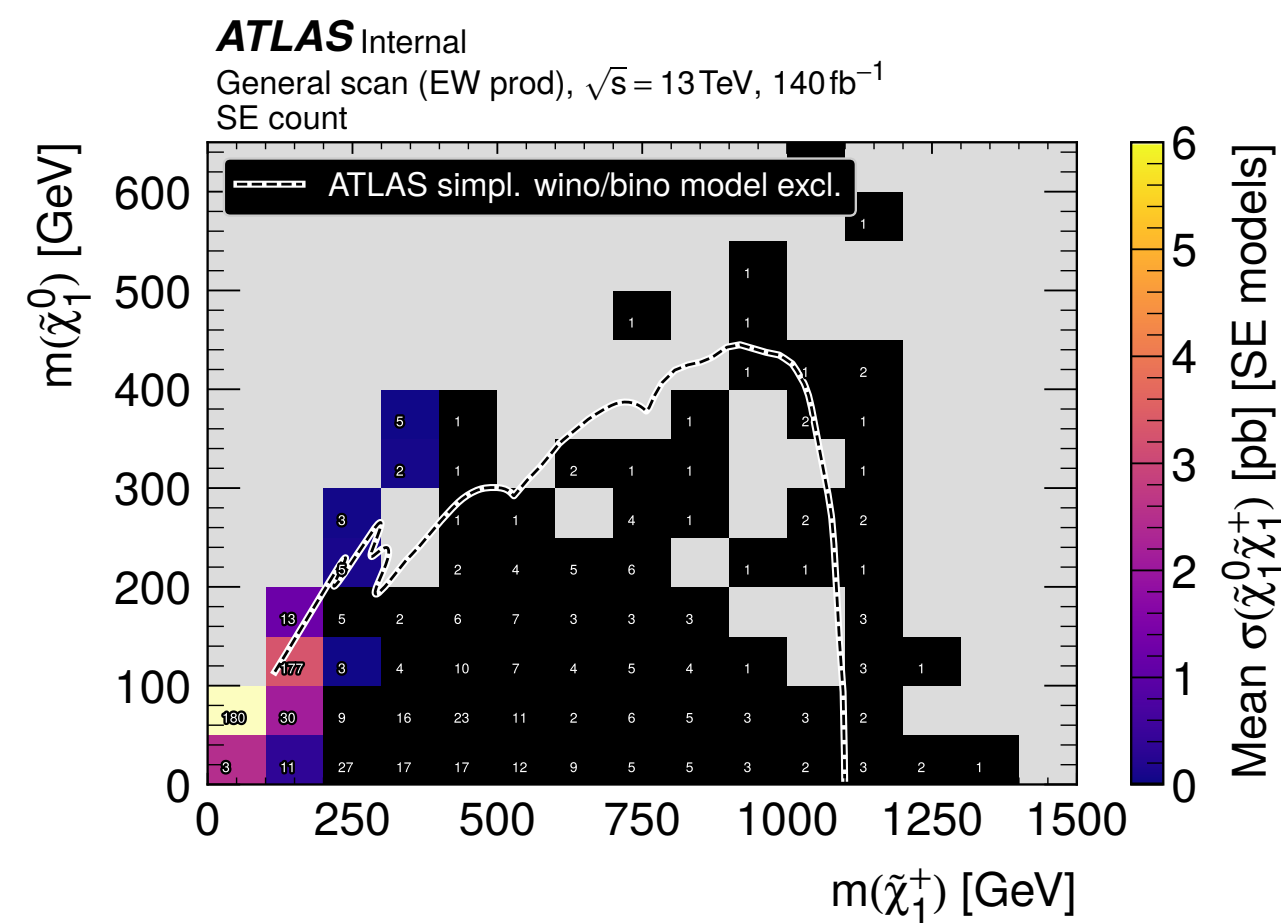
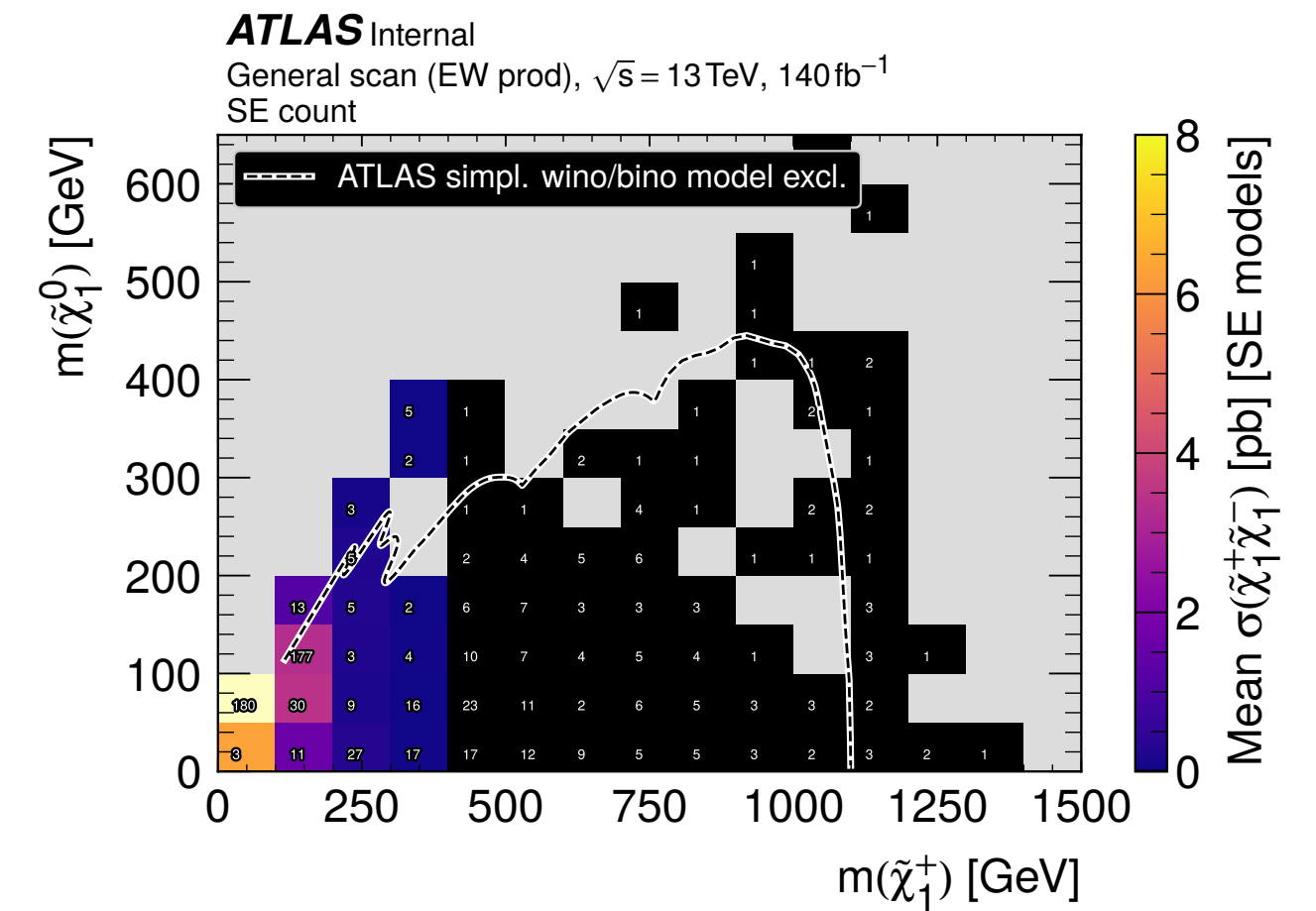
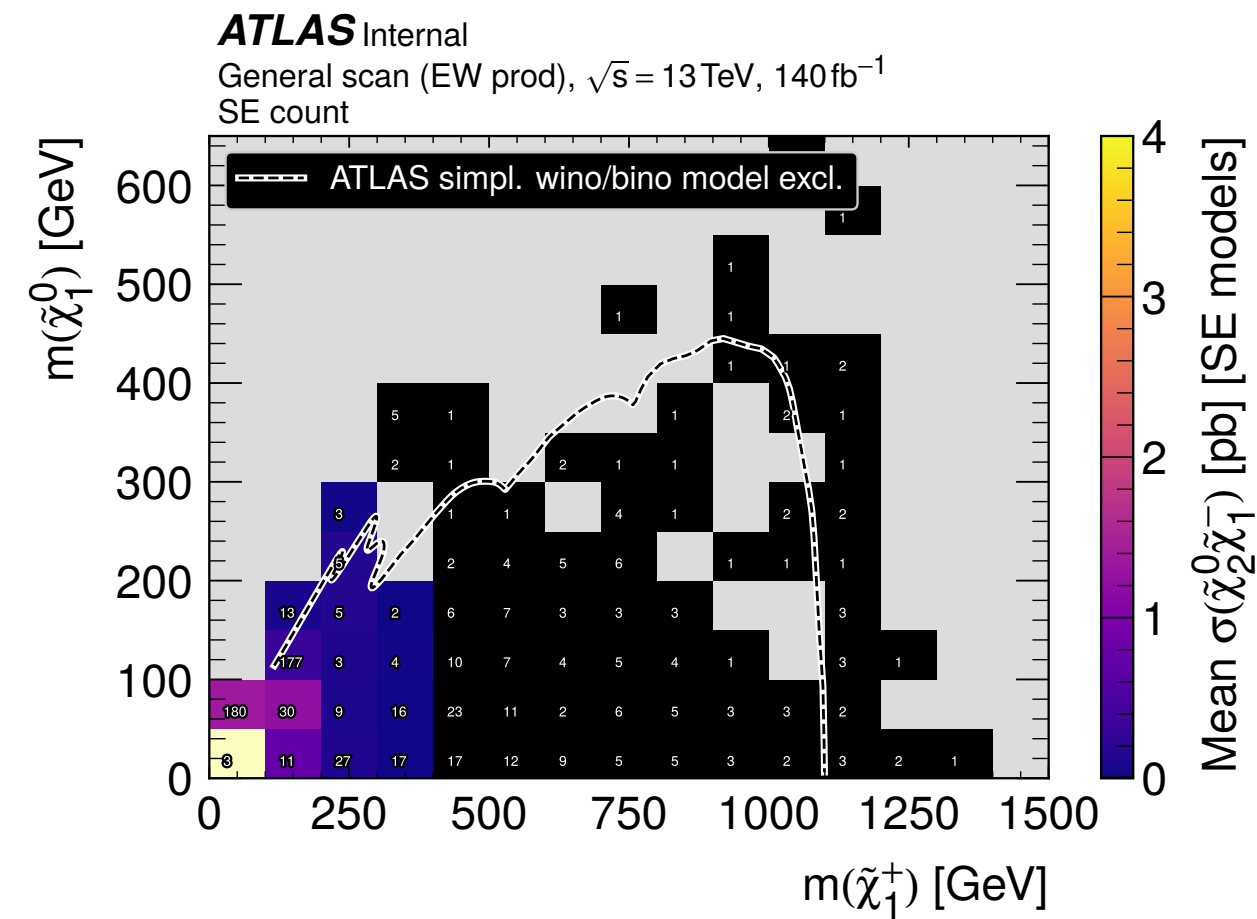
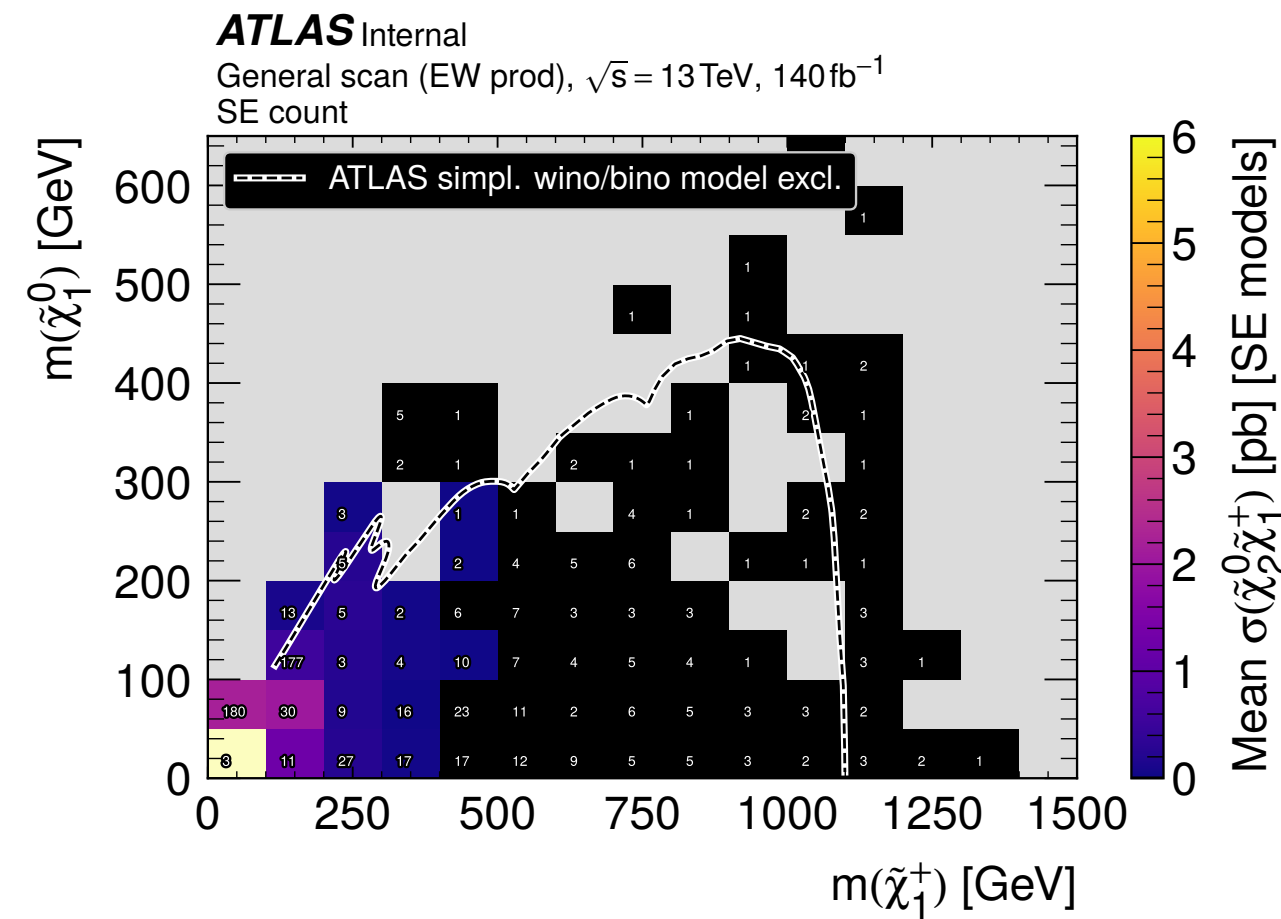
## 1.3 Branching fractions



# INPUT FEATURES: SPARSE CONTINUOUS

SPARSE

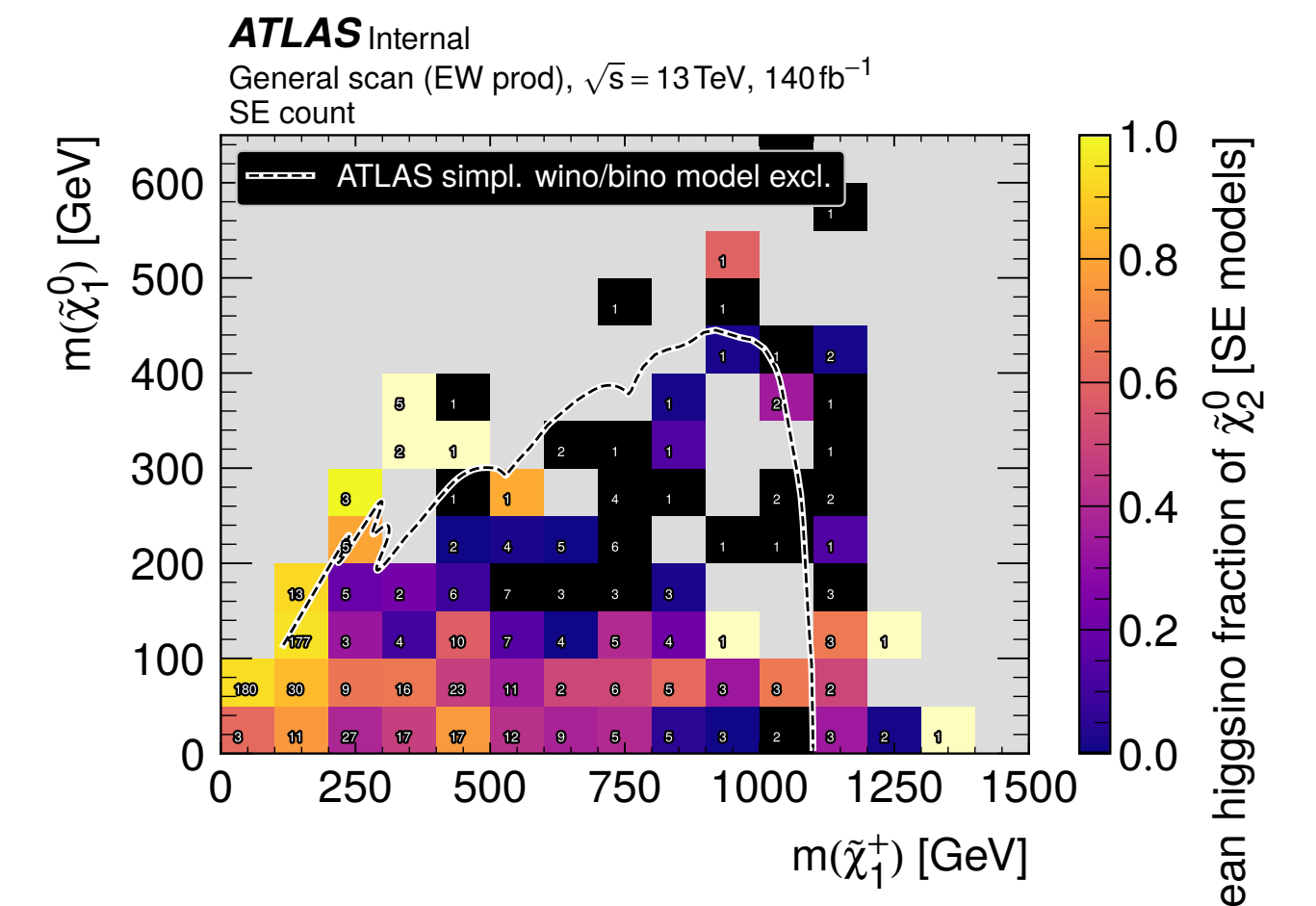
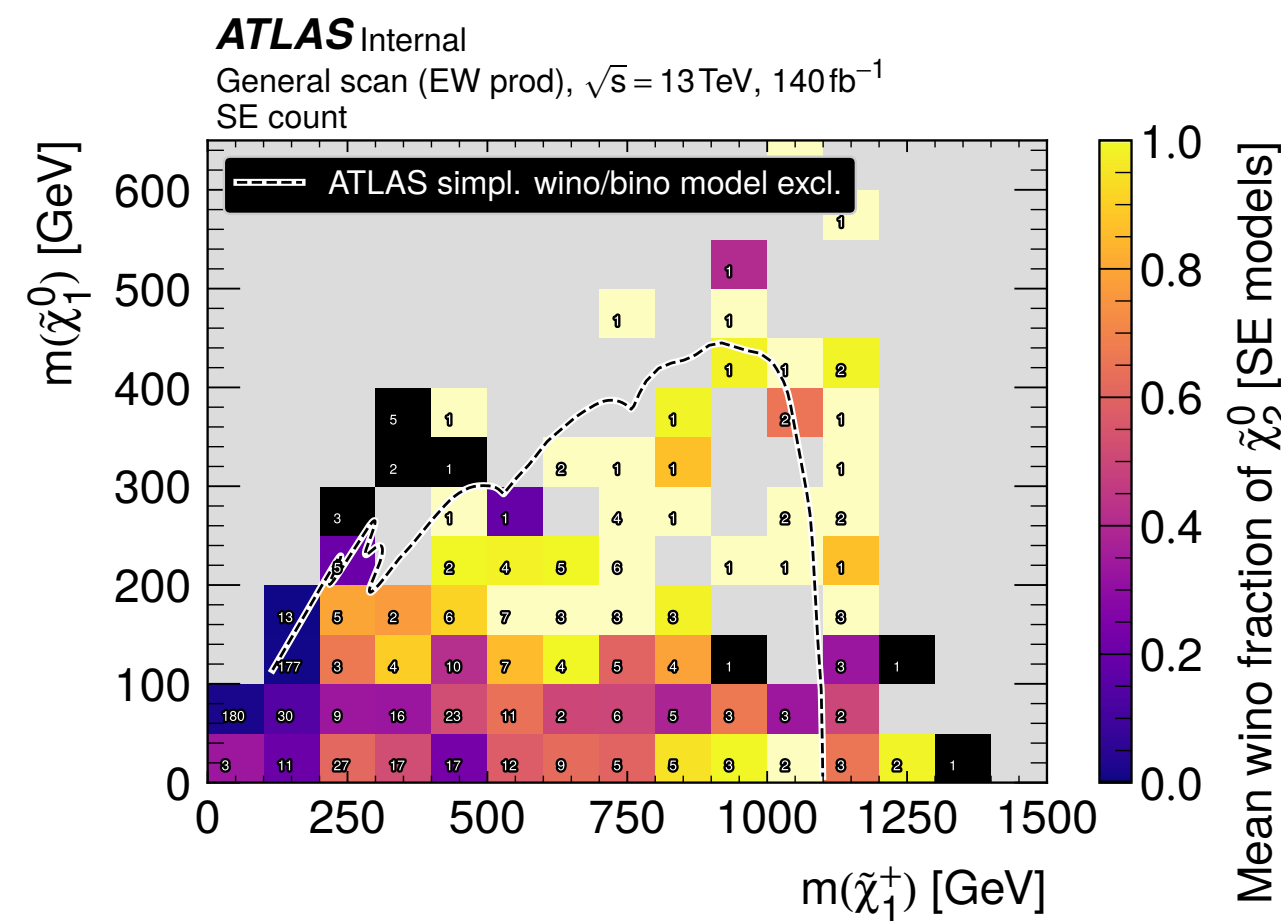
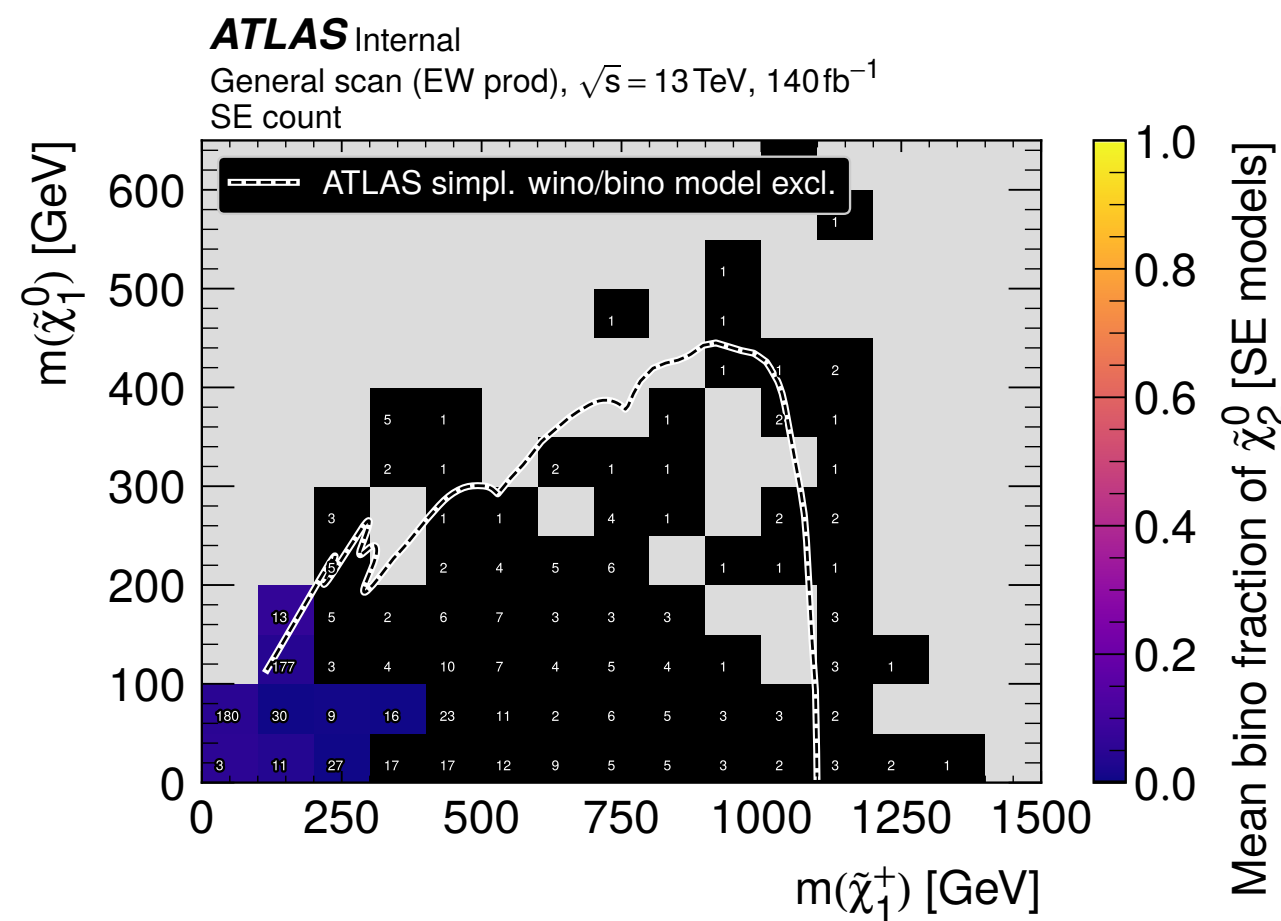
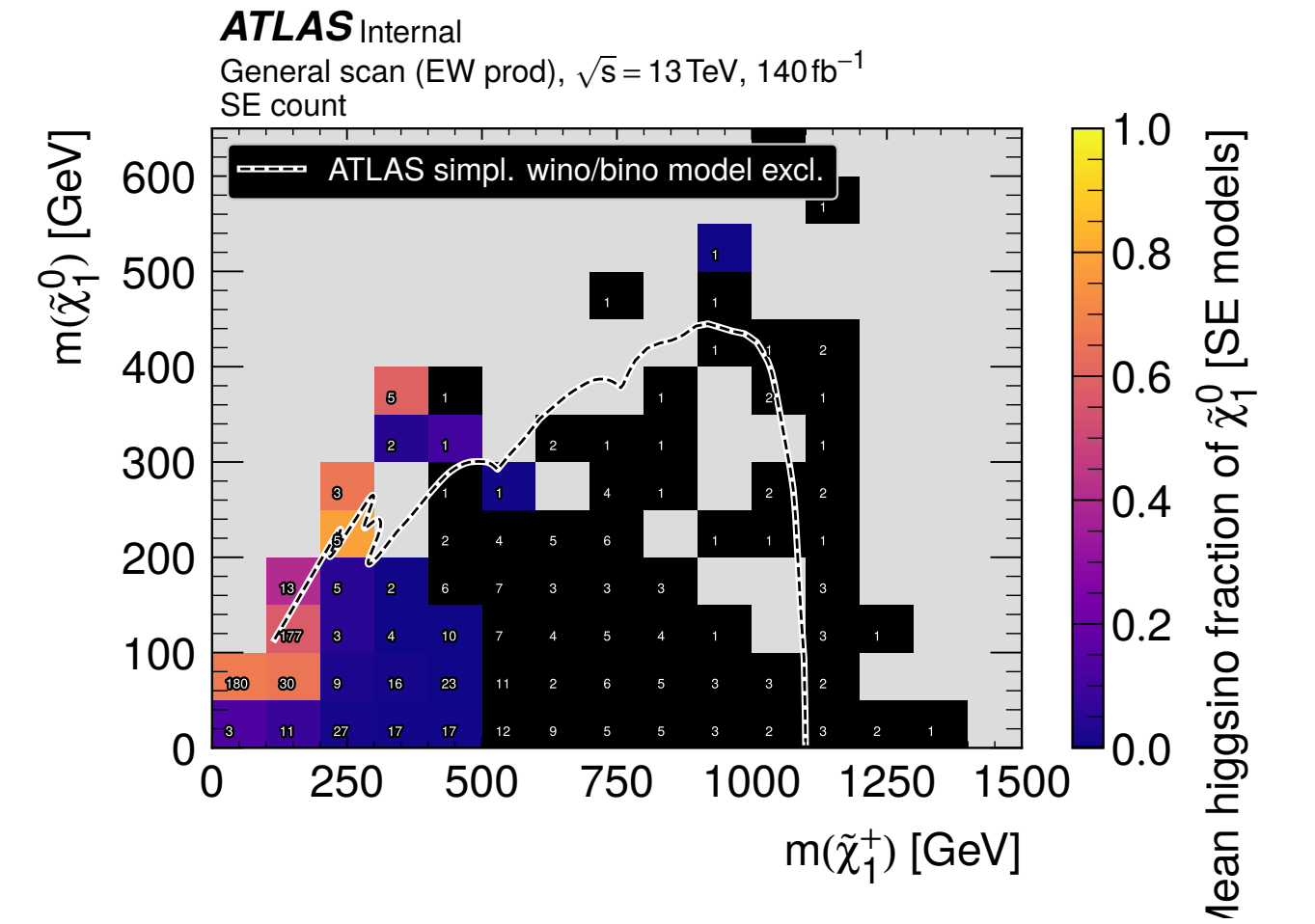
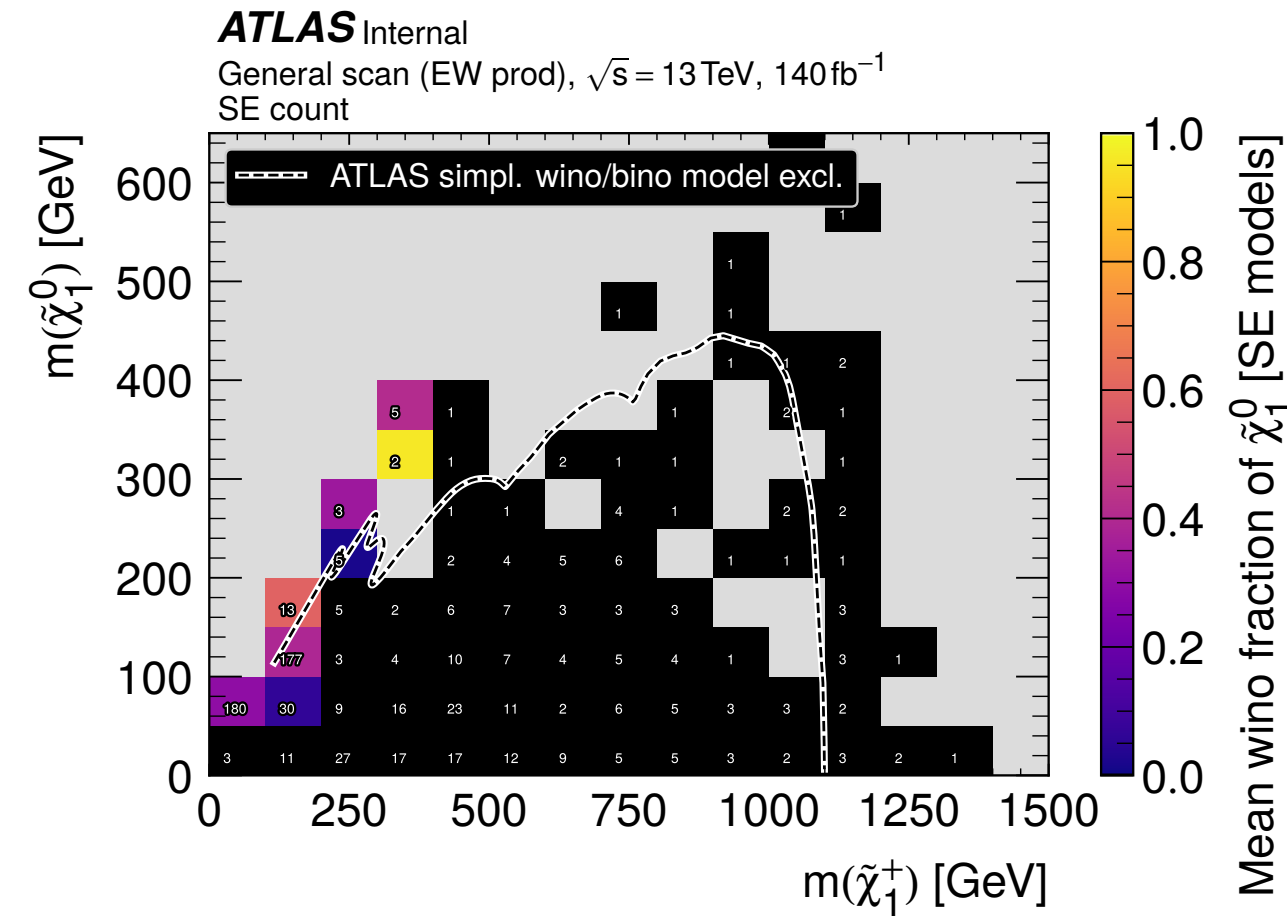
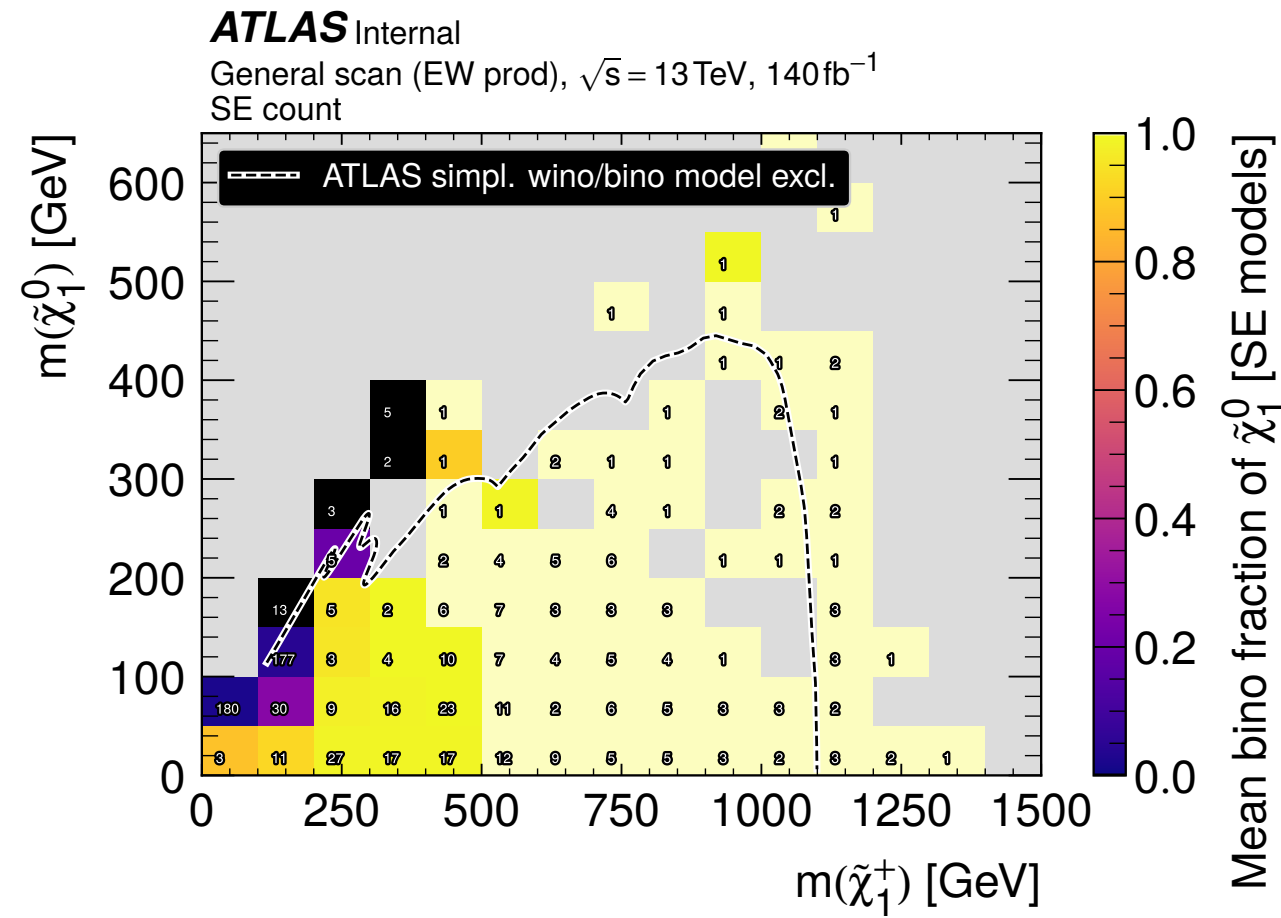
## 1.4 Cross-sections





# INPUT FEATURES: CONTINUOUS

## 1.5 LSP and NLSP mixing



# PREPROCESSING

PREPROCESSING

- Some input features can benefit from preprocessing. Some clustering algorithms handle, e.g. sparse distributions better than others, but proper preprocessing should improve the clustering algorithm's performance.
- Investigating different types of preprocessing for different types of input, and evaluating whether there is improvement.

| Input type          | Test                                           | Comment                                                                                                                                                                                                                                                                                             |
|---------------------|------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| (Sparse) Continuous | Binary                                         | Number of distinct patterns                                                                                                                                                                                                                                                                         |
|                     | Hopkins statistics                             | N models per distinct flag pattern — discrete structure a binary clusterer can actually find:<br><b>few heavy patterns = natural candidate clusters</b><br>many near-singleton patterns = little to cluster                                                                                         |
|                     | Pairwise distance and coefficient of variation | $\approx 0.5$ no structure, $\rightarrow 1$ clusterable, $\rightarrow 0$ regular grid (Hopkins & Skellam 1954, <i>*Ann. Bot.*</i> 18:213; Banerjee & Davé 2004, FUZZ-IEEE)                                                                                                                          |
|                     | VAT ordered disssimilarity                     | Measures how useful pairwise distances are for clustering using relative contrast and coeffiecient of variation, <b>very large values indicate likely near-duplicate rows.</b>                                                                                                                      |
|                     |                                                | Reorders the dissimilarity matrix so similar observations appear next to each other, making cluster structure visible. <b>Clear dark diagonal blocks suggest distinct clusters; sharper blocks after preprocessing indicate improved separation, while diffuse patterns suggest weak structure.</b> |

Transformations:

- **Log** — compresses large values and reduces skew.
- **PCA** — projects data onto directions of maximum variance.
- **UMAP** — creates a low-dimensional embedding that preserves local structure.
- **t-SNE** — creates a low-dimensional embedding emphasizing local neighborhood separation.

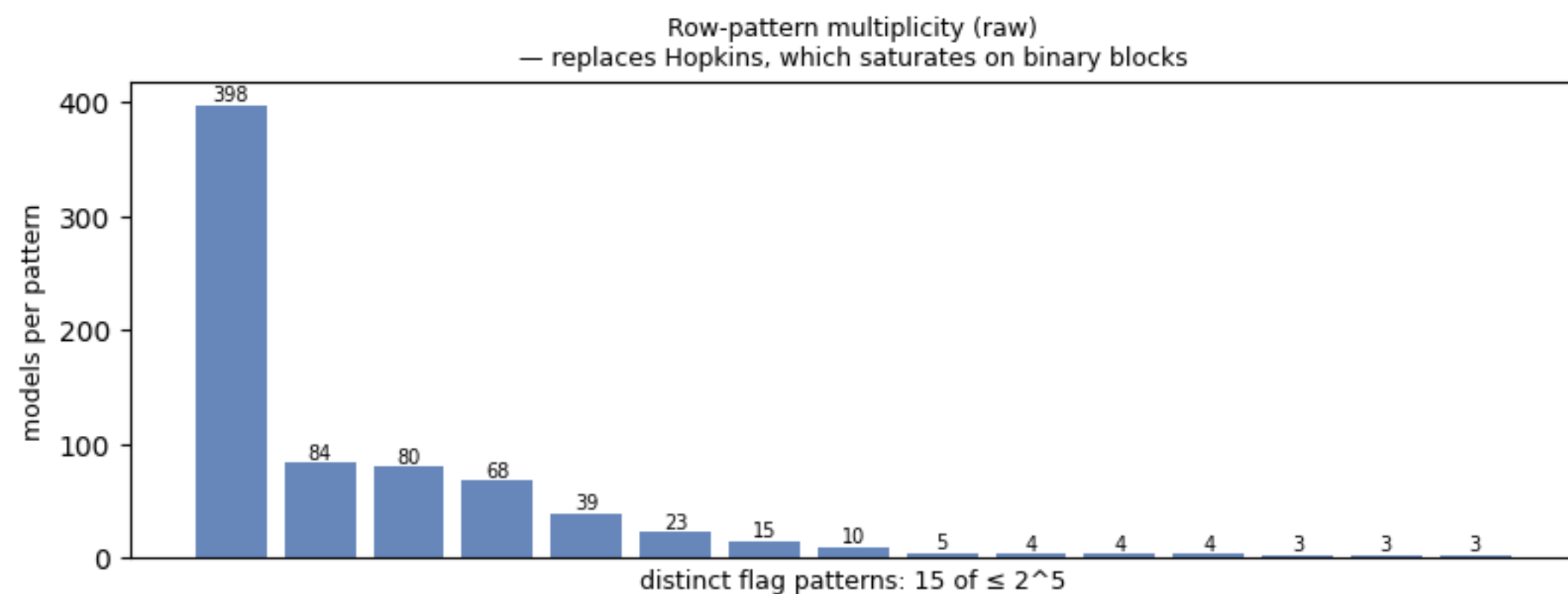


# PREPROCESSING: BINARY - NO PREPROCESSING

PREPROCESSING

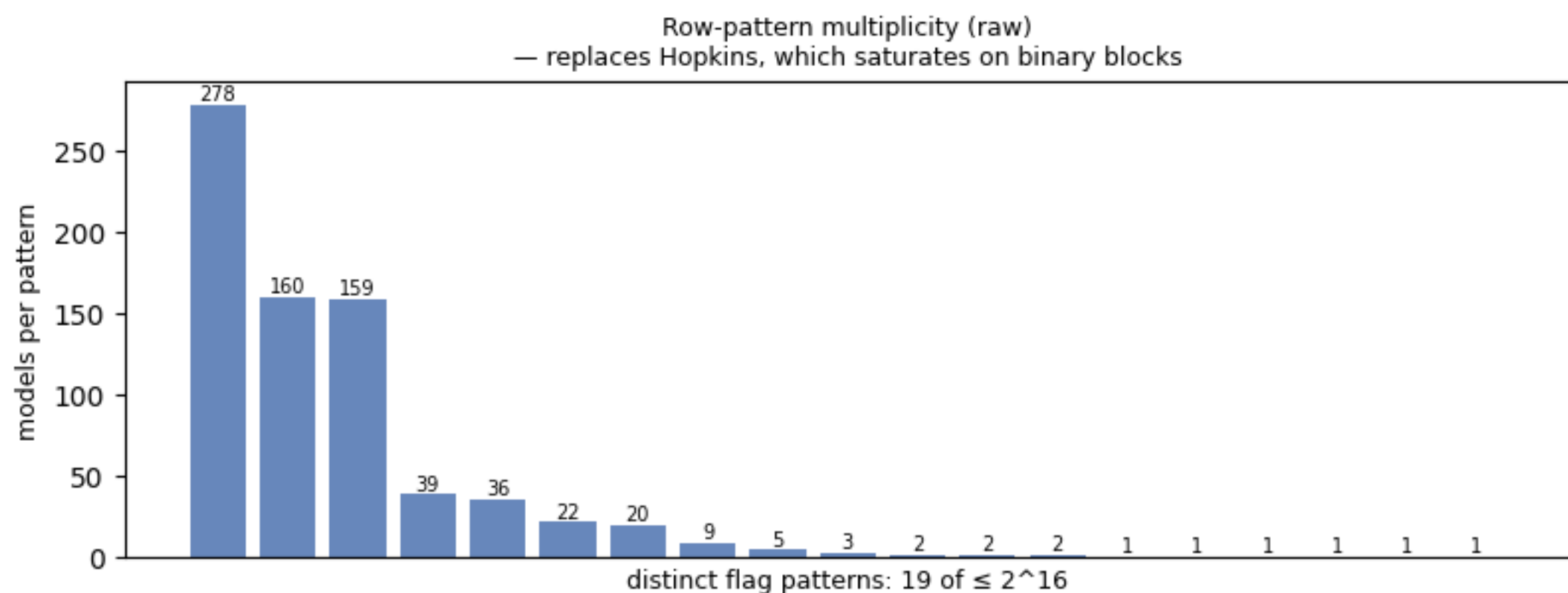
*Should be more high patterns, then clusterable*

Preprocessing usefulness · constr\_cols · 743 models



External constraints have good information, but some patterns are indistinguishable.  
15 categories

Preprocessing usefulness · prodana\_cols · 743 models



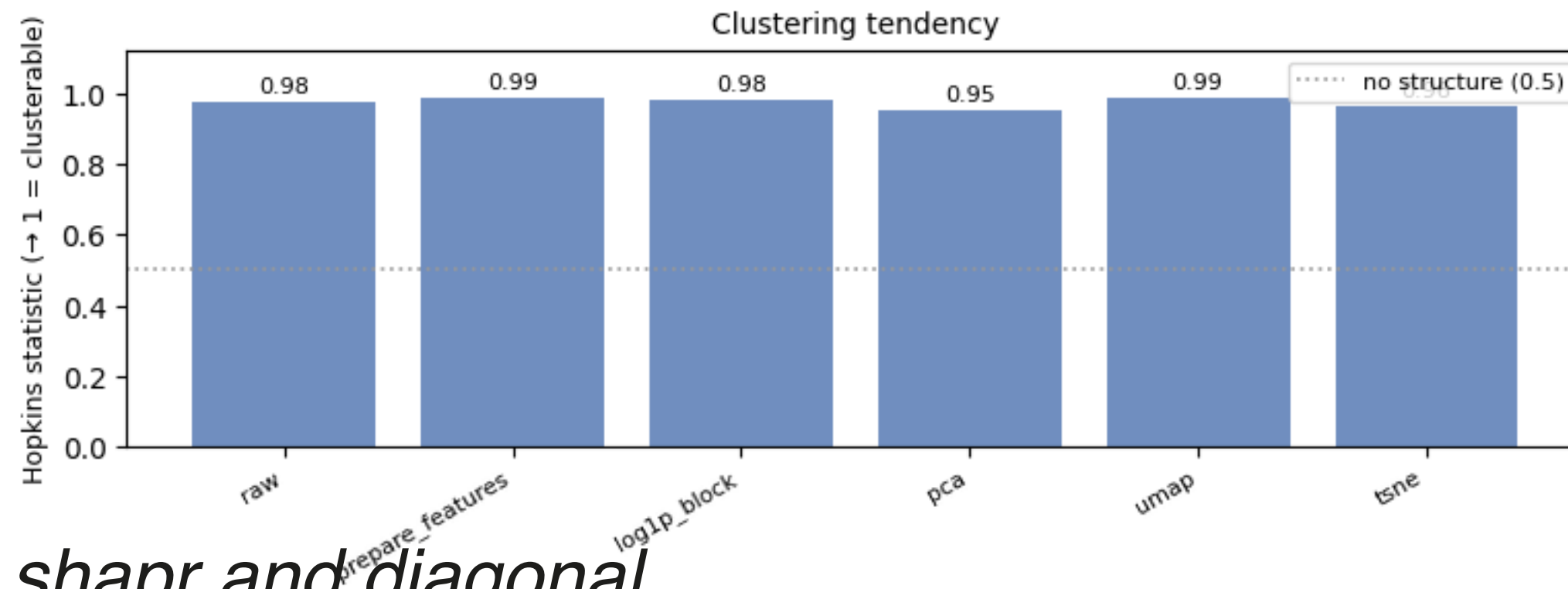
Analyses have good information, but some patterns are indistinguishable.  
19 categories

# PREPROCESSING: CONTINUOUS

PREPROCESSING

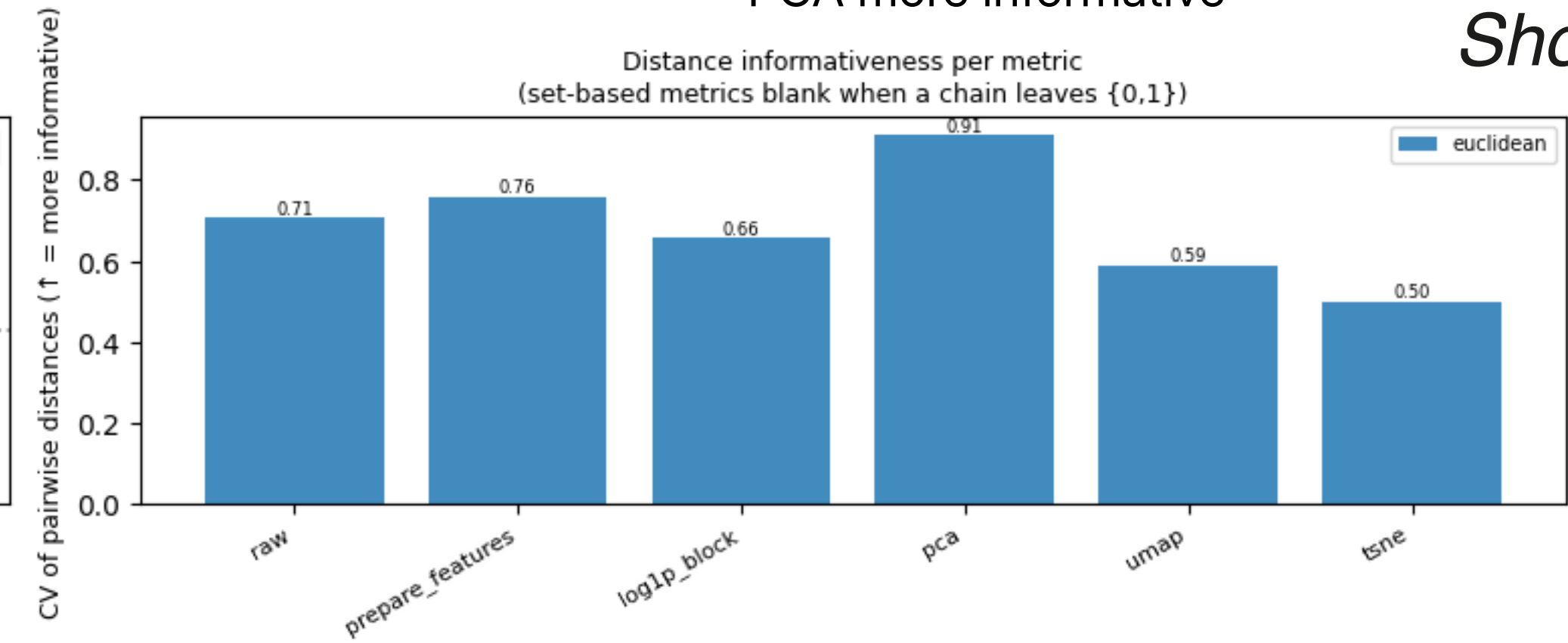
*Should be close to 1*

Clusterable for any transformation



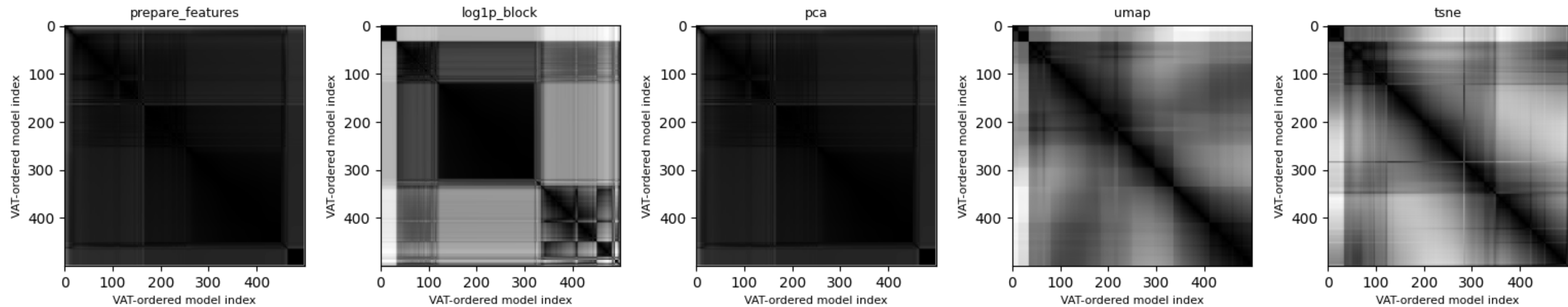
PCA more informative

*Should be high*



*Should be shapr and diagonal*

VAT ordered dissimilarity · bf\_cols (euclidean) — dark diagonal blocks = clusters · 743 models



No large difference

==> BF: Use no transformation.

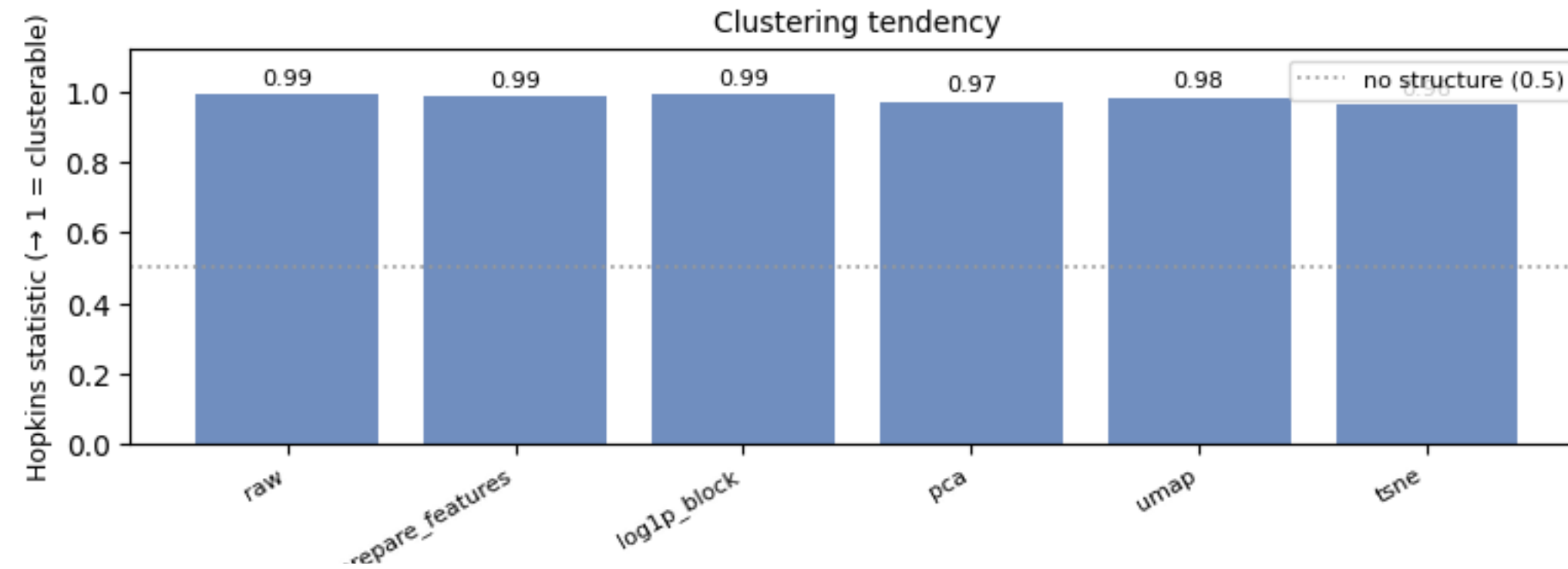


# PREPROCESSING: CONTINUOUS

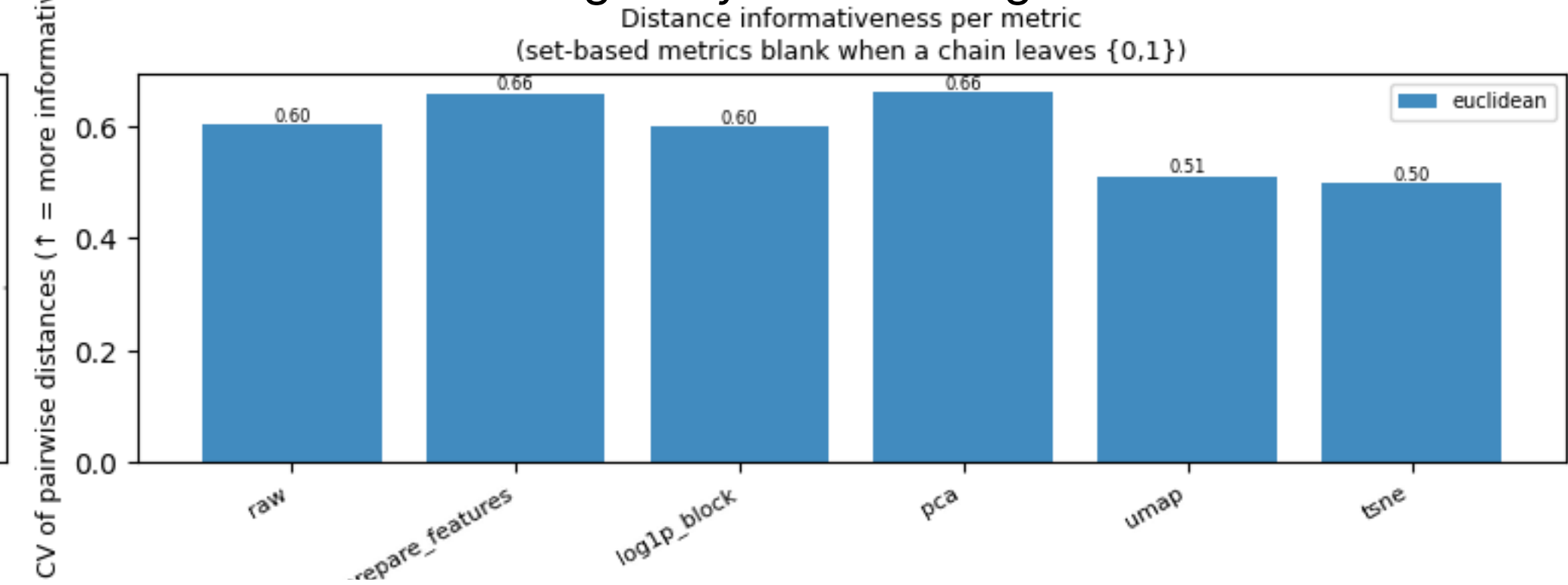
PREPROCESSING

Preprocessing usefulness · mix\_cols · 743 models

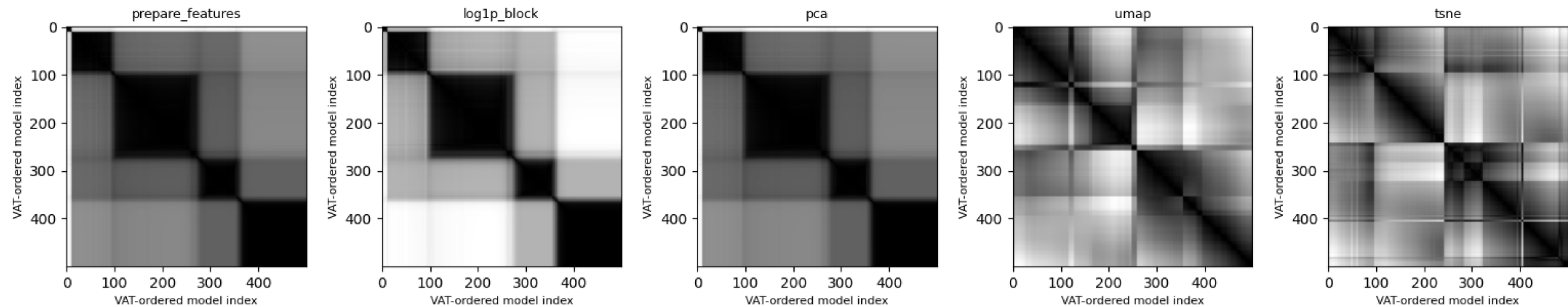
Clusterable for any transformation



No gain by transforming



VAT ordered dissimilarity · mix\_cols (euclidean) — dark diagonal blocks = clusters · 743 models



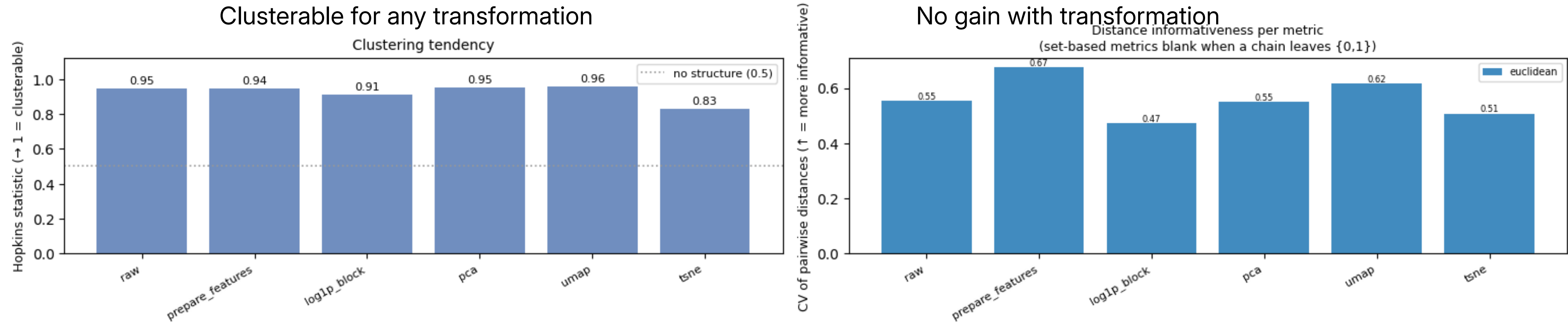
No large gain by transforming

==> Mixing: Use no transformation.

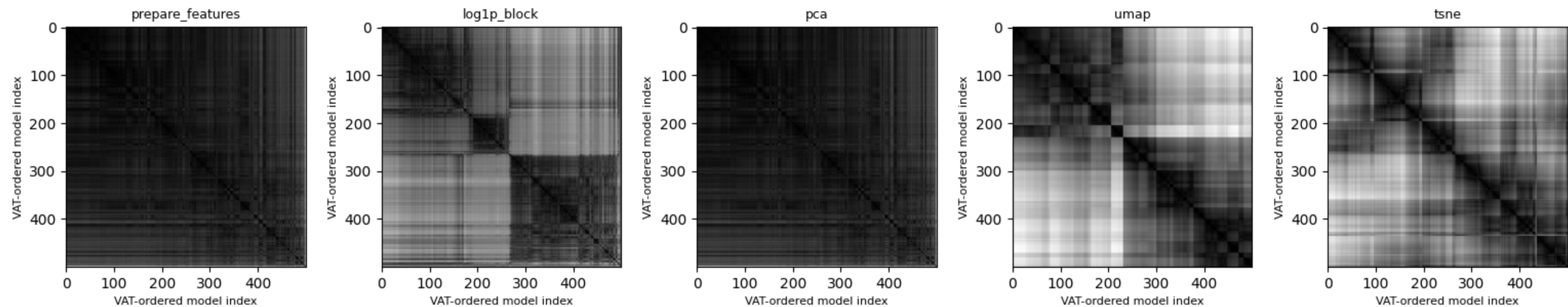
# PREPROCESSING: CONTINUOUS SPARSE

PREPROCESSING

Preprocessing usefulness · mass\_cols · 743 models



VAT ordered dissimilarity · mass\_cols (euclidean) — dark diagonal blocks = clusters · 743 models



Diagonal distinguishable, no gain with transformation

==> Mass: Use no transformation.



# PREPROCESSING: CONTINUOUS SPARSE

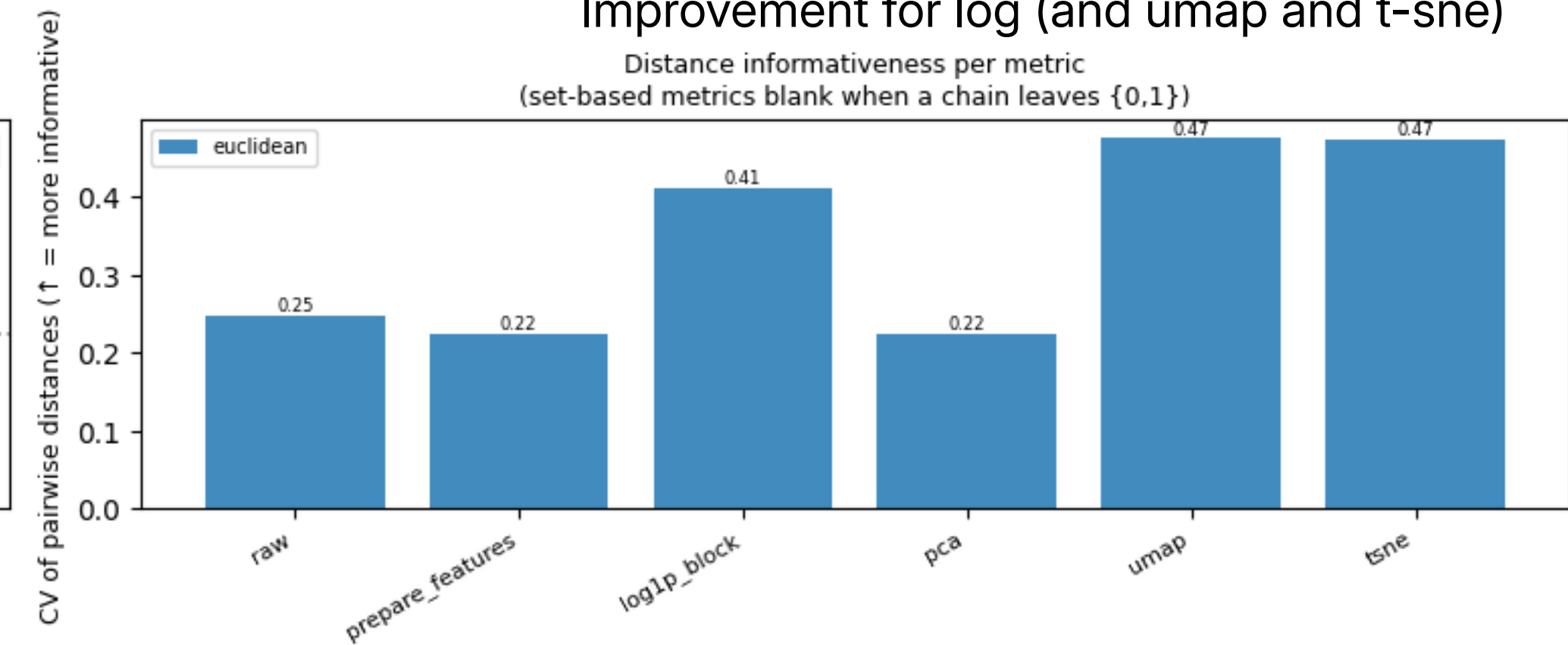
PREPROCESSING

Preprocessing usefulness · input\_param\_cols · 743 models

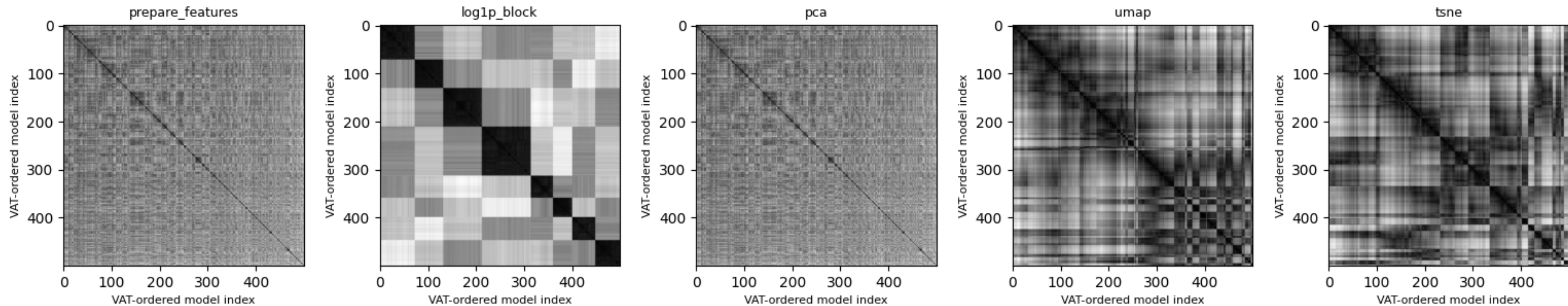
Clusterable for any transformation, improvement for log



Improvement for log (and umap and t-sne)



VAT ordered dissimilarity · input\_param\_cols (euclidean) — dark diagonal blocks = clusters · 743 models



Good clustering pattern in the diagonal

Good clustering pattern in the diagonal, higher number of clusters

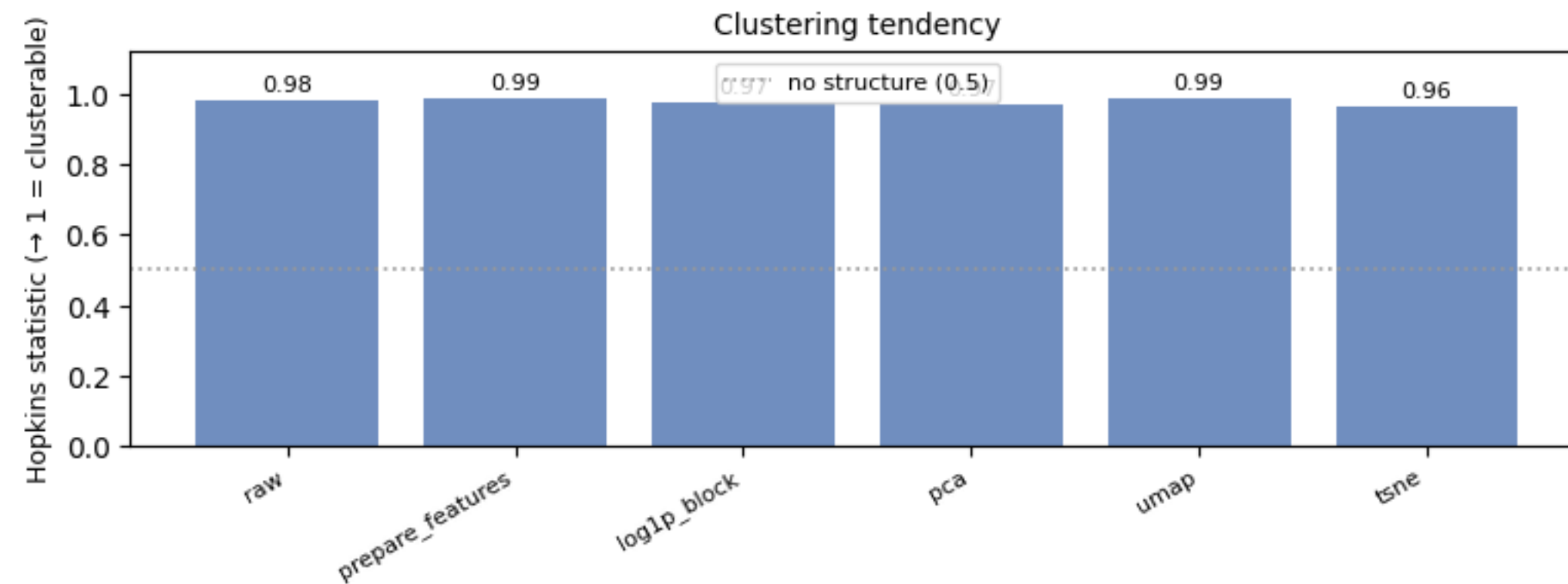
==> Input parameters: Use Log transformation.

# PREPROCESSING: CONTINUOUS SPARSE

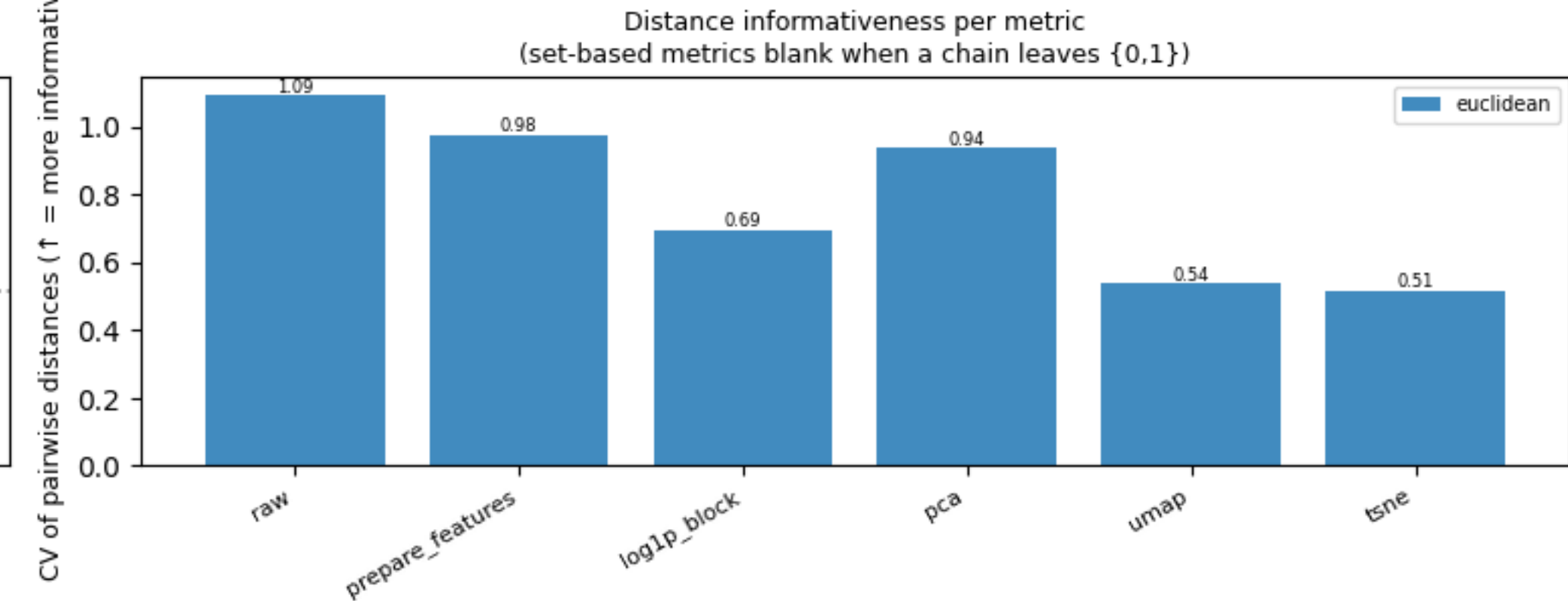
PREPROCESSING

Preprocessing usefulness · xsec\_cols · 743 models

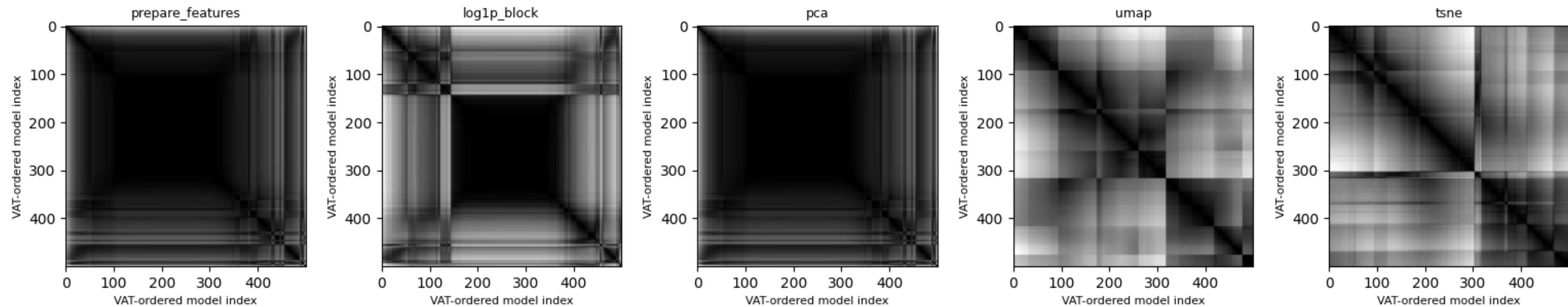
Clusterable for any transformation



No gain with transformation



VAT ordered dissimilarity · xsec\_cols (euclidean) — dark diagonal blocks = clusters · 743 models



No gain with transformation

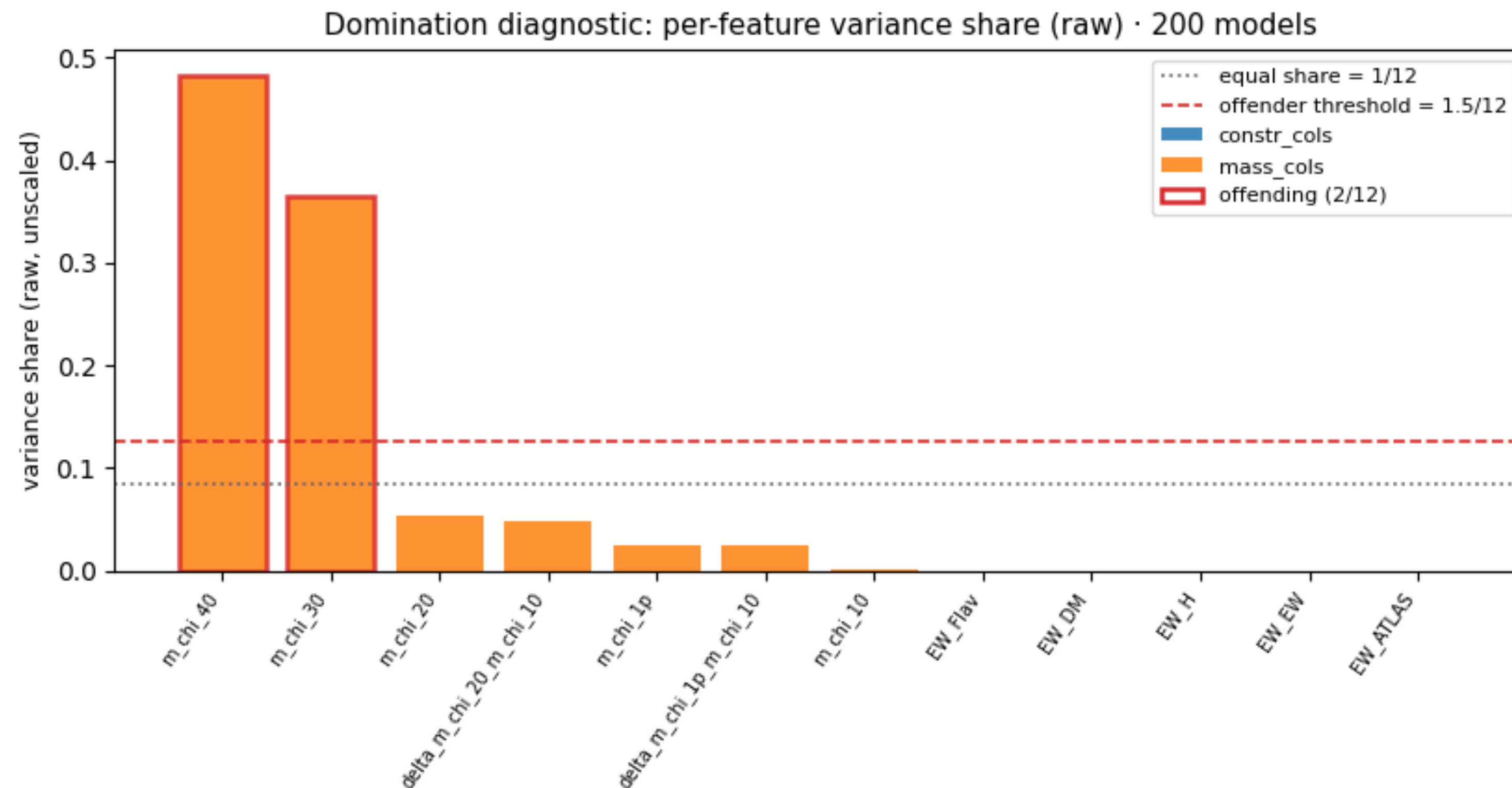
==> Xsec: Use no transformation.



# DOMINANCE

DOMINANCE

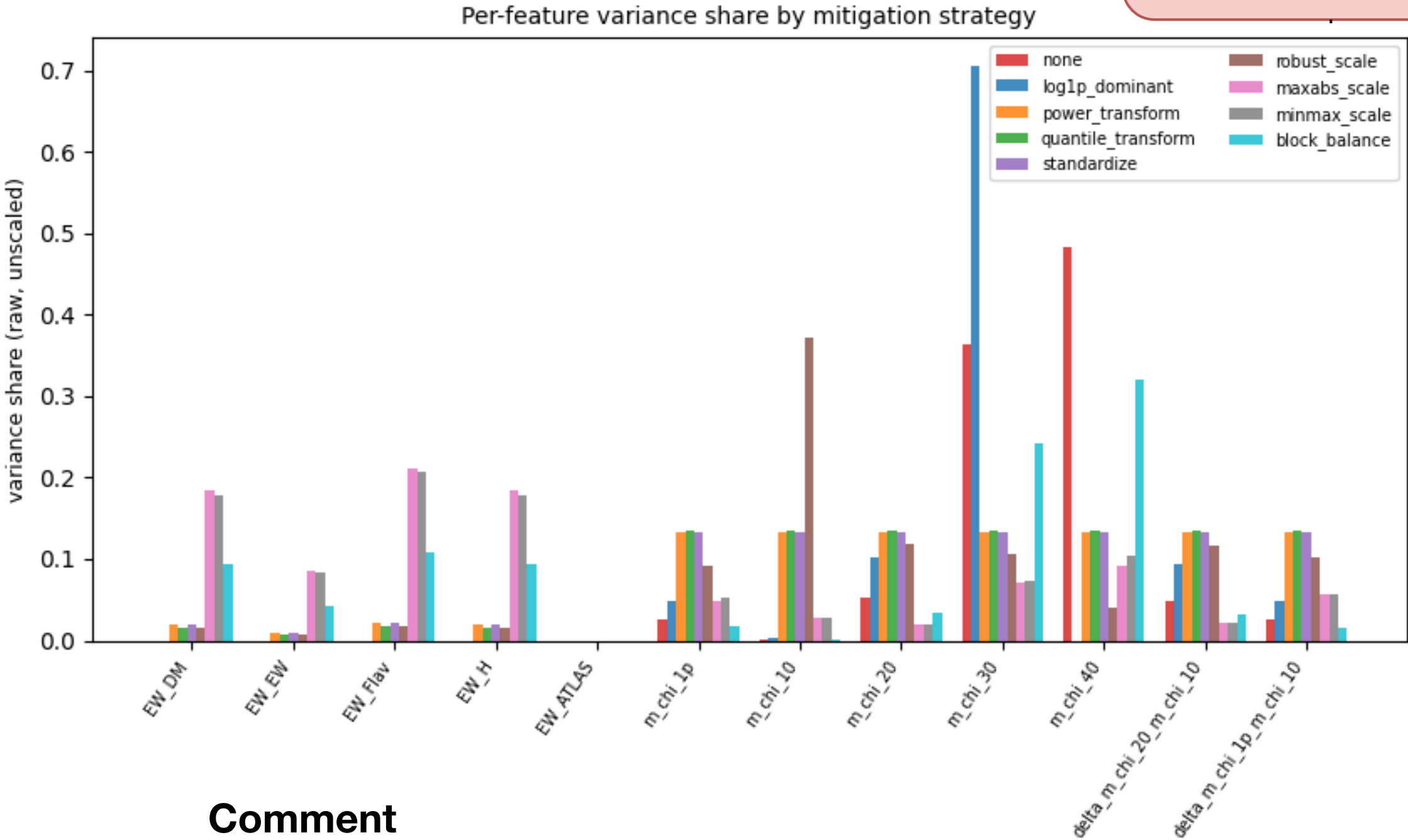
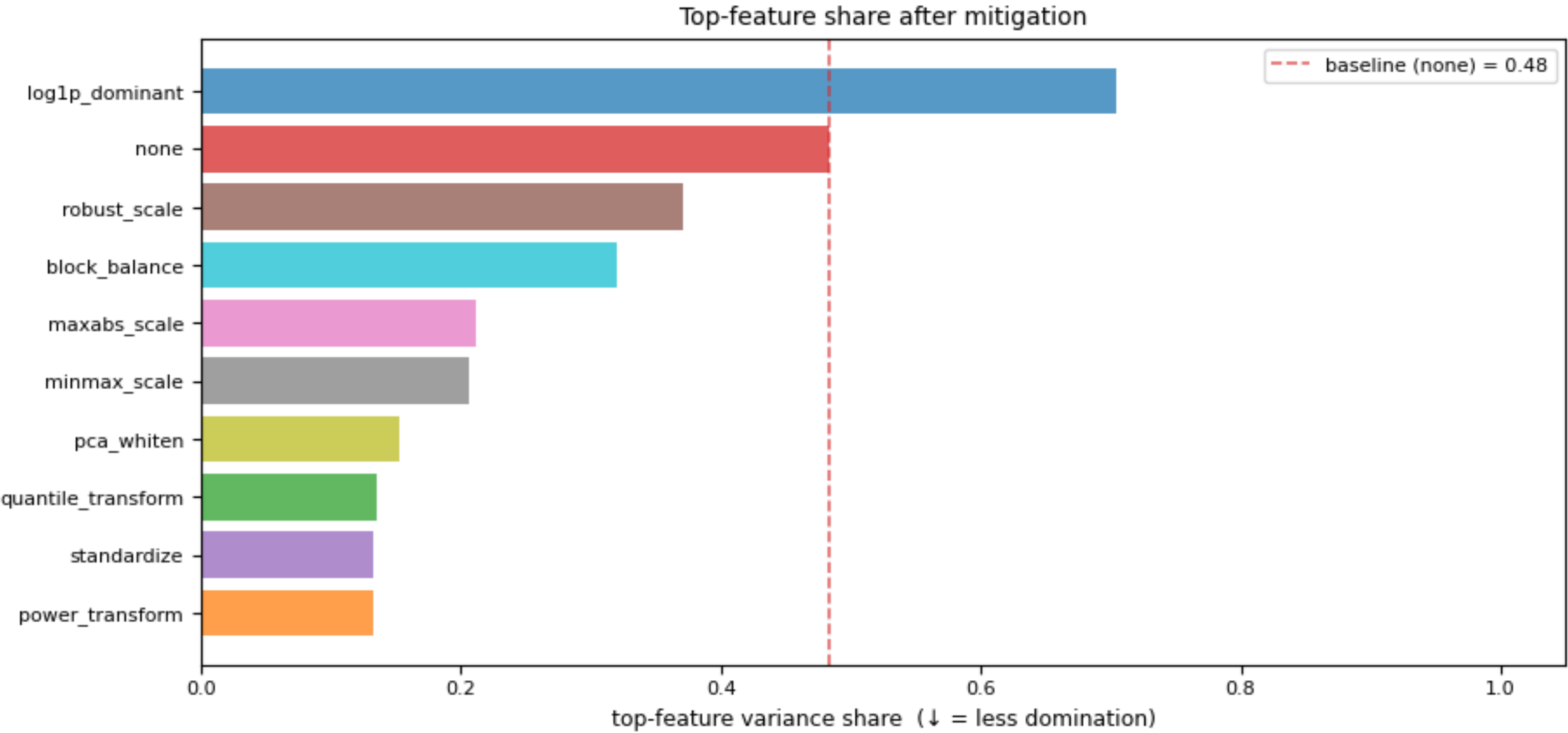
- Need for domination mitigation, e.g. no contribution from additional information if one block of input dominant (mass).
- Apply input feature normalisation.
- **Diagnosis:**
  - **Per-feature:** Each feature's fractional contribution to total variance and classifies features as dominant when their variance share exceeds  $\text{factor}/n\_features$ , threshold factor = 1.5
  - **Block-level:** Variance by feature block  $\text{factor} \times n\_block / n\_features$  (per-feature applied after block mitigation)



# DOMINANCE

Domination mitigation on offending block(s) only · 200 models

DOMINANCE



| Mitigation         | Comment                                                                                                          |
|--------------------|------------------------------------------------------------------------------------------------------------------|
| Log                | Compresses large values with log transformation..                                                                |
| Robust scale       | Rescales using the median and interquartile range.                                                               |
| Block balance      | Rescales each dominant feature block to unit total variance while preserving relative variance within the block. |
| Max-absolute scale | Normalizes each feature by its maximum absolute value.                                                           |
| Min-max scale      | Maps each feature to a fixed bounded range.                                                                      |
| PCA Whiten         | Logarithmically compresses only the most dominant feature(s).                                                    |
| Quantile transform | Maps features to a common target distribution.                                                                   |
| Standardize        | Rescales features to zero mean and unit variance.                                                                |
| Power transform    | Gaussianizes feature distributions and stabilizes variance.                                                      |

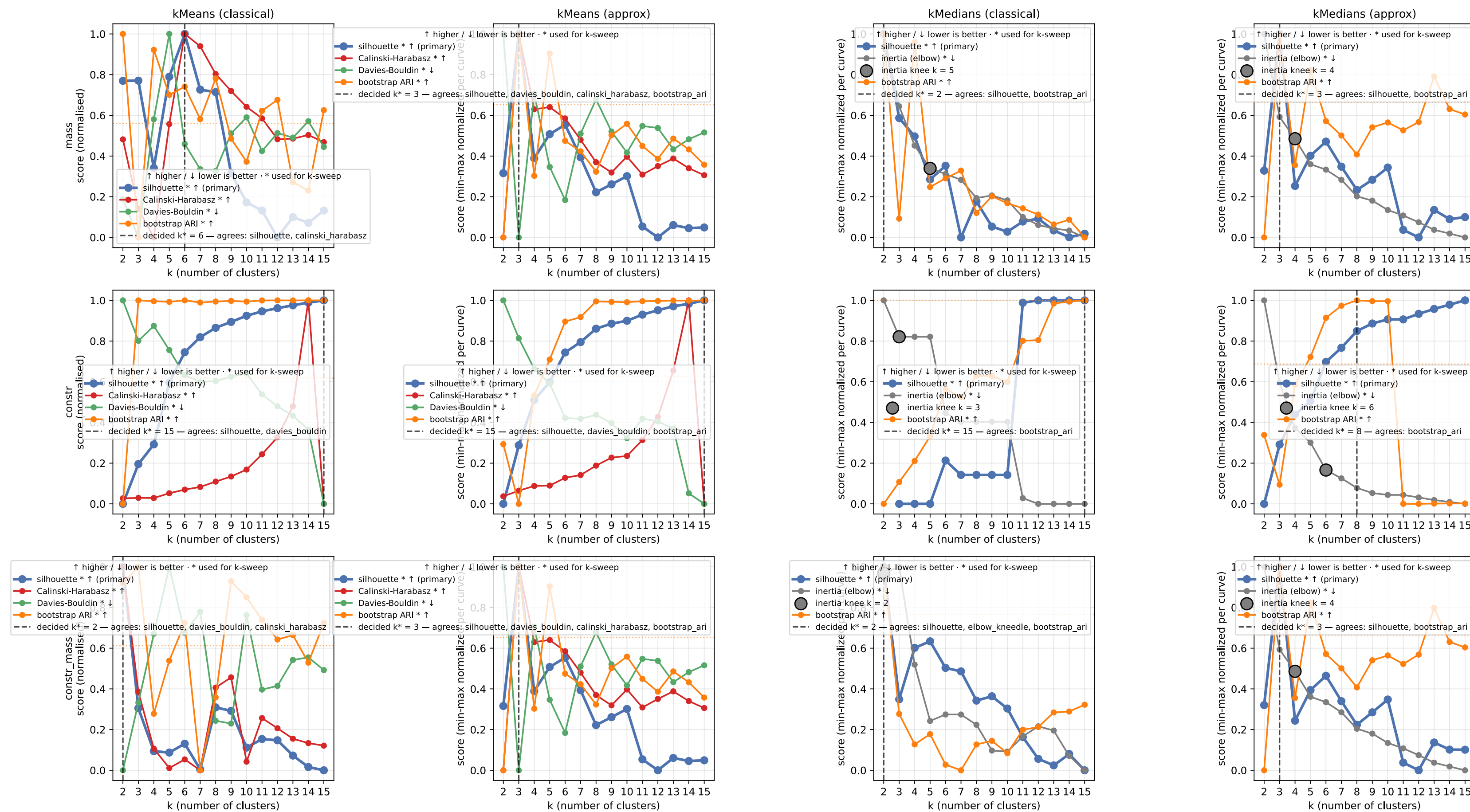
==> Mass dominance mitigation is overcome with Max-abs or Min-max scale mitigation.

==> The transformation is applied for any input block, Max-abs is a good choice.

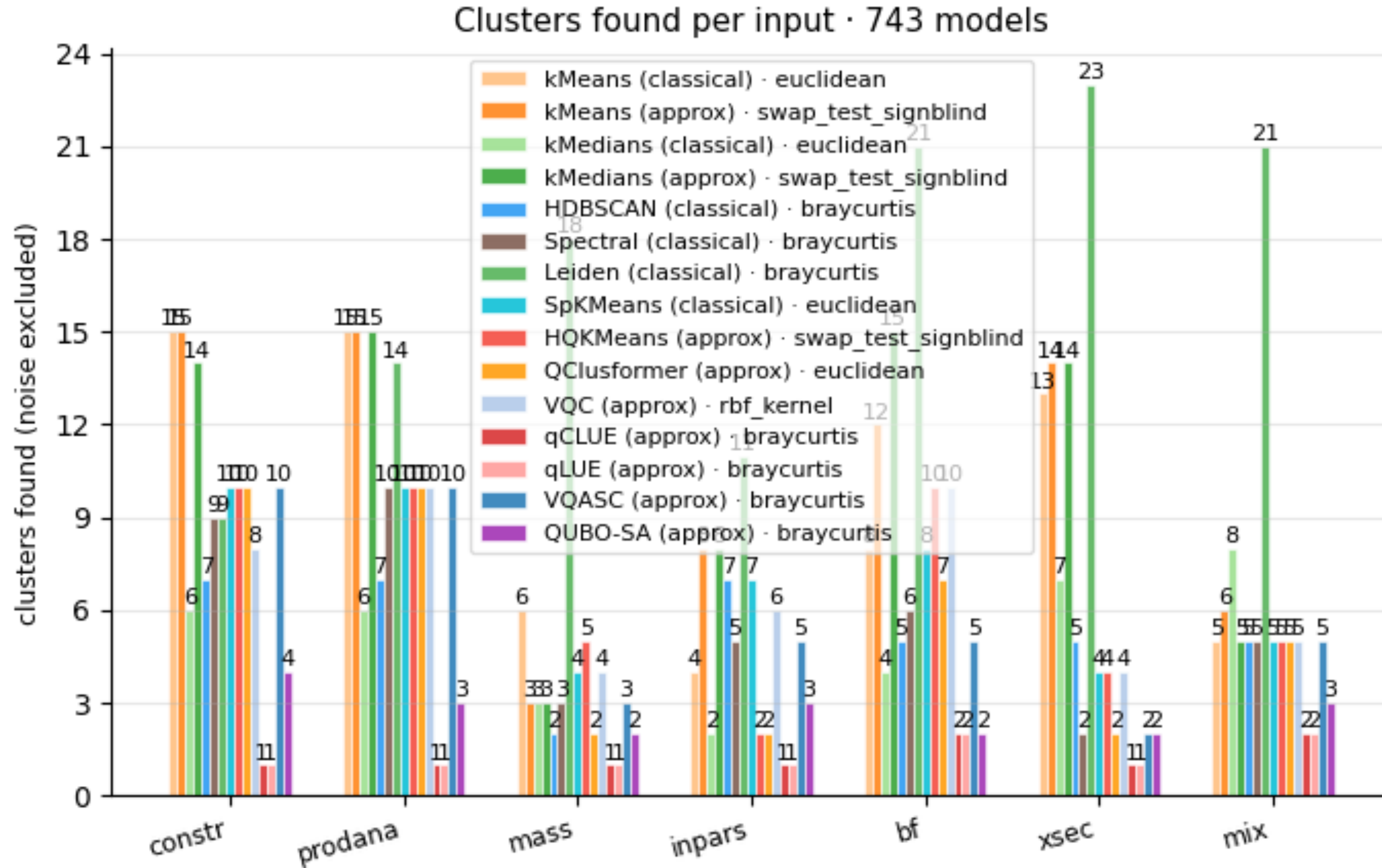


## k-sweep diagnostics — combined

K-sweep diagnostics — algorithm/variant × input · 743 models



# N CLUSTERS

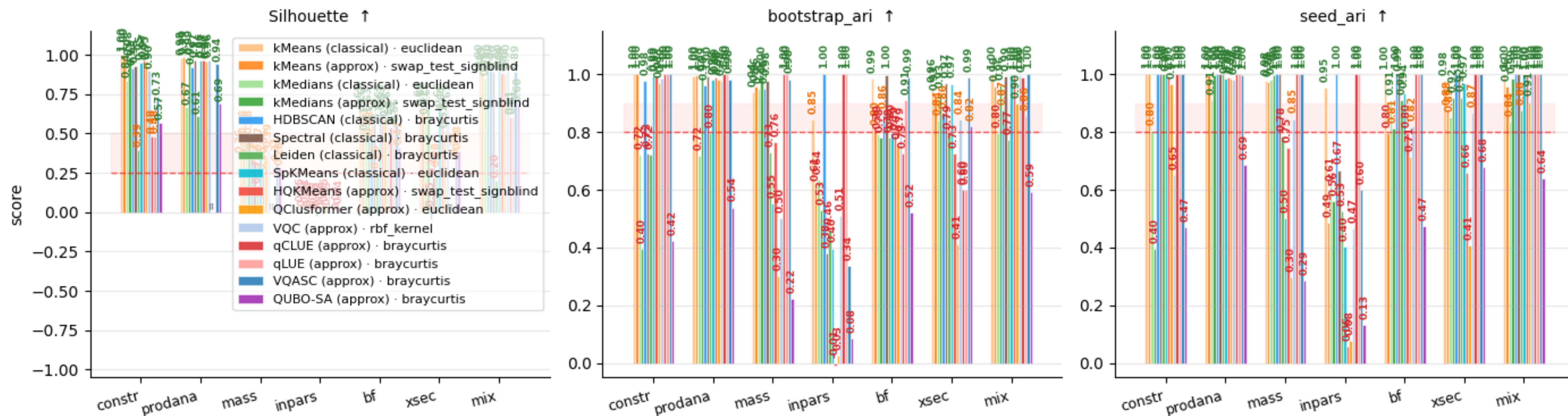




# CLUSTERING ALGORITHM PERFORMANCE

k-SWEEP  
(some algorithms)

- Check on clustering algorithm in given setup stability for seed and distance.



**Silhouette score** — measures how well each sample fits within its assigned cluster relative to neighboring clusters, with higher values indicating better separation and cohesion.

**Bootstrap ARI** — measures clustering stability across resampled datasets using the Adjusted Rand Index, with higher values indicating robust cluster assignments.

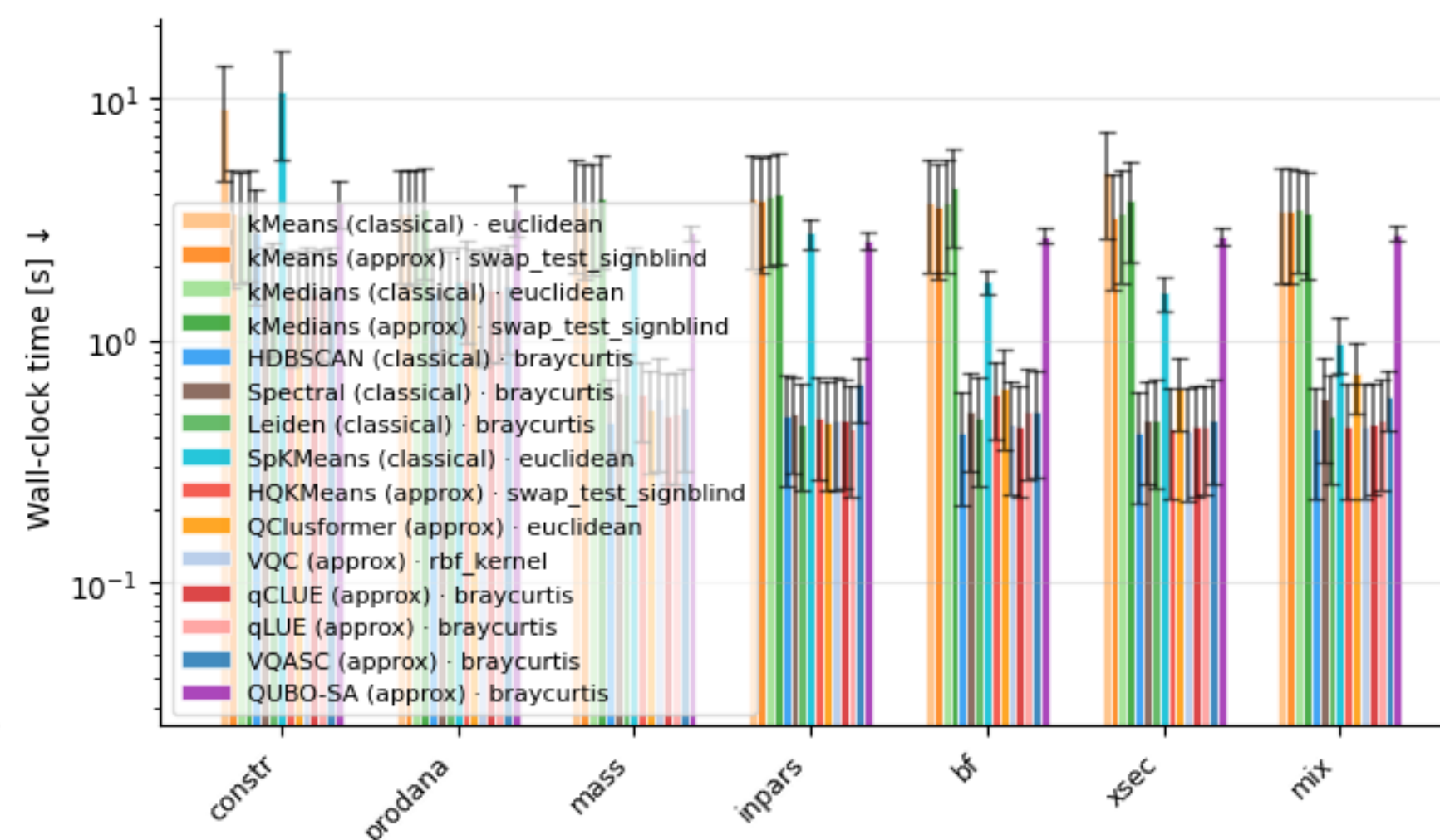
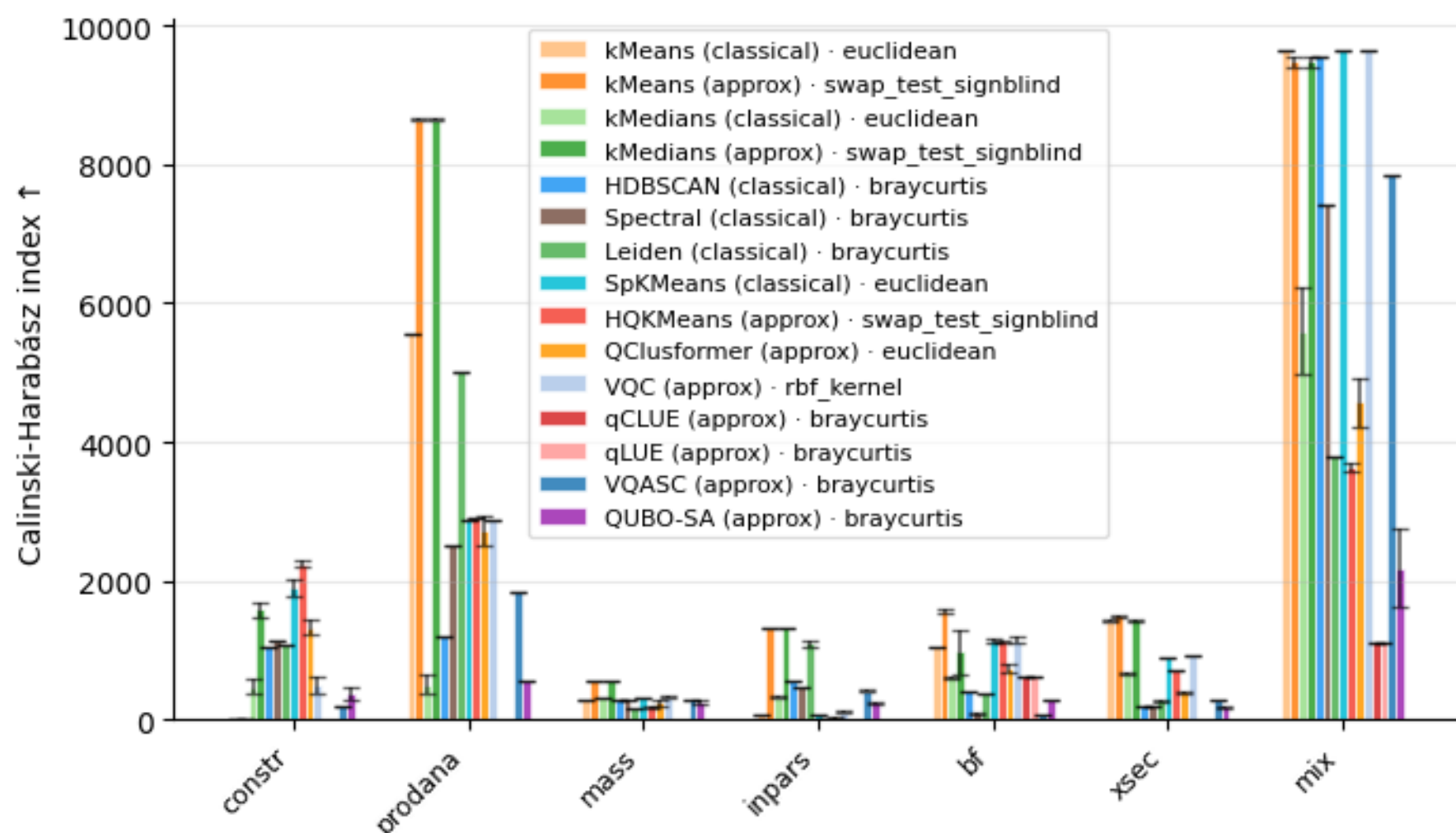
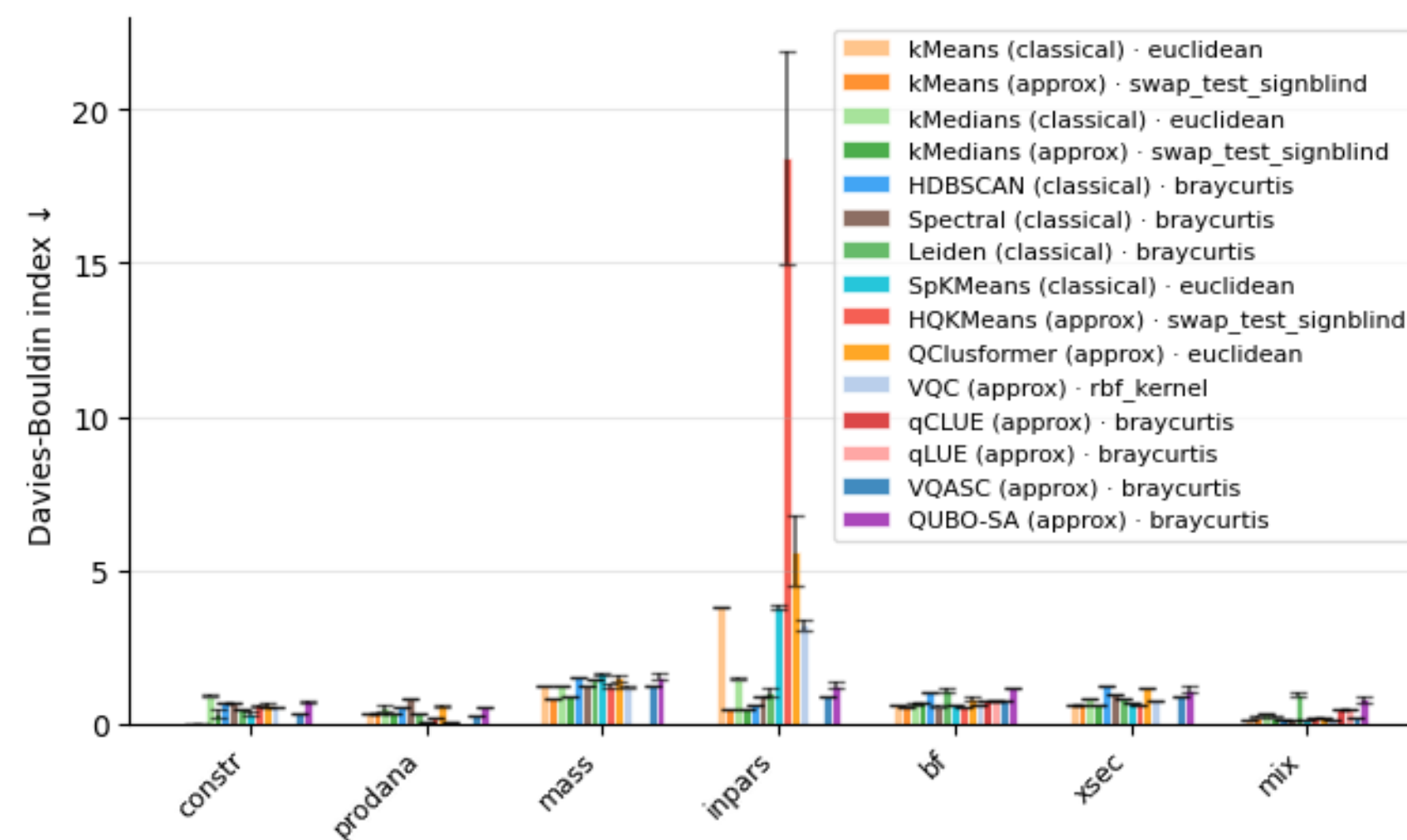
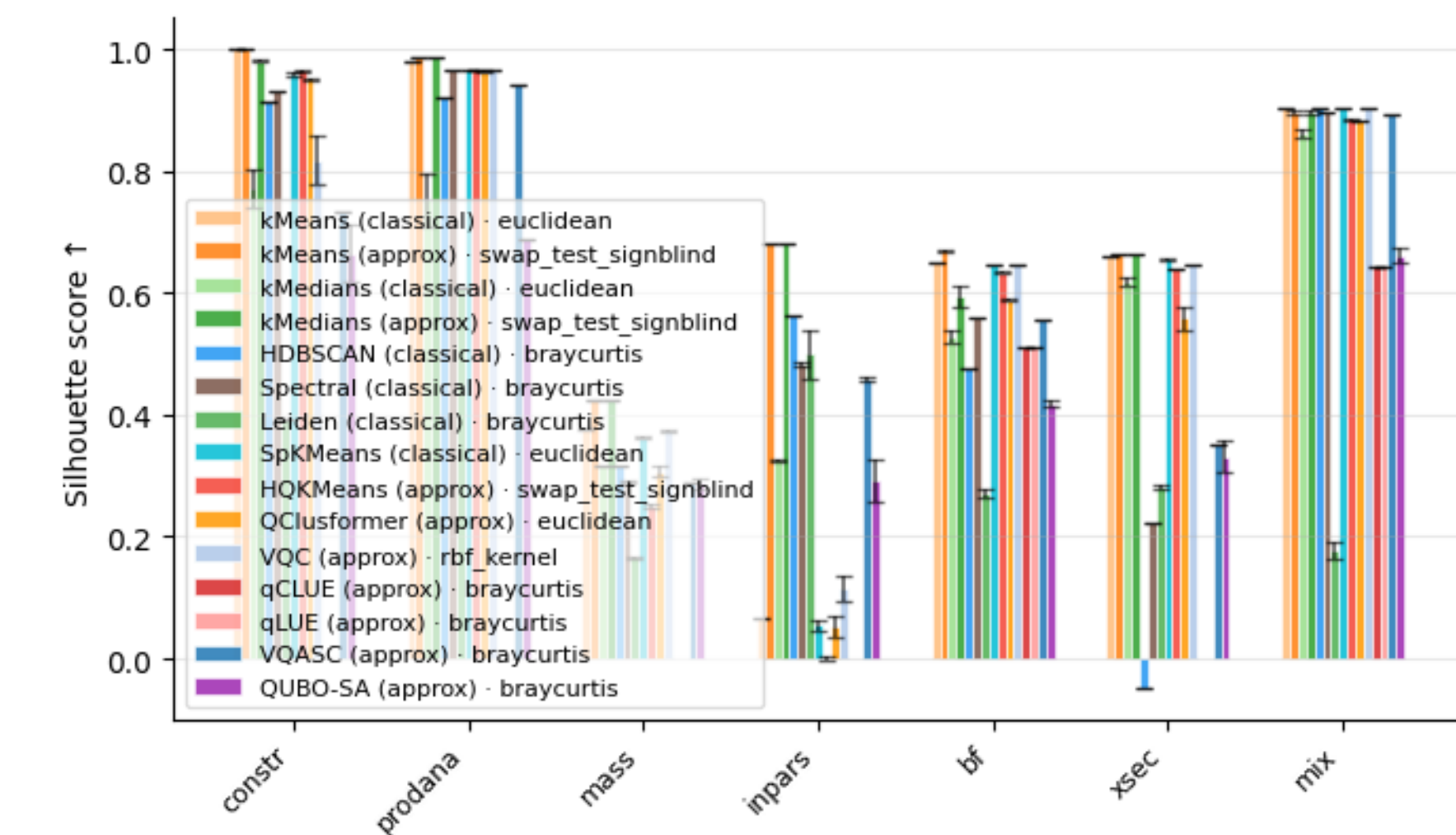
**Seed ARI** — measures clustering stability across different random initializations using the Adjusted Rand Index, with higher values indicating low sensitivity to the random seed.



# CLUSTERING PERFORMANCE

CLUSTERING  
Performance evaluation  
Benchmarking

- Check on clustering performance given this sample and determined number of clusters.



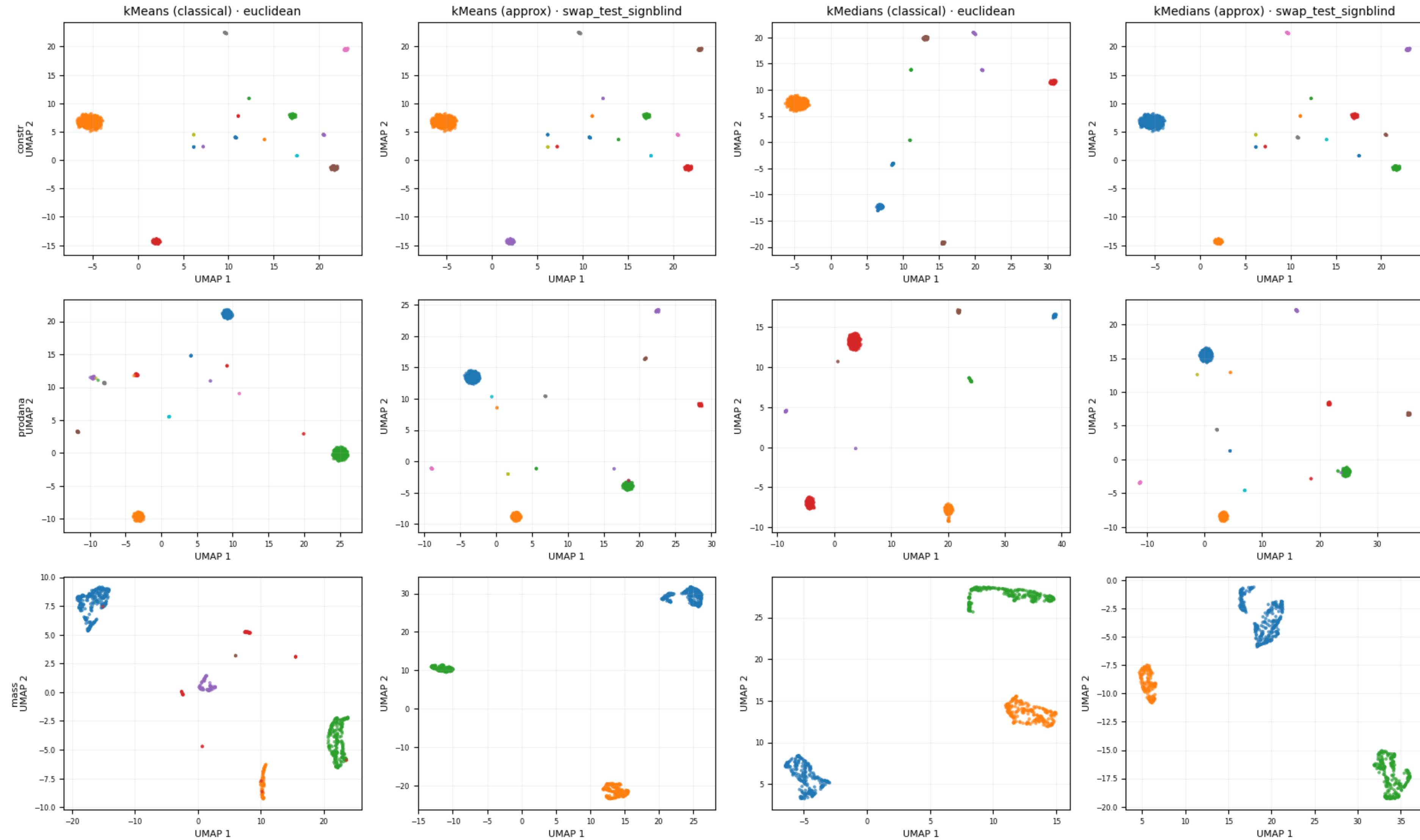
- **Silhouette score** — measures **within-cluster cohesion** relative to **separation** from neighboring clusters; **higher is better**.
- **Davies–Bouldin index** — measures **average similarity** between each **cluster** and its most similar competing cluster; **lower is better**.
- **Calinski–Harabasz score** — compares between-cluster **dispersion** with within-cluster dispersion; **higher is better**.



# CLUSTERING

**CLUSTERING**  
Performance evaluation  
Benchmarking

UMAP embedding grid — algorithm × input · 743 models

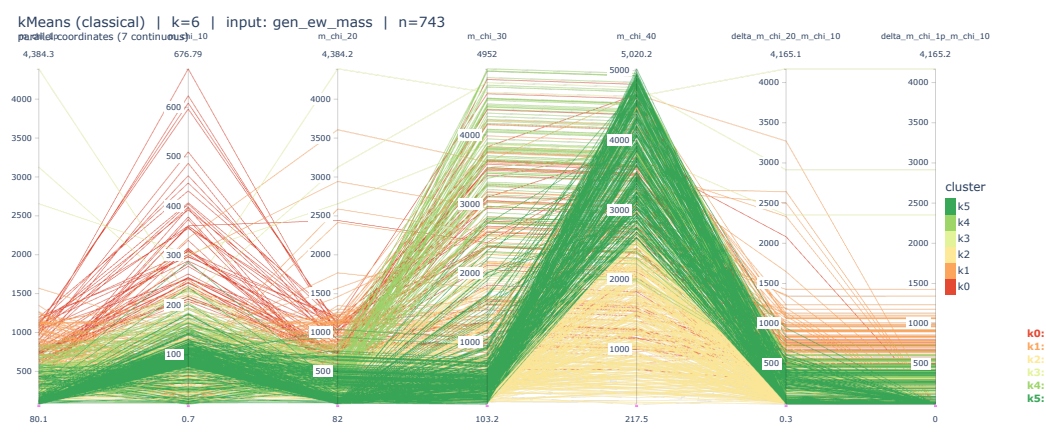




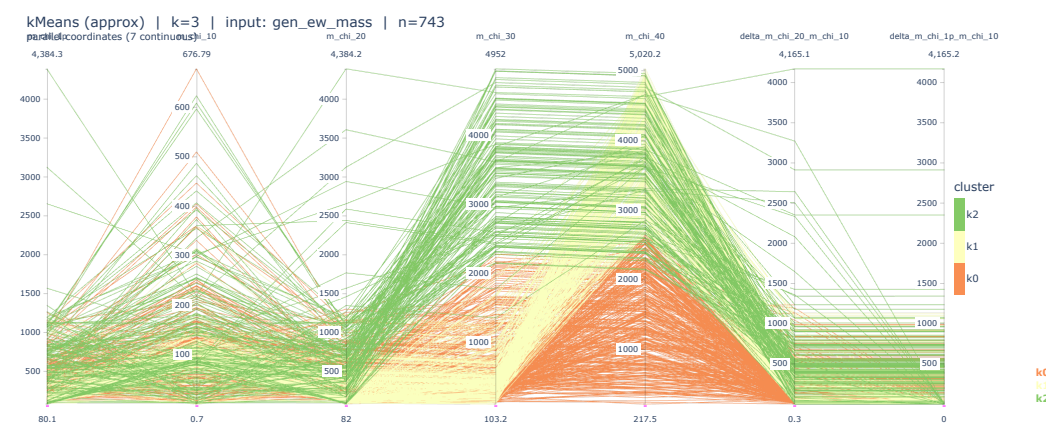
# CLUSTERING

CLUSTERING  
Performance evaluation  
Benchmarking

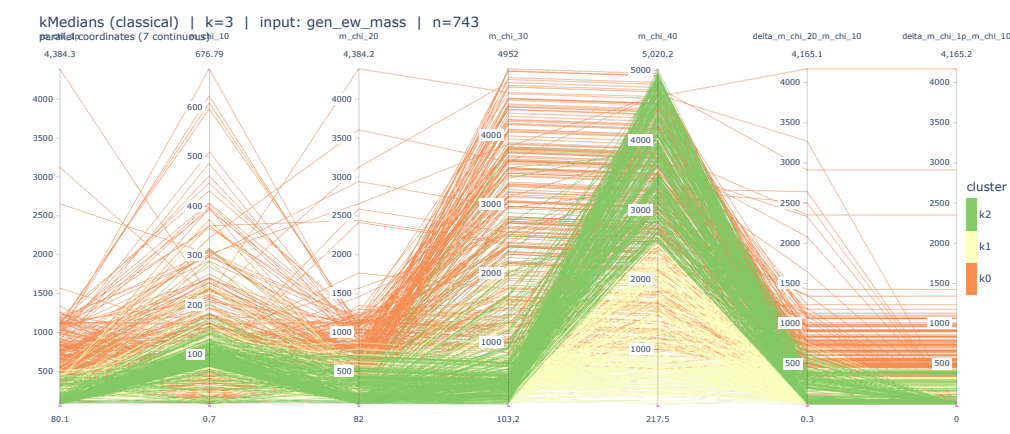
kMeans (classical)



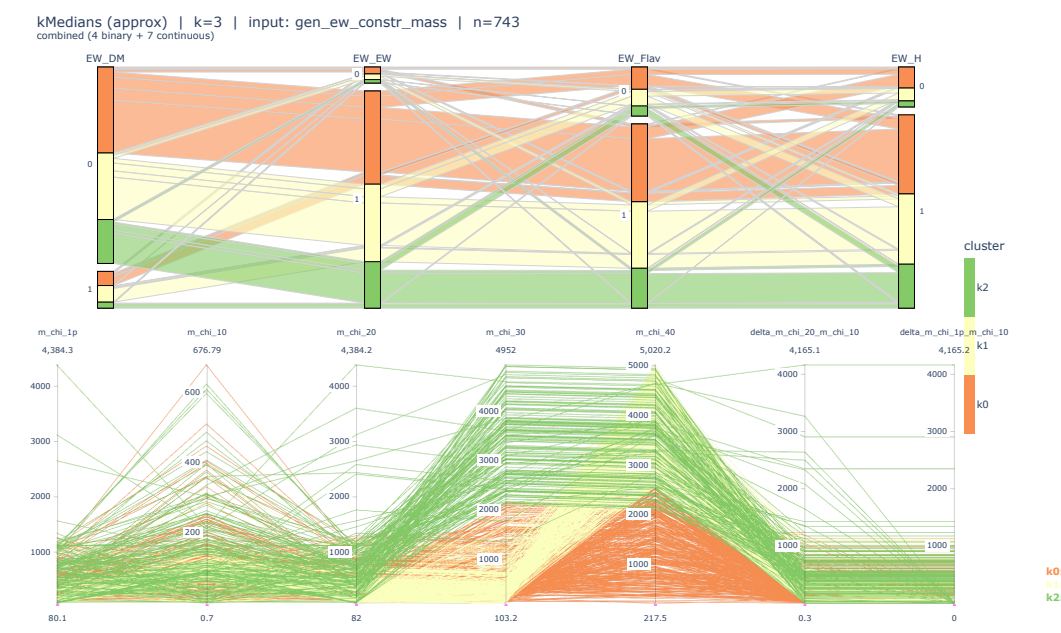
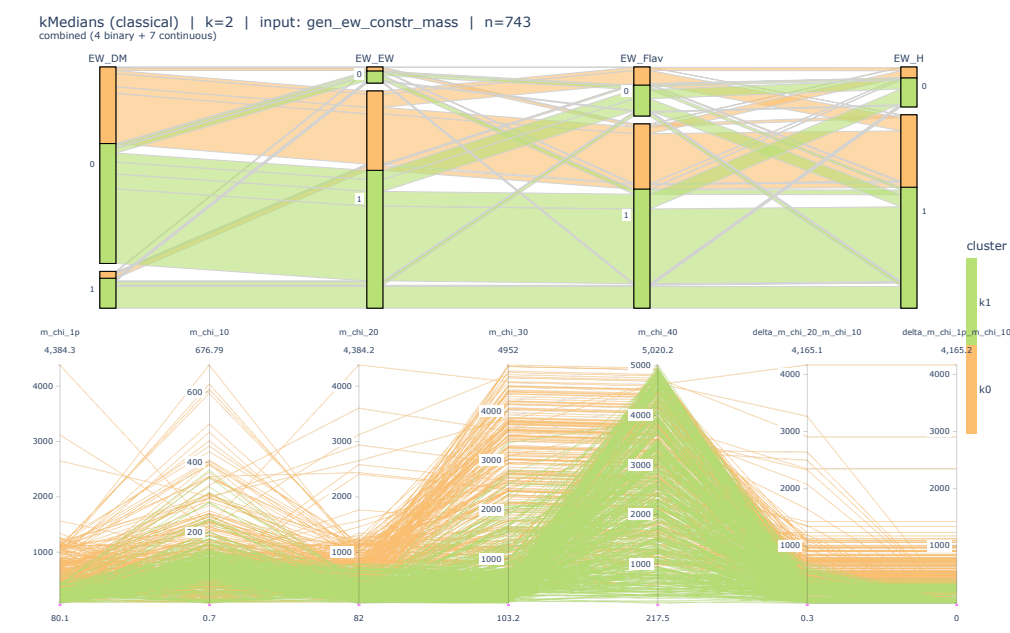
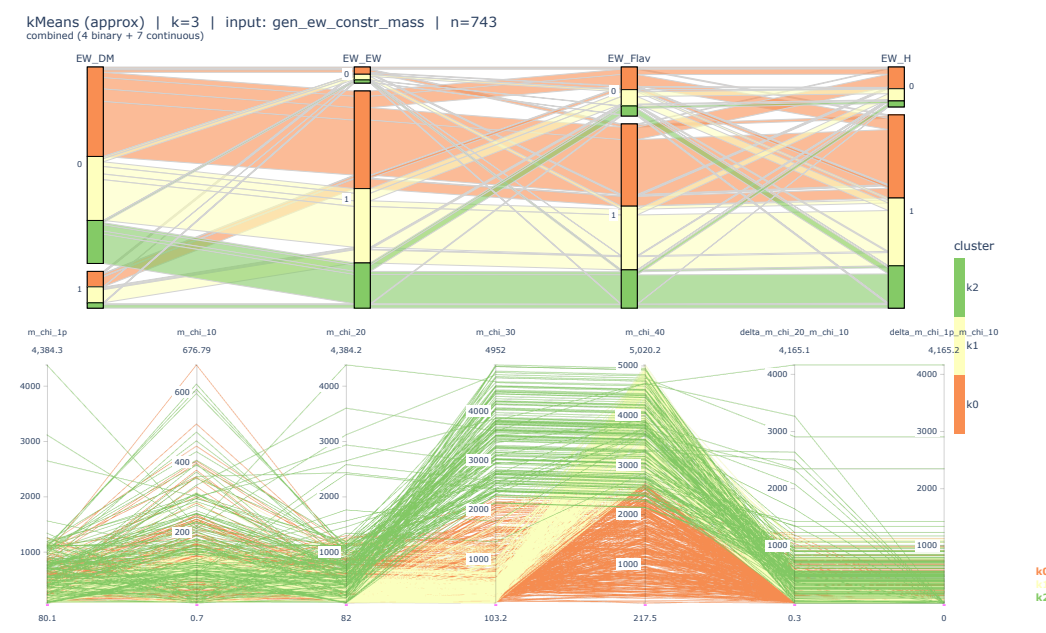
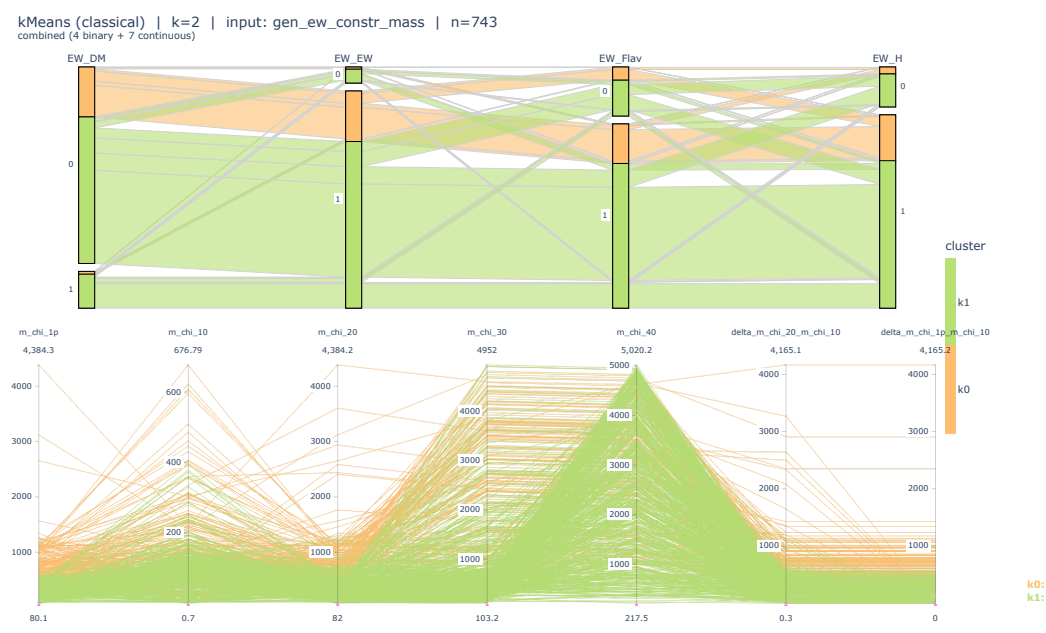
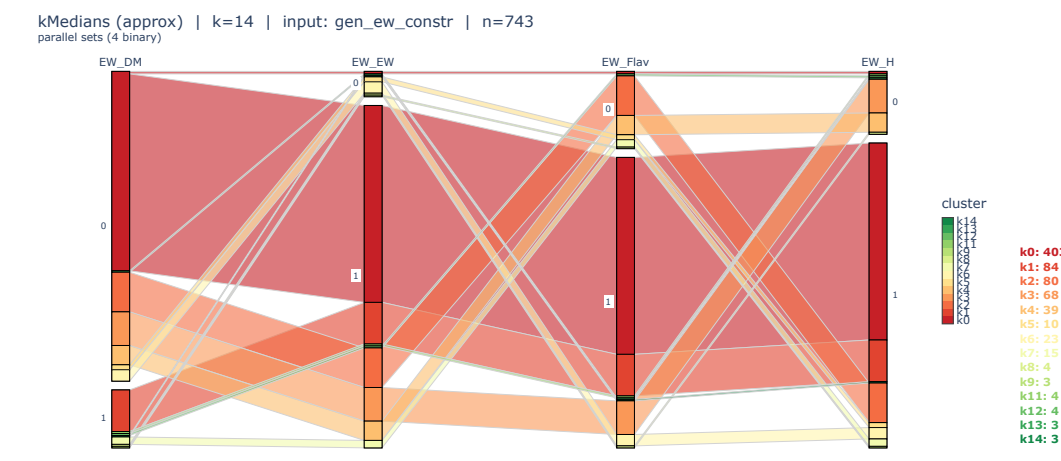
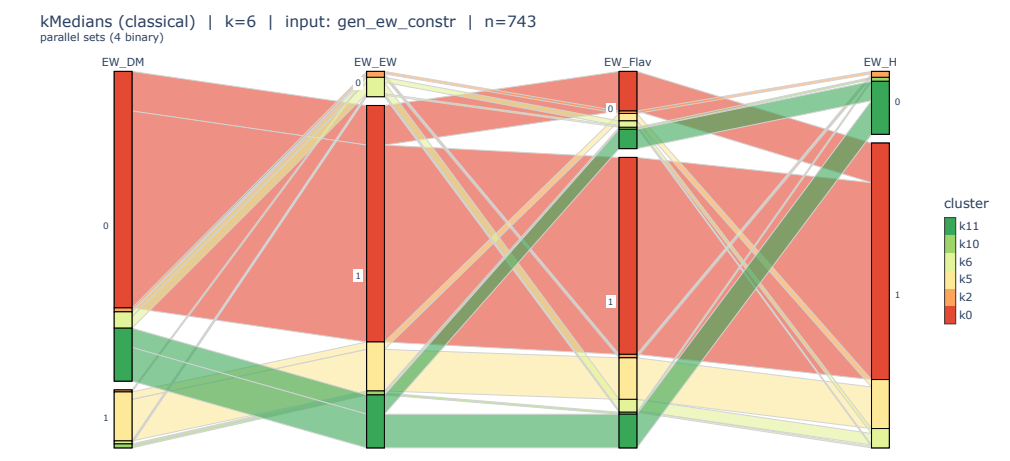
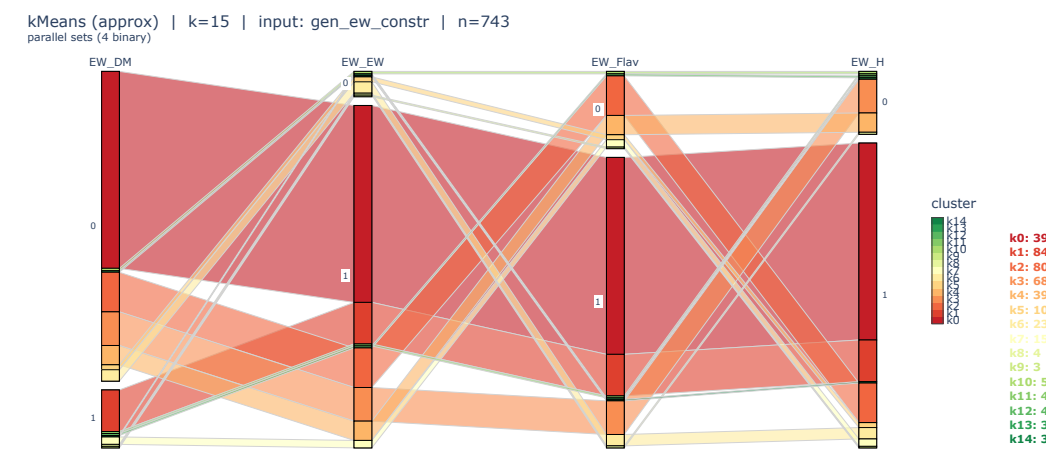
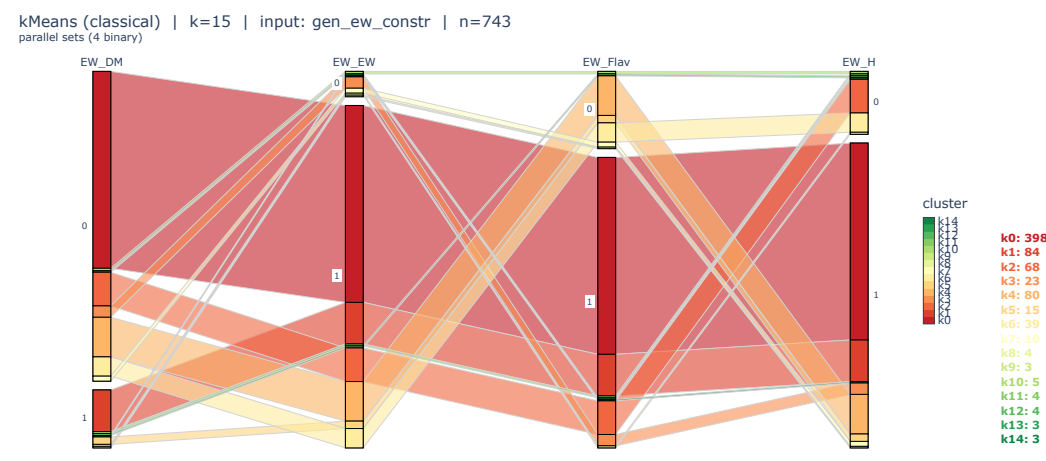
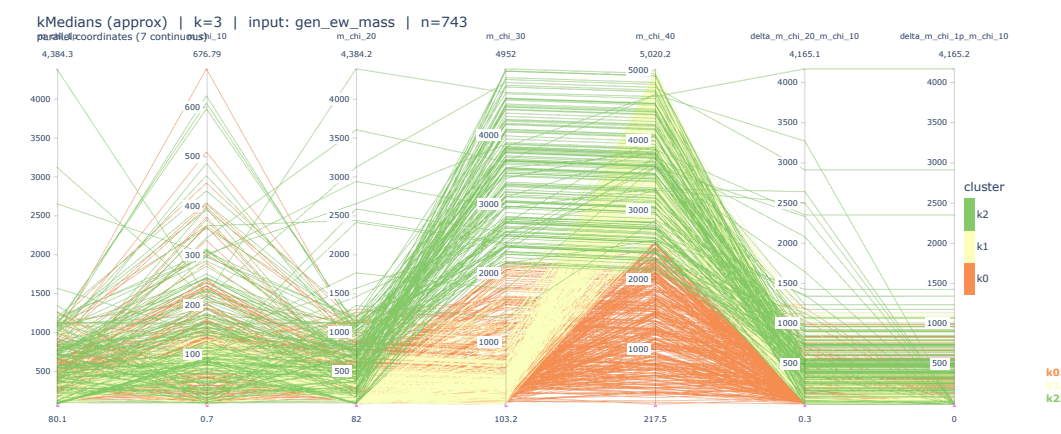
kMeans (approx)



kMedians (classical)



kMedians (approx)

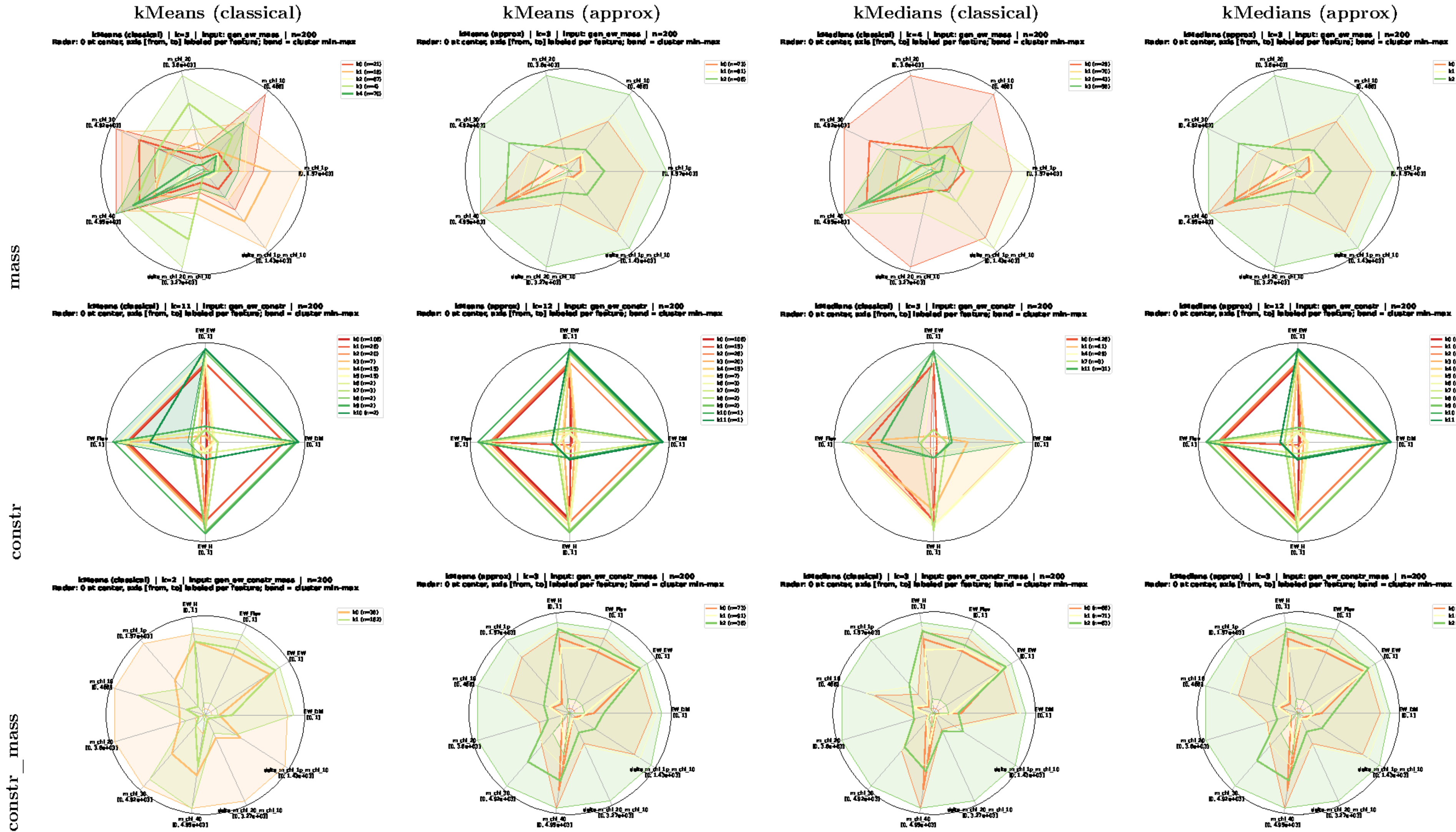




# CLUSTERING

CLUSTERING  
Performance evaluation  
Benchmarking

Radar grid — algorithm × input · 200 models

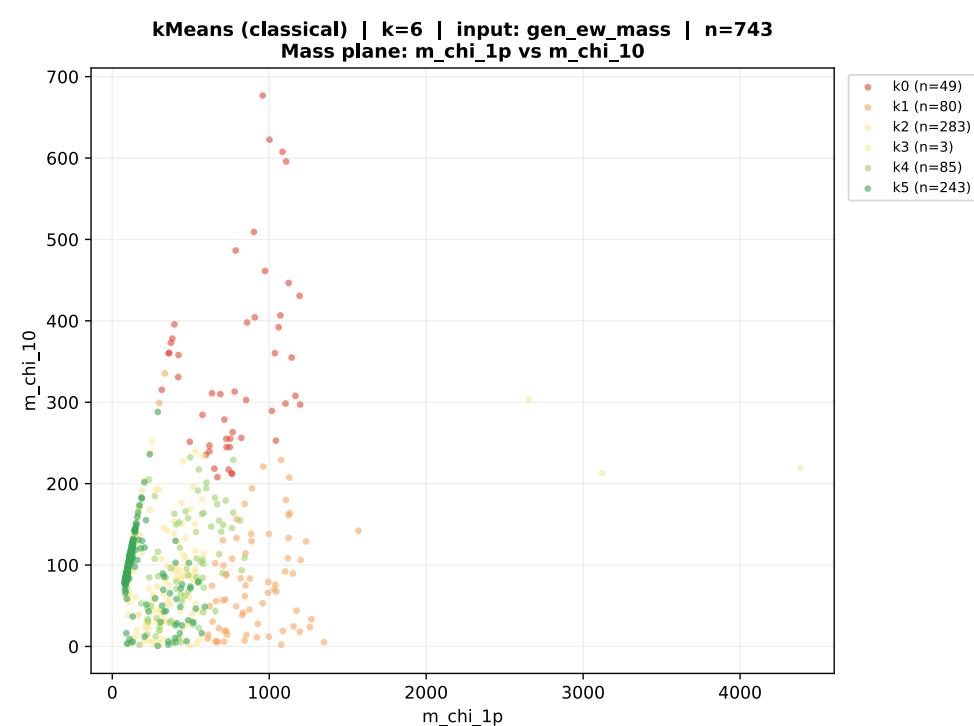


# CLUSTERING

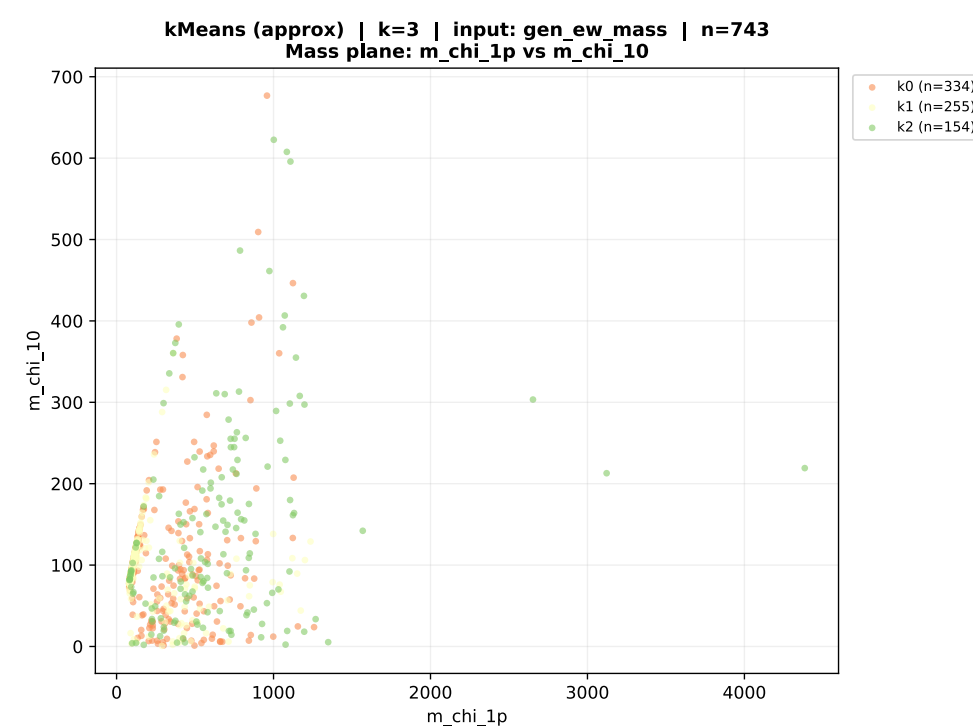
**CLUSTERING**  
Performance evaluation  
Benchmarking

mass

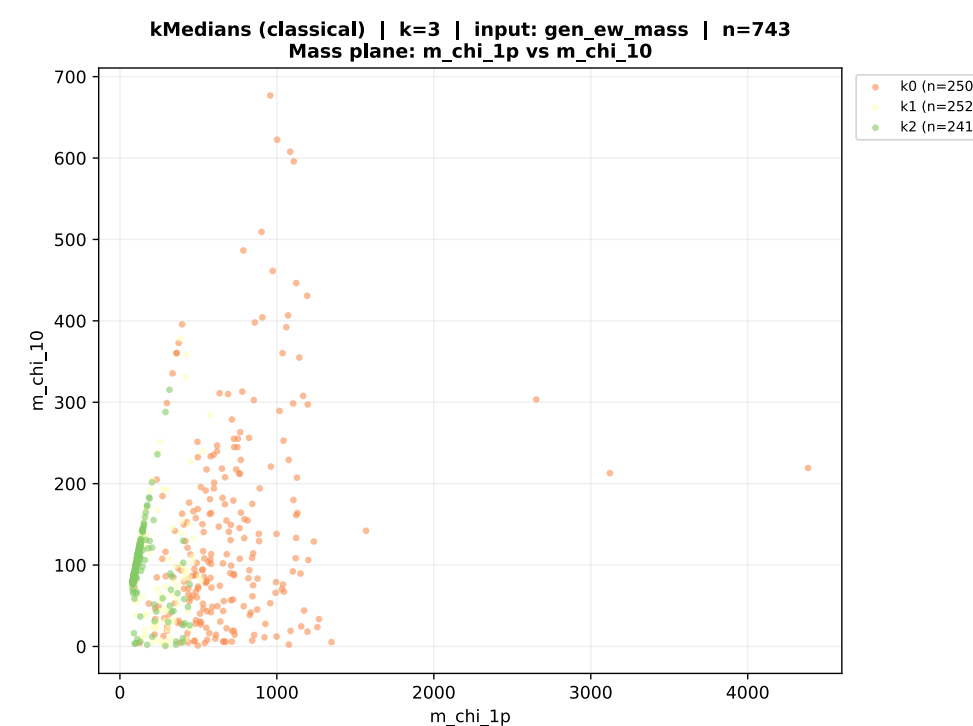
## kMeans (classical)



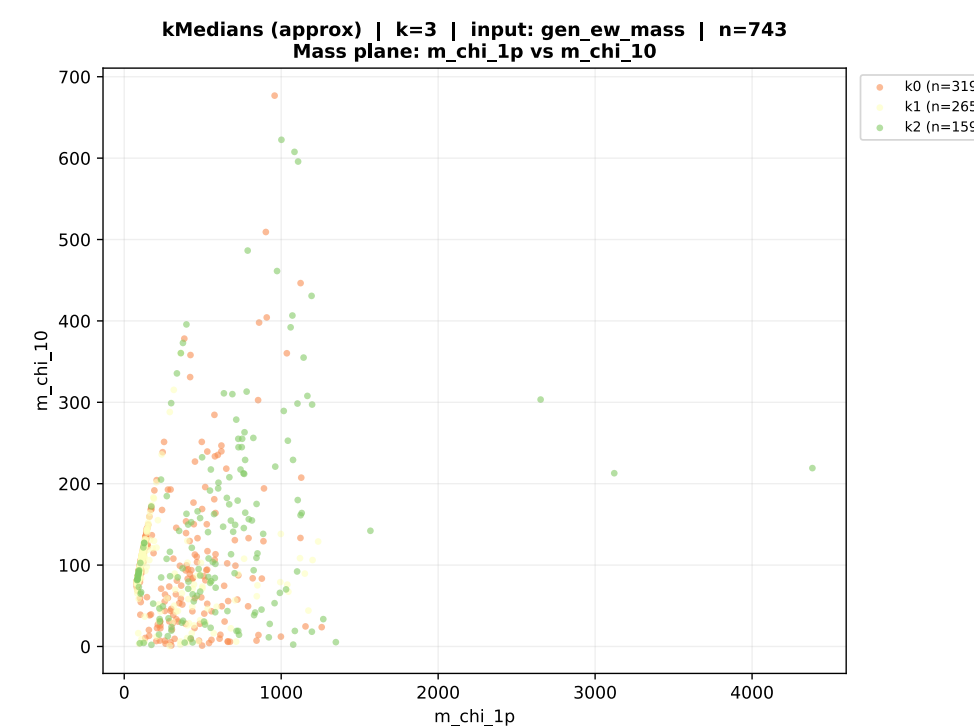
## kMeans (approx)



## kMedians (classical)

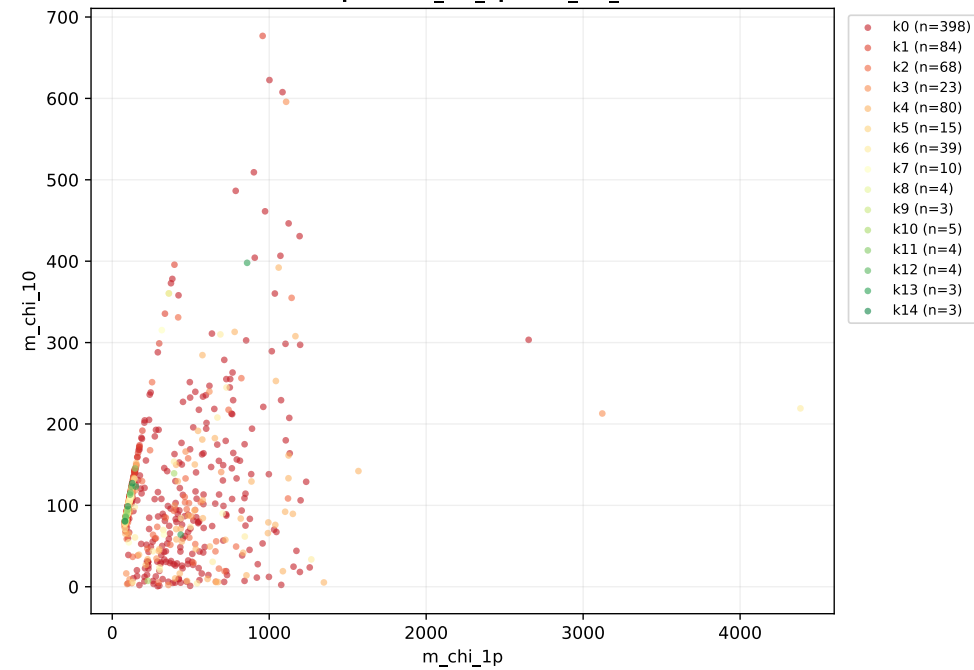


## kMedians (approx)

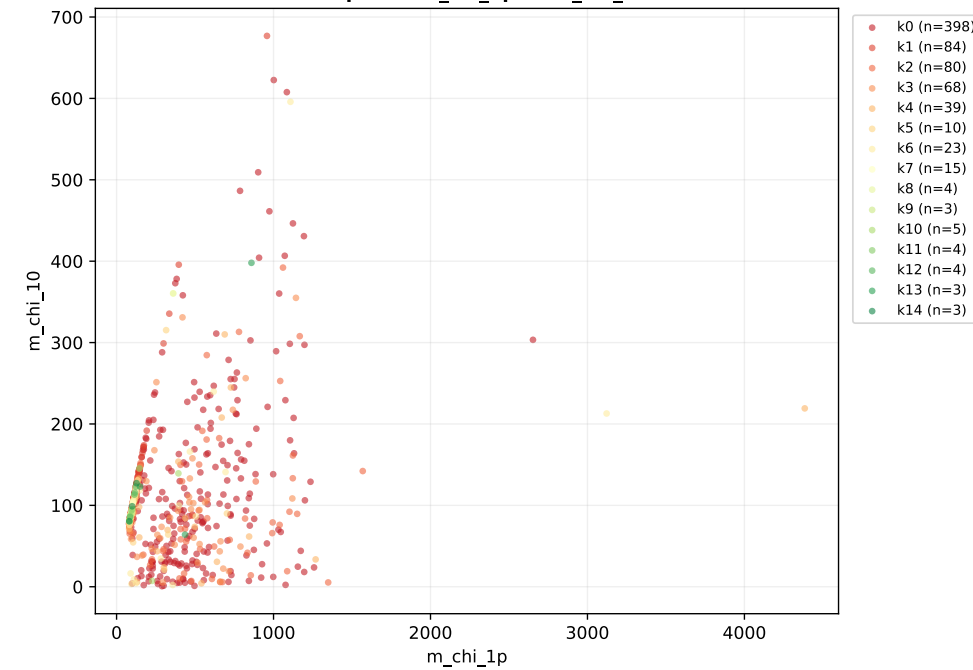


constr

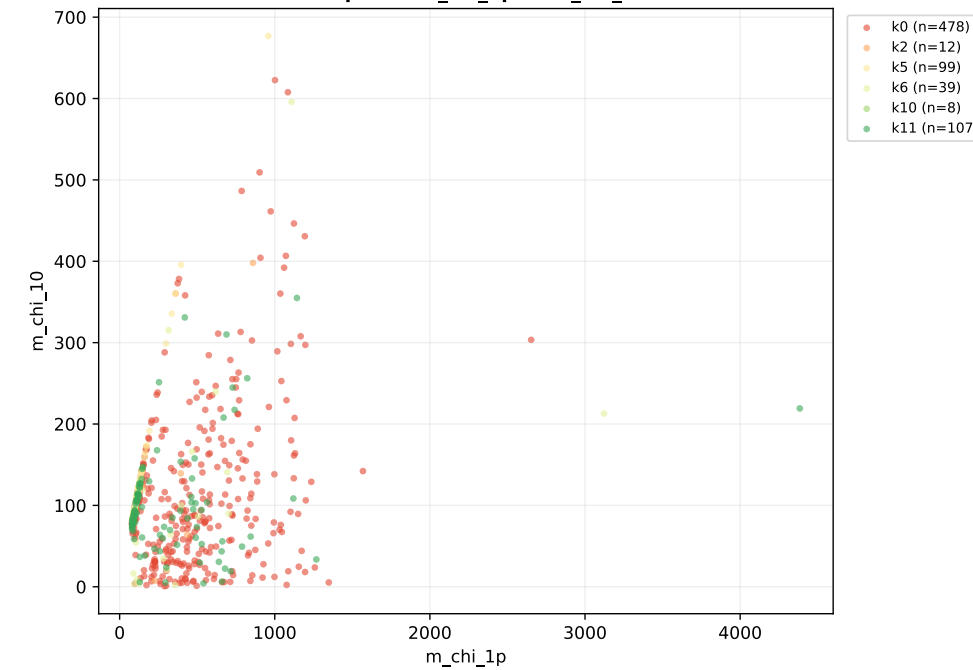
## kMeans (classical)



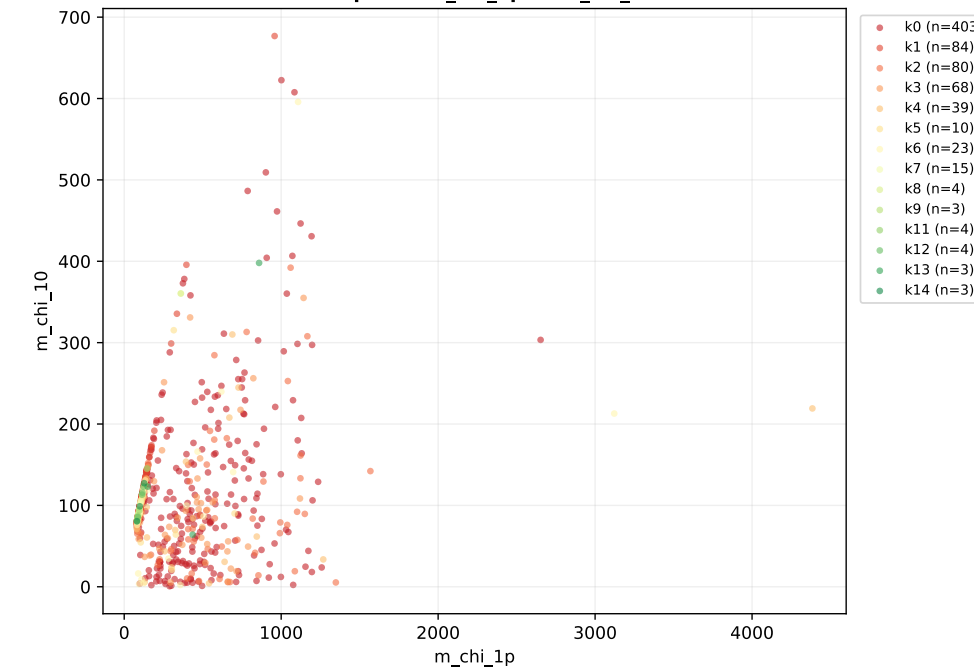
## kMeans (approx)



## kMedians (classical)

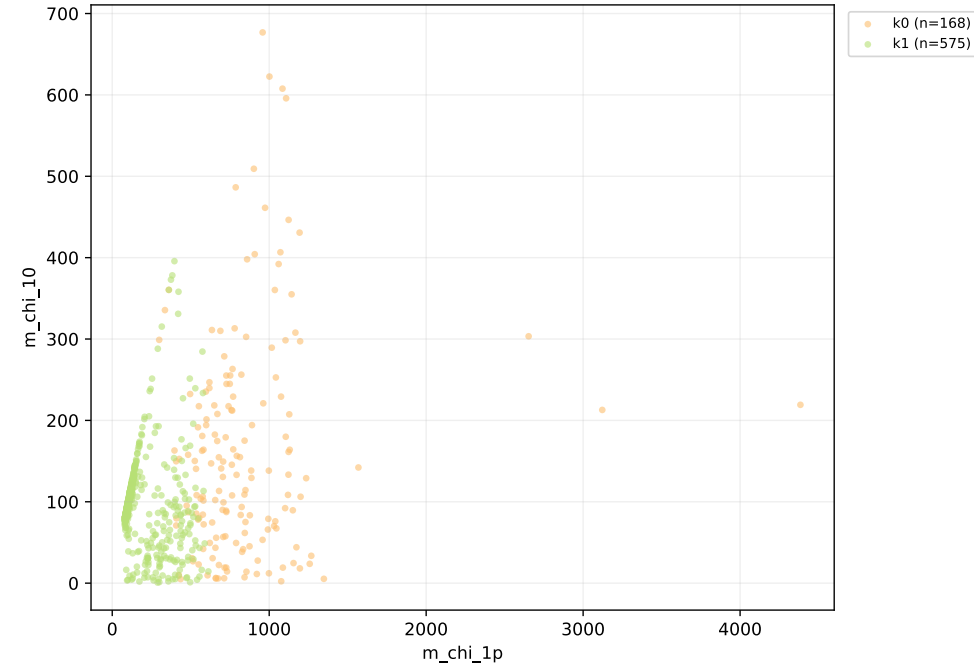


## kMedians (approx)

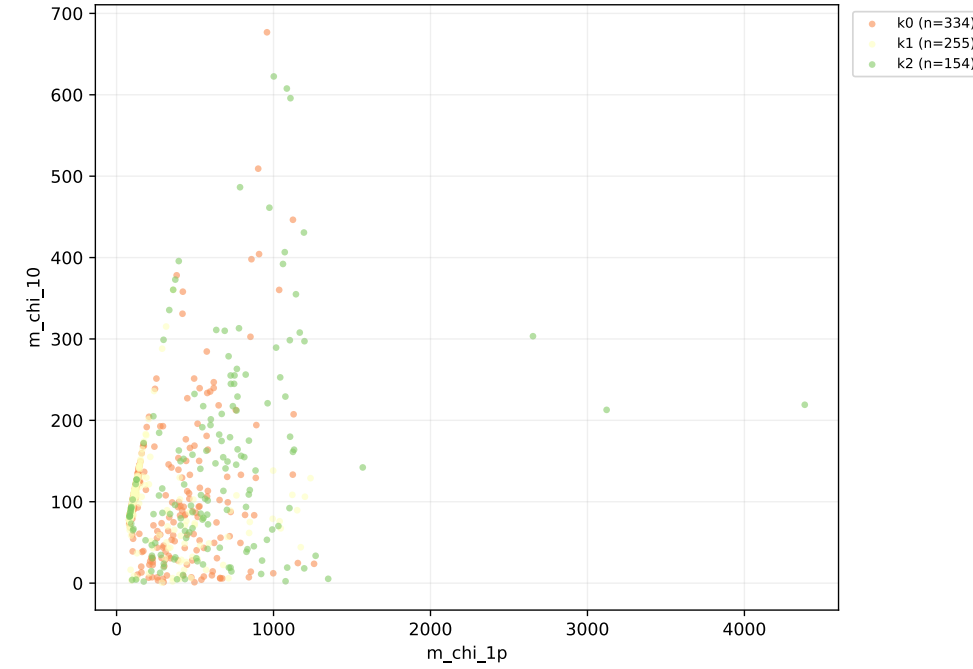


constr\_mass

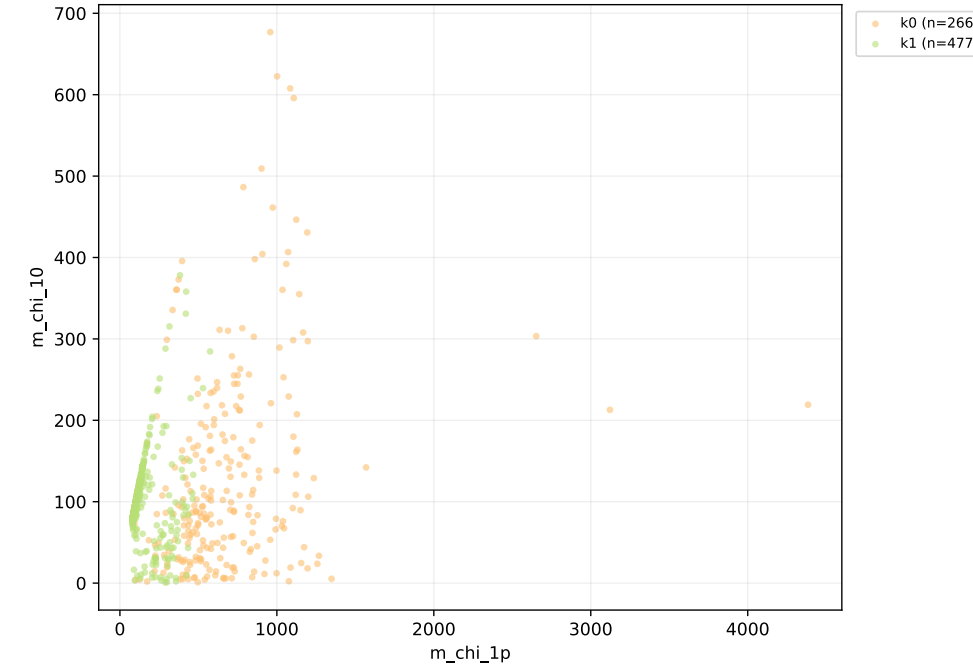
## kMeans (classical)



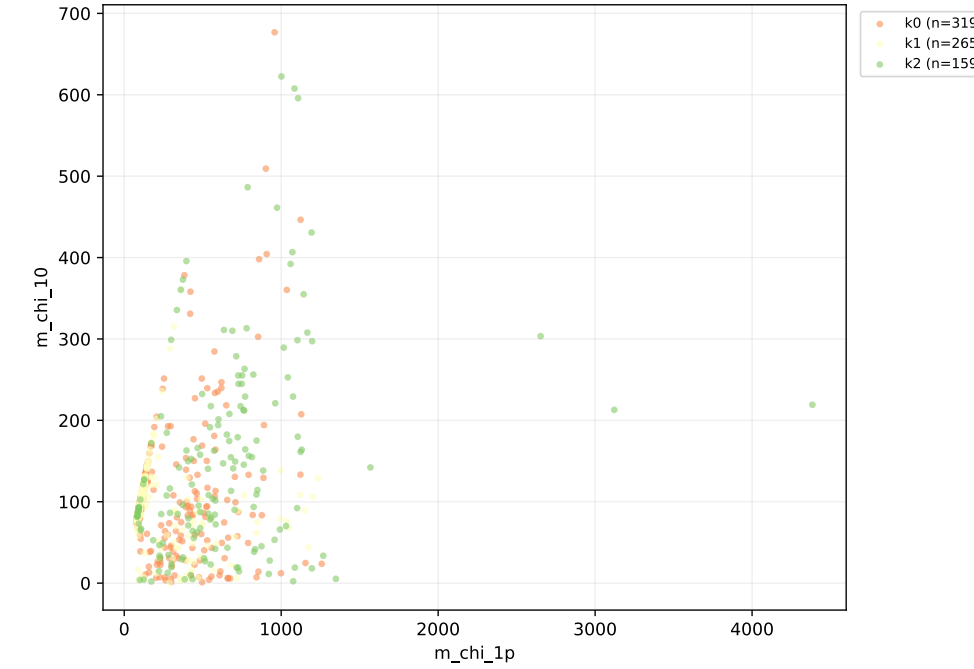
## kMeans (approx)



## kMedians (classical)

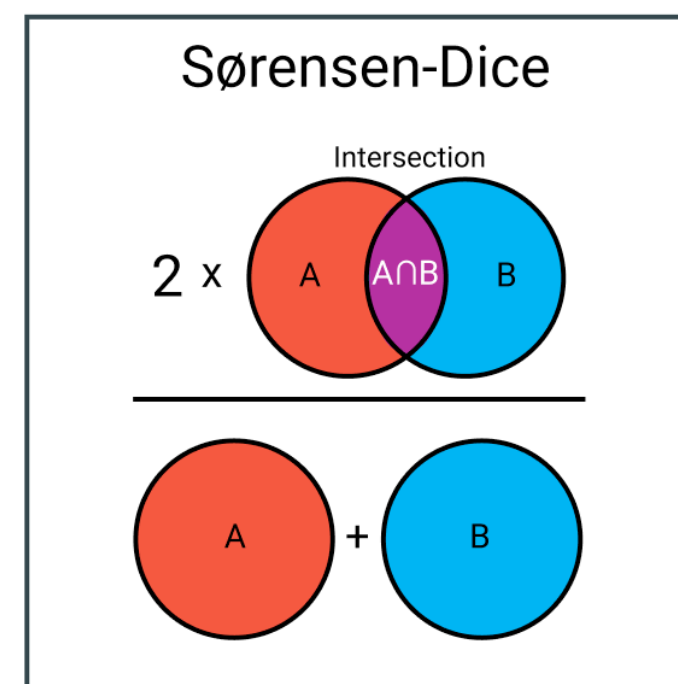
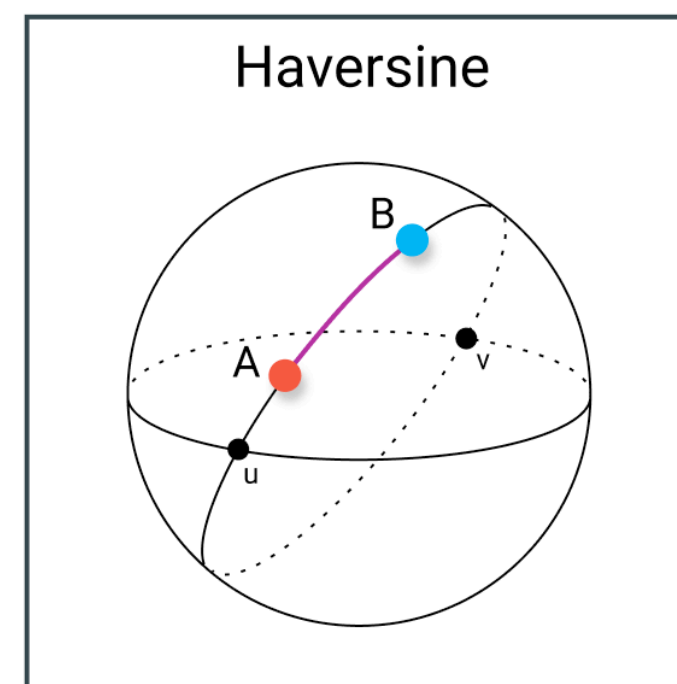
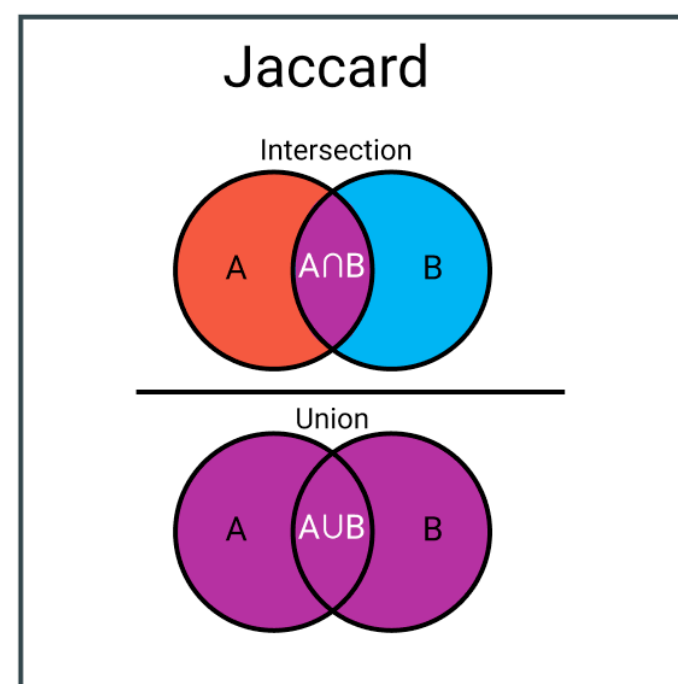
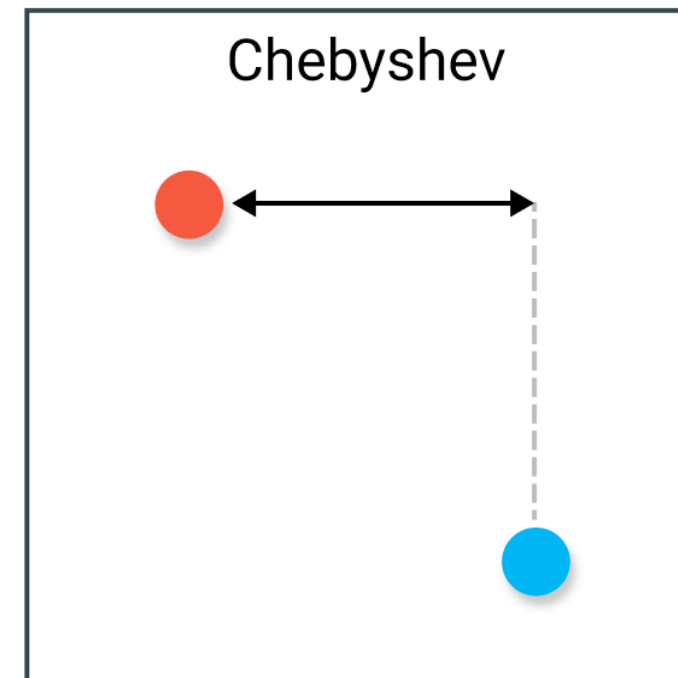
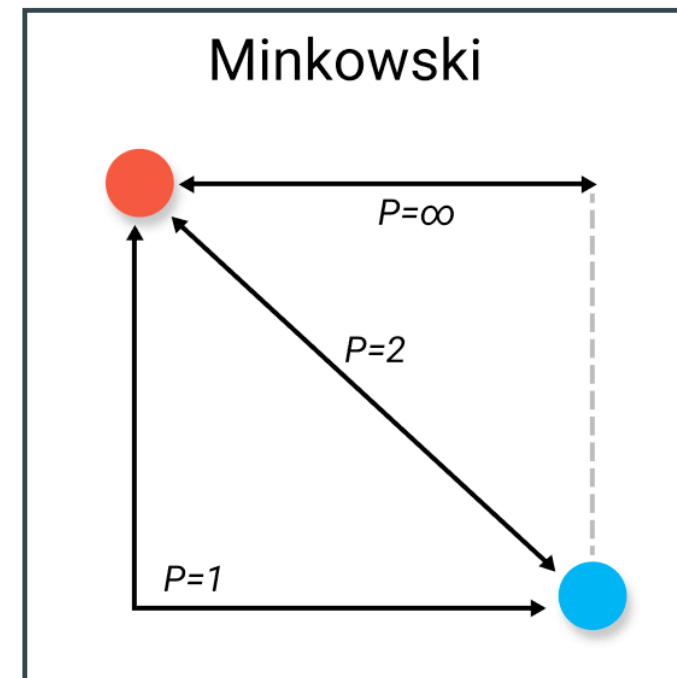
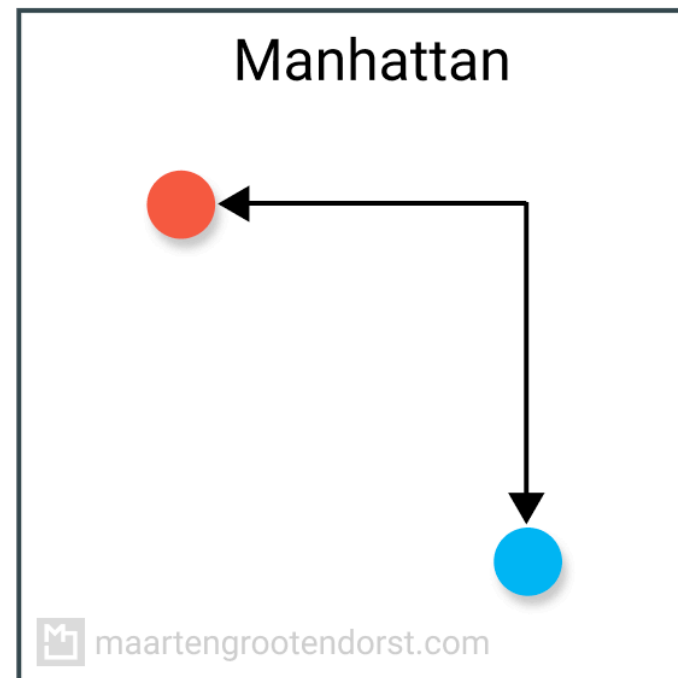
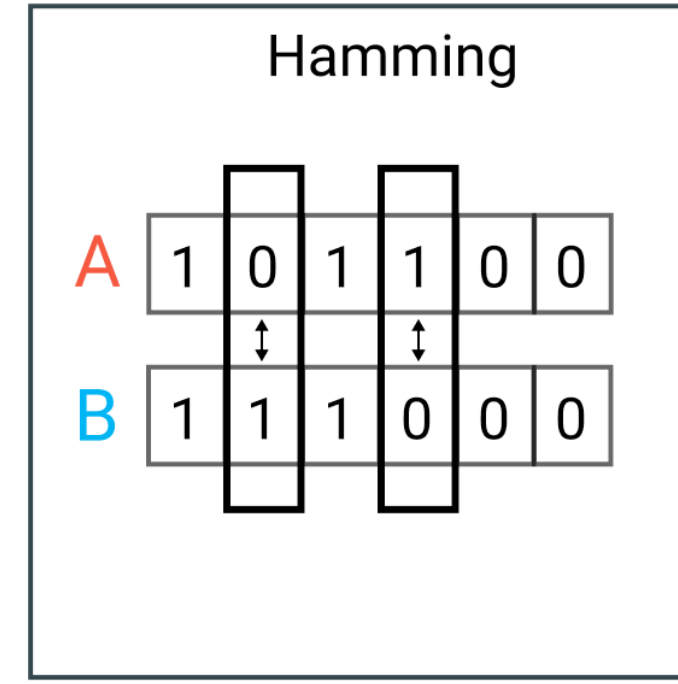
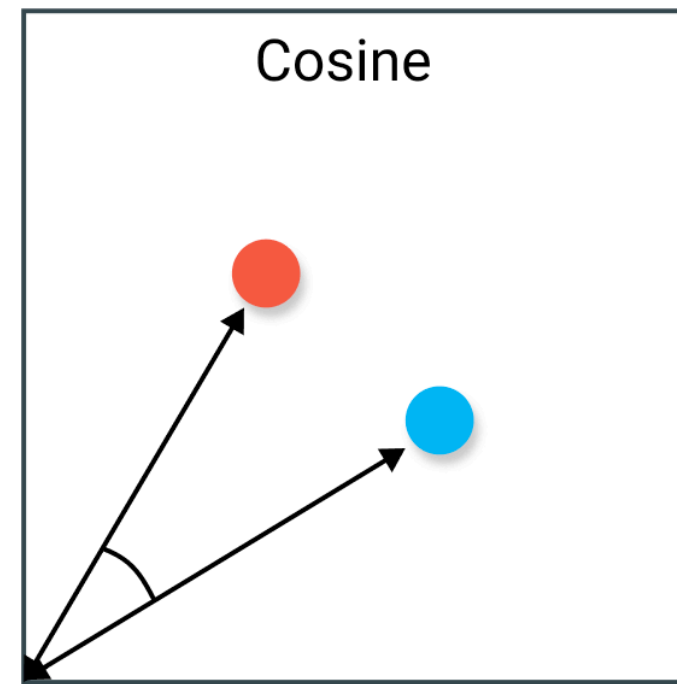
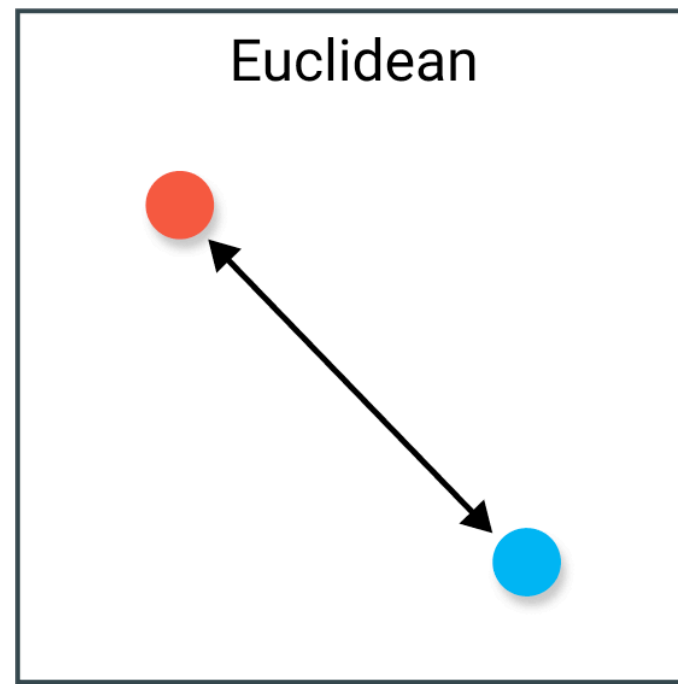


## kMedians (approx)





# DISTANCE MEASURES



## Euclidean distance

- Cartesian coordinates distance
- Not scale invariant (units change the result), should use normalization
- Con: Less useful with high number of dimensions
- Works best for low dimensional data, keeps info on magnitude of vectors
- Usage: Simple and intuitive, used for simple examples with these constraints

## Cosine similarity

- Cosine of angle between vectors (two vectors in the same orientation have cos of the angle 1, opposite have -1)
- Counteracts the Euclidean high dimensionality problem
- Con: Magnitude of vectors not taken into account
- **Usage: For high dimensional data where magnitude is disregarded**

## Hamming distance

- Number of values that are different between two vectors
- Con: Difficult to use for vectors of different length
- Usage: Error detection when data is transmitted over network, or for measuring distance between categorical variables

## Manhattan distance

- Distance between real-valued vectors (distance between two vectors in right angles, no diagonal movement)
- **Can be used with high dimensional data**
- **Con: Less intuitive than Euclidean distance**
- **Usage: For discrete or binary data**

## Minkovski distance

- Used in Normed vector space (n-dimensional real space), i.e. for vectors with length
- It has 3 requirements: zero vector, scalar factor and triangle inequality
- It projects with parameter p to E, M and C
- Con: Same restrictions as E, M and C
- Usage: iterate on p easily and find the best option

## Cherished distance

- Maximum distance along one axis (Chessboard distance, minimum number of King moves from one square to another)
- Con: For specific use cases only
- Usage: Useful measure in games, e.g. warehouse crane making movements from single spot

## Jaccard index

- **Intersection over union, calculates similarity and diversity of the sample**
- **Con: Influenced by the size of the data**
- Usage: For binary data, for similarity of images and text, parameter sets comparison

## Haversine distance

- Distance of two points on a sphere, given longitude and latitude (similar to Euclidean distance without a straight line)
- Con: Assumes points lie on a sphere
- Usage: in navigation, distance between two points when flying over Earth, no advantage for short distances

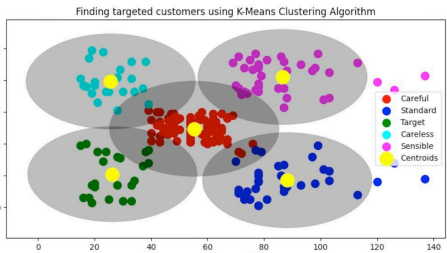
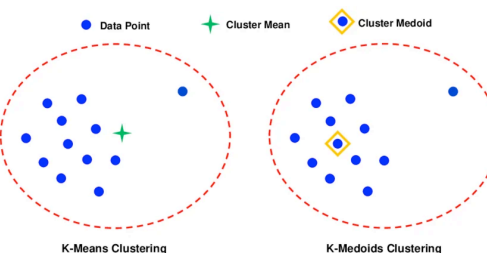
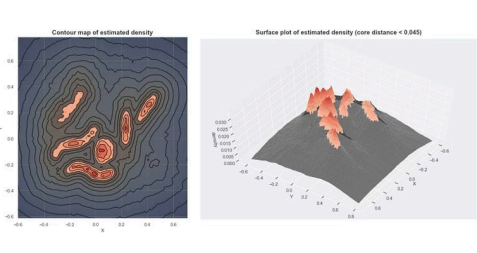
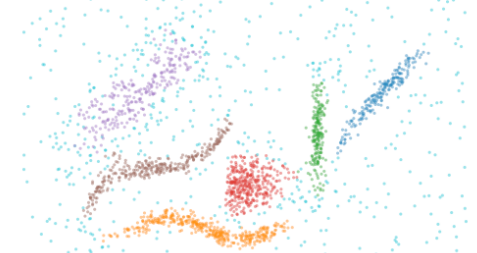
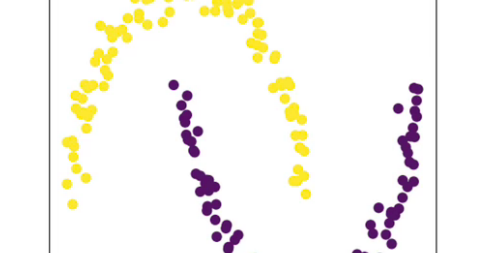
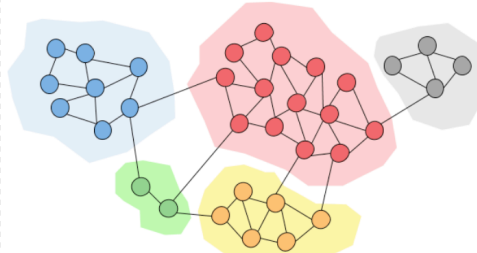
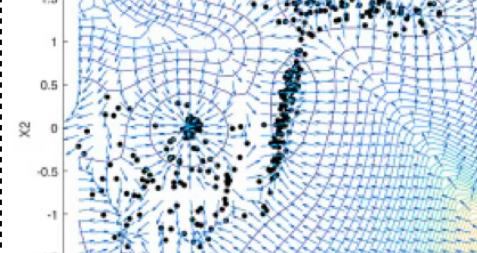
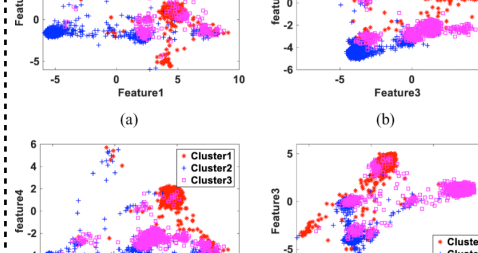
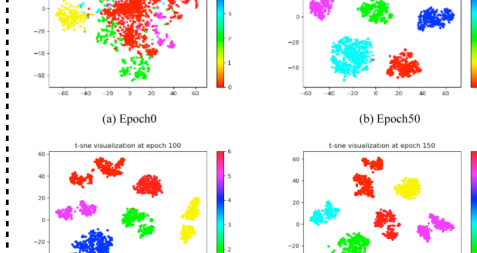
## Sørensen-Dice index

- **Similarity and diversity of sample sets. Percentage of overlap between two sets (similar to J), value [0,1]**
- **Con: Higher weight for bigger sets (same as J)**
- Usage: In image segmentation or text similarity analysis

# CLUSTERING ALGORITHMS

Binary: Cosine, Jaccard, Dice  
Continuous sparse: Cosine, Bray-Curtis, Canberra  
Mixed: Gower

\* Colored are the best performing combinations

|             |                                                                                   |                                                                                   |                                                                                     |                                                                                     |                                                                                     |                                                                                     |                                                                                     |                                                                                     |                                                                                     |
|-------------|-----------------------------------------------------------------------------------|-----------------------------------------------------------------------------------|-------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------|
|             |  |  |  |  |  |  |  |  |  |
| Principle   | Centroid                                                                          | Medoid                                                                            | Density + variable density<br>+ works with distance matrix                          | Density, automatically finds the optimal minimum cluster size                       | Graph-based, works with similarity (cosine, Jaccard)                                | Graph community                                                                     | Kernel-based, finds potential minima, good for complex shapes                       | Deep (autoencoder)                                                                  | Deep (contrastive)                                                                  |
| Best for    | Distance concentration                                                            | Mixed data                                                                        | Clusters with different density, simple clusters                                    | Variable density, very high dimensional, stable across scales                       | Non-linear / manifold clusters.                                                     | Network / graph clusters                                                            | Complex potential landscape / overlapping clusters                                  | Learned representation                                                              | Learned representation                                                              |
| High-dim    | ★                                                                                 | ★★                                                                                | ★★★                                                                                 | ★★★                                                                                 | ★★★                                                                                 | ★★★                                                                                 | ★★★                                                                                 | ★★★                                                                                 | ★★★                                                                                 |
|             | k-means                                                                           | K-medoids                                                                         | HDBSCAN                                                                             | PLSCAN                                                                              | Spectral                                                                            | Leiden                                                                              | FM-PQC                                                                              | IDECA                                                                               | Contrastive                                                                         |
| Gower       | -                                                                                 | ★★★★                                                                              | ★★★                                                                                 | ★★★                                                                                 | ★★                                                                                  | ★★                                                                                  | ★★★★                                                                                | ★★★                                                                                 | ★★★                                                                                 |
| Cosine      | ★                                                                                 | ★★★                                                                               | ★★★                                                                                 | ★★★                                                                                 | ★★★                                                                                 | ★★★                                                                                 | ★★★                                                                                 | ★★★                                                                                 | ★★★                                                                                 |
| Jaccard     | -                                                                                 | ★★★                                                                               | ★★★                                                                                 | ★★★                                                                                 | ★★                                                                                  | ★★★                                                                                 | ★★★                                                                                 | ★★★                                                                                 | ★★★                                                                                 |
| Dice        | -                                                                                 | ★★★                                                                               | ★★★                                                                                 | ★★★                                                                                 | ★★                                                                                  | ★★★                                                                                 | ★★★                                                                                 | ★★★                                                                                 | ★★★                                                                                 |
| Bray-Curtis | -                                                                                 | ★★★                                                                               | ★★★                                                                                 | ★★★                                                                                 | ★★                                                                                  | ★★                                                                                  | ★★★                                                                                 | ★★★                                                                                 | ★★★                                                                                 |
| Canberra    | -                                                                                 | ★★★                                                                               | ★★★                                                                                 | ★★★                                                                                 | ★★                                                                                  | ★★                                                                                  | ★★★                                                                                 | ★★★                                                                                 | ★★★                                                                                 |
| Manhattan   | ★                                                                                 | ★★★                                                                               | ★★★                                                                                 | ★★★                                                                                 | ★★                                                                                  | ★★                                                                                  | ★★★                                                                                 | ★★★                                                                                 | ★★★                                                                                 |
| Correlation | -                                                                                 | ★★★                                                                               | ★★★                                                                                 | ★★★                                                                                 | ★★★                                                                                 | ★★★                                                                                 | ★★★                                                                                 | ★★★                                                                                 | ★★★                                                                                 |
| Euclidean   | ★★★                                                                               | ★★★                                                                               | ★★                                                                                  | ★★                                                                                  | ★★★                                                                                 | ★★                                                                                  | ★★★                                                                                 | ★★★                                                                                 | ★★★                                                                                 |
| Hamming     | -                                                                                 | ★★★                                                                               | ★★★                                                                                 | ★★★                                                                                 | ★                                                                                   | ★★★                                                                                 | ★★                                                                                  | ★★                                                                                  | ★★                                                                                  |
| Mahalanobi  | ★★                                                                                | ★★★                                                                               | ★★                                                                                  | ★★                                                                                  | ★                                                                                   | ★★                                                                                  | ★★                                                                                  | ★★★                                                                                 | ★★★                                                                                 |



# DISTANCE MEASURES

Input: binary, sparse continuous, mixed

| Distance              | High-dim | Binary | Sparse | Mixed | Handles Zeros                | Measures                  | When to Use                       | Comment for binary not sparse |                                                     | Comment for continuous and sparse |                                                 | Comment for mixed |                                    |
|-----------------------|----------|--------|--------|-------|------------------------------|---------------------------|-----------------------------------|-------------------------------|-----------------------------------------------------|-----------------------------------|-------------------------------------------------|-------------------|------------------------------------|
| Gower                 | **       | **     | **     | ***   | Yes                          | Mixed                     | Mixed binary and continuous data  | **                            | Works, but mainly useful when mixed with continuous | **                                | Works (normalizes per feature)                  | ***               | <b>Best choice for mixed data</b>  |
| Cosine                | ***      | **     | ***    | **    | Yes, ignores shared zeros    | Pattern (angle)           | High-dimensional sparse vectors,  | **                            | <b>Works for high-dim binary</b>                    | ***                               | <b>Very good for sparse high-dim continuous</b> | **                | Works if data vectorized           |
| Jaccard               | **       | ***    | ***    | –     | Yes, ignores shared zeros    | Overlap                   | Sparse binary data                | **                            | <b>Good for binary</b>                              | *                                 | Not suitable                                    | *                 | Not suitable                       |
| Dice                  | **       | ***    | ***    | –     | Yes, ignores shared zeros    | Overlap (weight)          | Sparse binary                     | **                            | <b>Good for binary</b>                              | *                                 | Not suitable                                    | *                 | Not suitable                       |
| Bray-Curtis           | **       | **     | ***    | **    | Yes, ignores shared zeros    | Counts                    | Sparse counts or histograms       | **                            | Works if binary treated as counts                   | ***                               | <b>Very good for sparse counts/continuous</b>   | **                | Can work if variables non-negative |
| Canberra (~Manhattan) | **       | **     | ***    | **    | Yes, emphasizes small values | Relative difference       | Data with many zeros              | **                            | Works but uncommon                                  | ***                               | <b>Very good for sparse continuous</b>          | **                | Works if variables numeric         |
| Manhattan             | **       | **     | **     | **    | No                           | Magnitude                 | General high-dimensional distance | **                            | Works                                               | **                                | Works but not ideal for sparse                  | **                | Works but not ideal                |
| Correlation           | **       | –      | **     | **    | Yes, mean centered           | Pattern/shape             | Pattern similarity                | *                             | Not suitable                                        | **                                | Good if pattern matters                         | **                | Works for numeric mixed            |
| Euclidean             | *        | –      | –      | **    | No                           | Magnitude                 | Low-dimensional continuous        | *                             | Not suitable                                        | *                                 | Bad for sparse                                  | **                | Works after normalization          |
| Hamming               | **       | ***    | **     | –     | No                           | Bit differences           | Binary vectors (dense)            | ***                           | Best for dense binary                               | *                                 | Not suitable                                    | *                 | Not suitable                       |
| Mahalanobis           | –        | –      | –      | **    | No                           | Distribution (covariance) | Low-dimensional continuous        | *                             | Not suitable                                        | **                                | Works for correlated variables                  | **                | Works for continuous mixed         |

# SATE-OF-THE-ART CLUSTERING ALGORITHMS

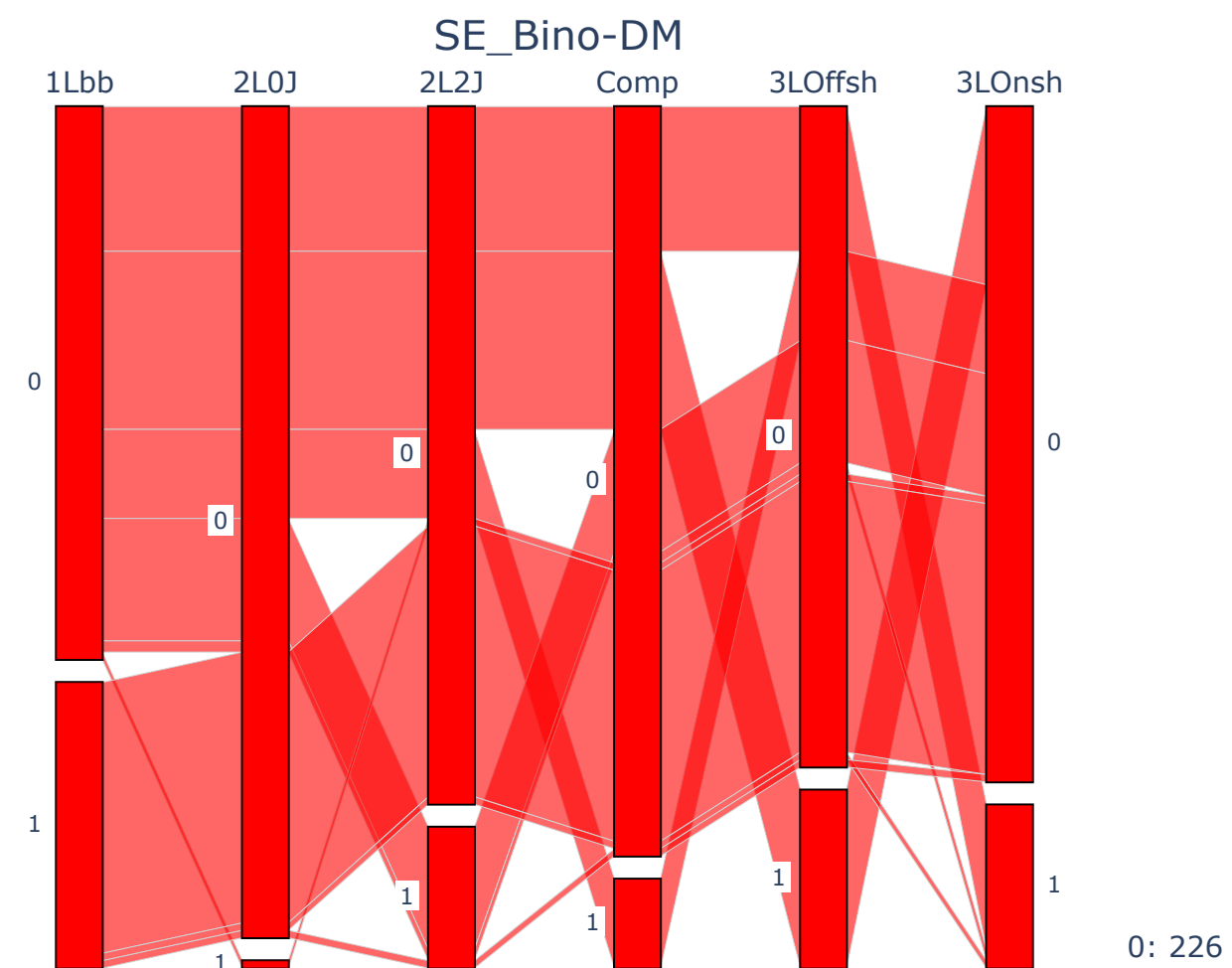
| Density-based (modern SOTA)   |                                                      | Graph-based / Spectral (modern SOTA) |                                               | Kernel / Quantum / Potential-based<br>(Modern SOTA) |                               | Deep clustering (learned representation)<br>(Very modern SOTA) |                               |
|-------------------------------|------------------------------------------------------|--------------------------------------|-----------------------------------------------|-----------------------------------------------------|-------------------------------|----------------------------------------------------------------|-------------------------------|
| DBSCAN                        | Classic density clustering                           | <b>Spectral Clustering</b>           | Graph Laplacian clustering                    | Quantum Clustering (QC)                             | Potential function clustering | <b>Deep Embedded Clustering (DEC)</b>                          | Autoencoder + clustering      |
| OPTICS                        | Density clustering with variable density             | <b>Leiden algorithm</b>              | Modern graph clustering (better than Louvain) | Probability Quantum Clustering (PQC)                | Quantum probability model     | IDEC                                                           | Improved DEC                  |
| <b>HDBSCAN</b>                | Hierarchical density clustering                      | Louvain                              | Community detection                           | <b>FM-PQC</b>                                       | Fast Multiscale PQC           | <b>Deep Spectral Clustering</b>                                | Deep + spectral               |
| <b>PLSCAN</b>                 | Persistent multiscale density clustering (new, 2025) | HCS                                  | Highly connected subgraphs                    | Mean Shift                                          | Kernel density clustering     | <b>Contrastive Clustering</b>                                  | Representation learning       |
| Density Peak Clustering (DPC) | Finds cluster centers via density peaks              | EVINGCA                              | Adaptive graph clustering (2025)              | Kernel k-means                                      | Kernel clustering             | Autoencoder + clustering                                       | Feature learning + clustering |
| CoreSPECT                     | Combines density + geometry                          | CluStRE                              | Streaming graph clustering (2025)             | Gaussian Mixture Models                             | Probabilistic clustering      | Graph Neural Network clustering                                | Graph deep clustering         |
| SUBCLU / PreDeCon             | Subspace density clustering                          | SSC / LRR                            | Subspace spectral clustering                  | Kernel Density Estimation Clustering                | Density via kernel            | Deep Subspace Clustering                                       | Subspace + deep learning      |

\* Highlighted are the state-of-the-art algorithms that work well with high number of dimensions.



# NUMBER OF CLUSTERS HUMAN VS CLUSTERING ALGORITHM

# BINO-DM



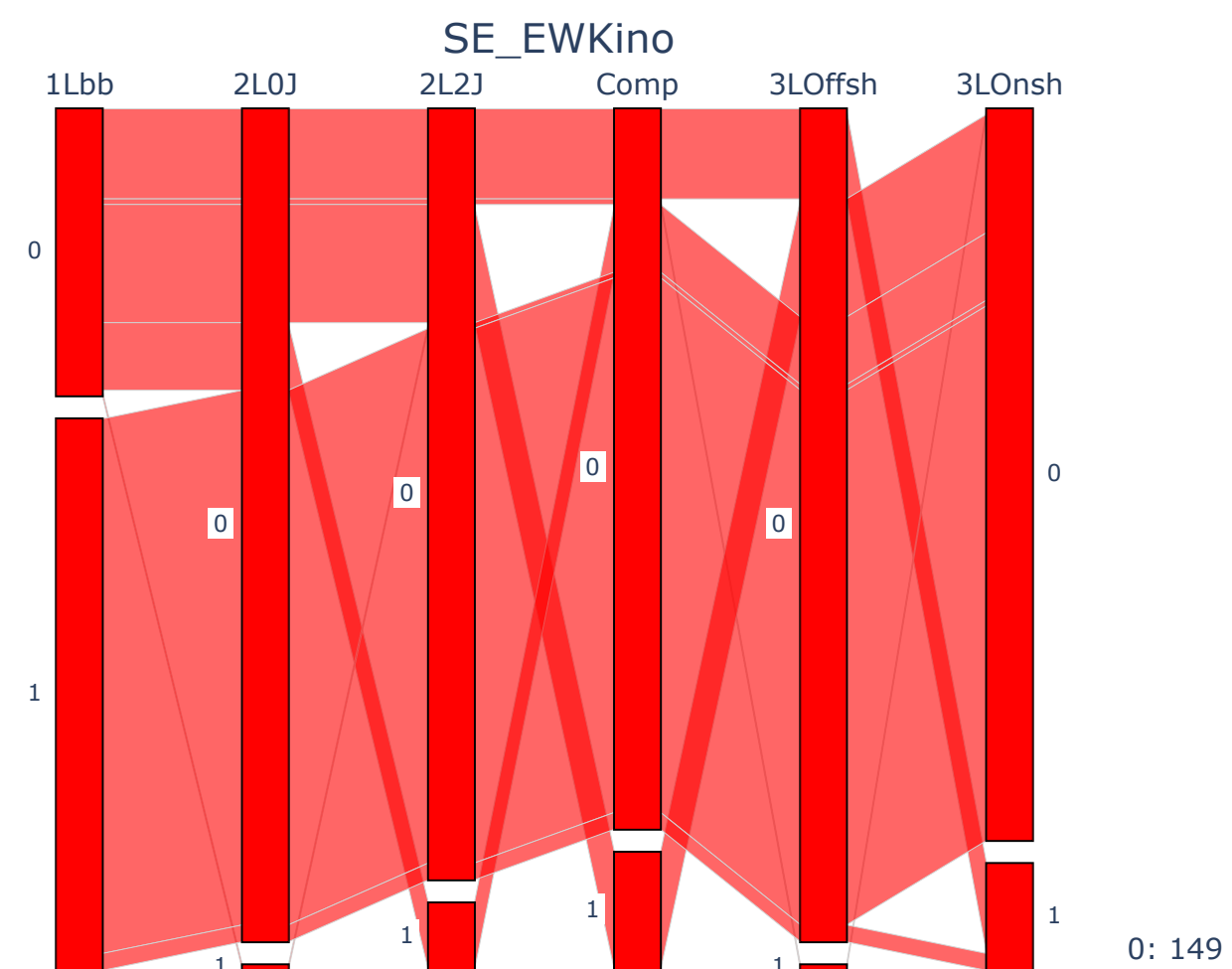
## EWK pMSSM scan

SE and external constraints 15 **categories** (combinations)

\* Missing Higgs mass constraint in this round.

```
['Constraints__DM']
['Constraints__EW']
['Constraints__Flavour']
['SurvATLAS']
['Constraints__DM', 'Constraints__EW']
['Constraints__DM', 'Constraints__Flavour']
['Constraints__DM', 'SurvATLAS']
['Constraints__EW', 'Constraints__Flavour']
['Constraints__EW', 'SurvATLAS']
['Constraints__Flavour', 'SurvATLAS']
['Constraints__DM', 'Constraints__EW', 'Constraints__Flavour']
['Constraints__DM', 'Constraints__EW', 'SurvATLAS']
['Constraints__DM', 'Constraints__Flavour', 'SurvATLAS']
['Constraints__EW', 'Constraints__Flavour', 'SurvATLAS']
['Constraints__DM', 'Constraints__EW', 'Constraints__Flavour', 'SurvATLAS']
```

# EWKINO



Look into  
models  
surviving any  
analysis

Look into  
models  
surviving  
multiple  
analyses

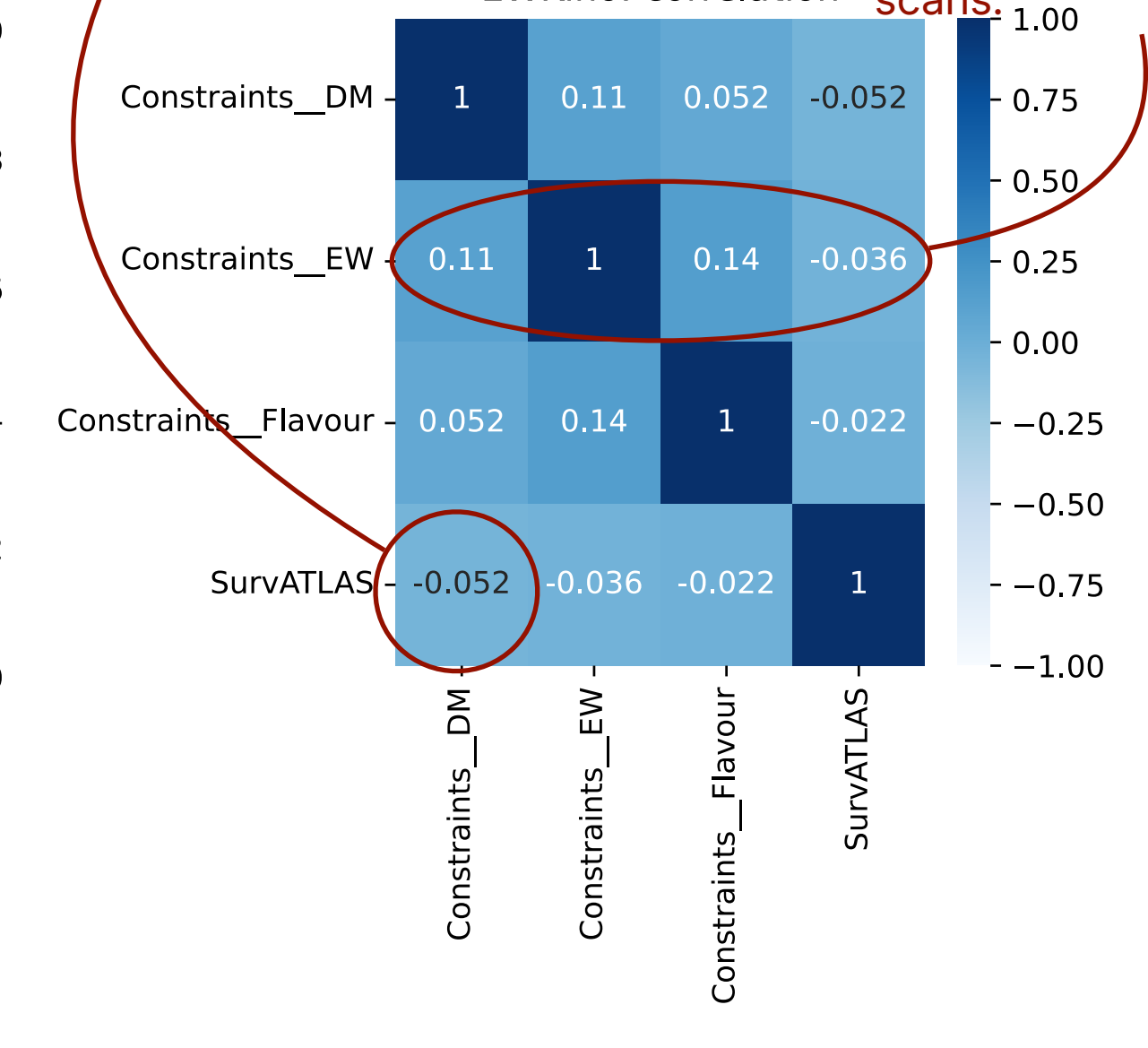
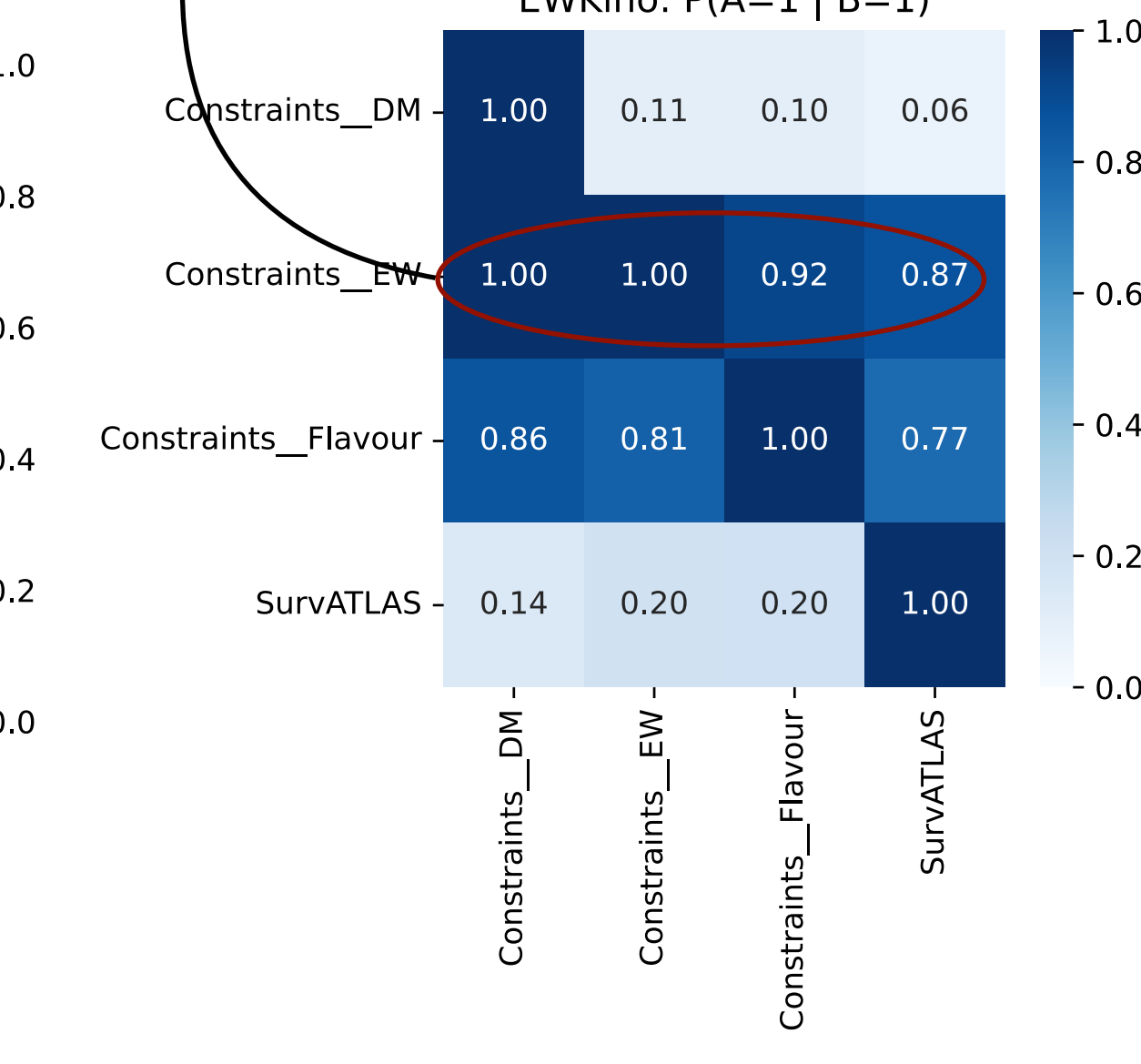
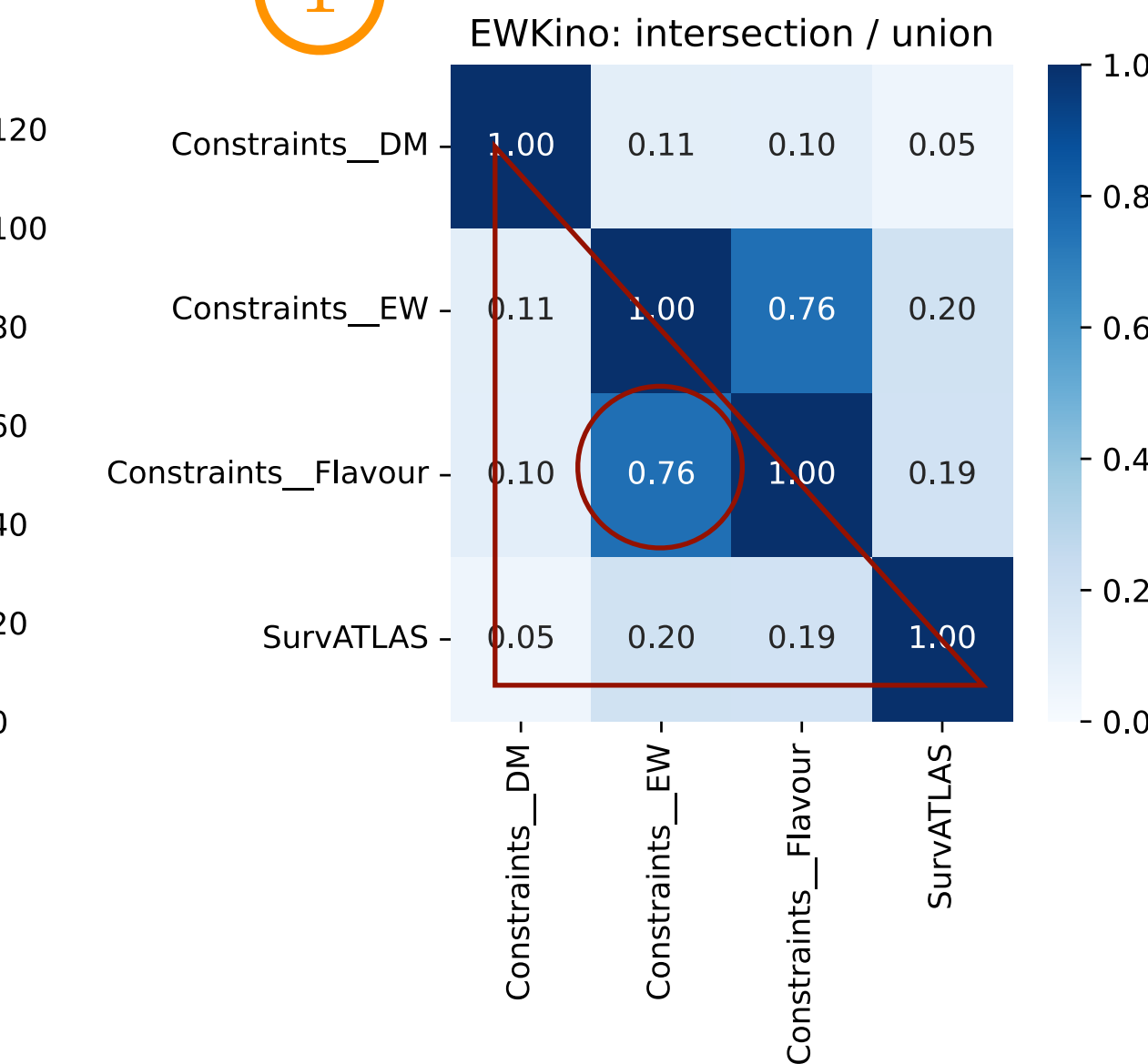
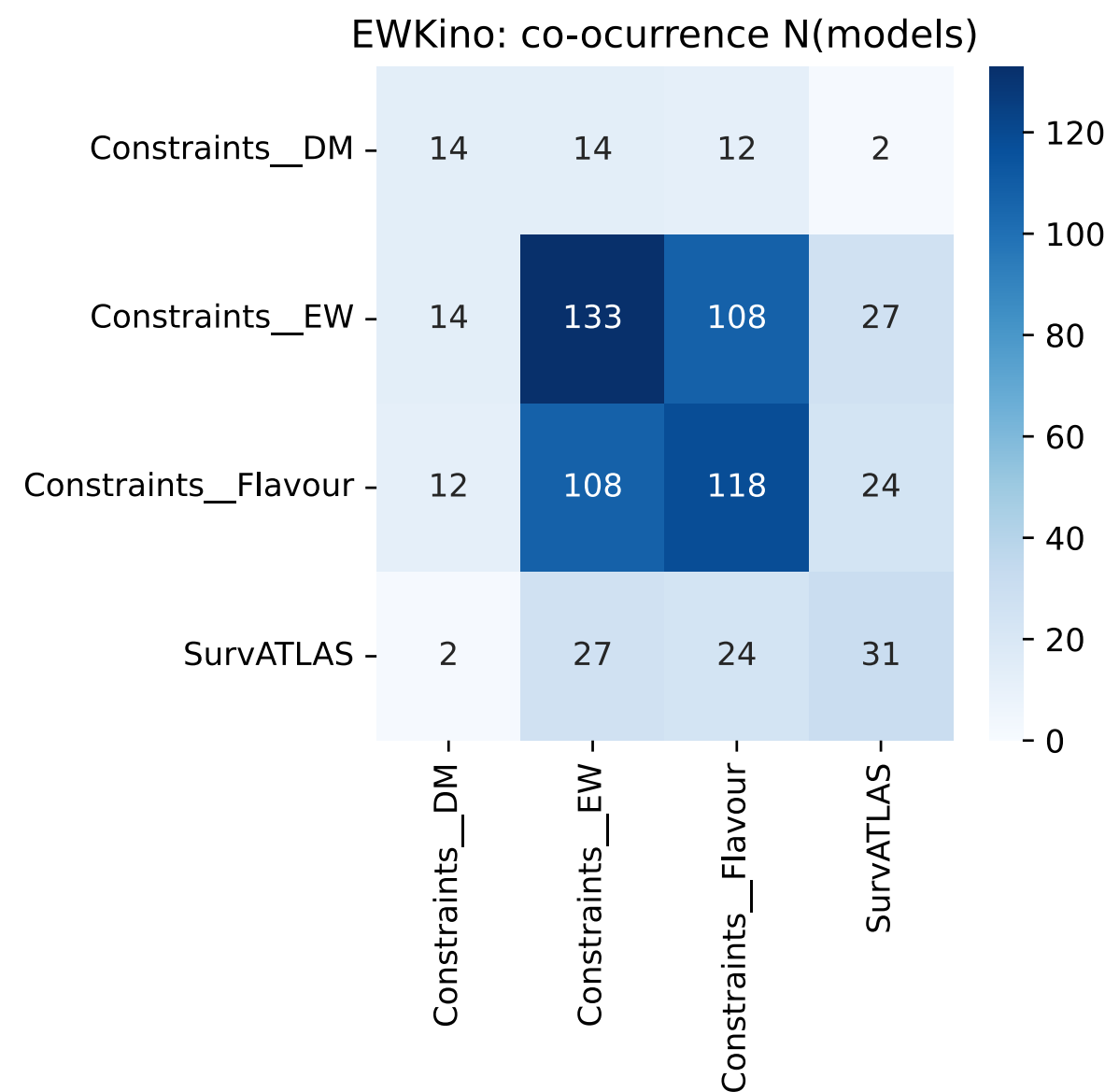
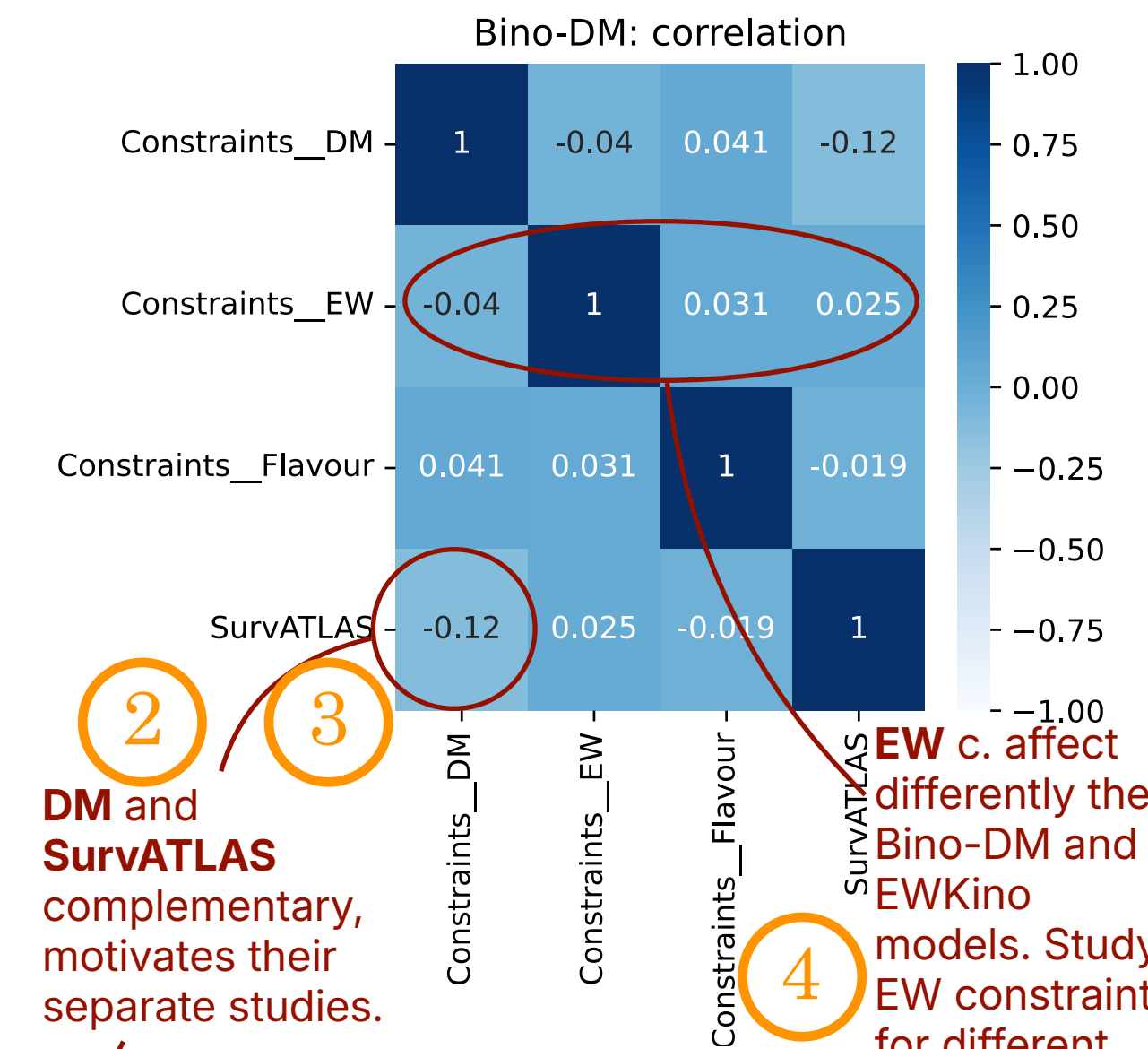
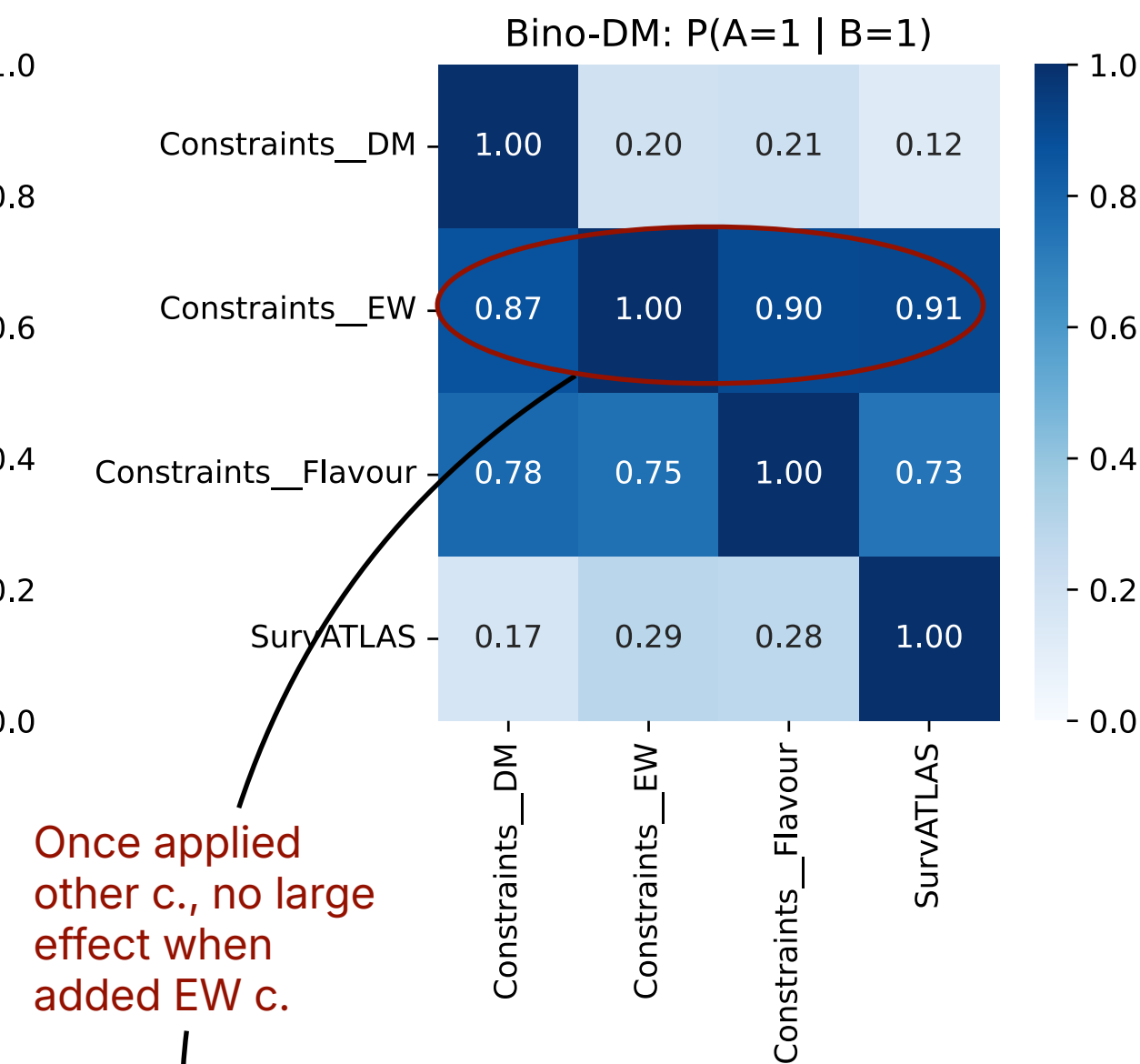
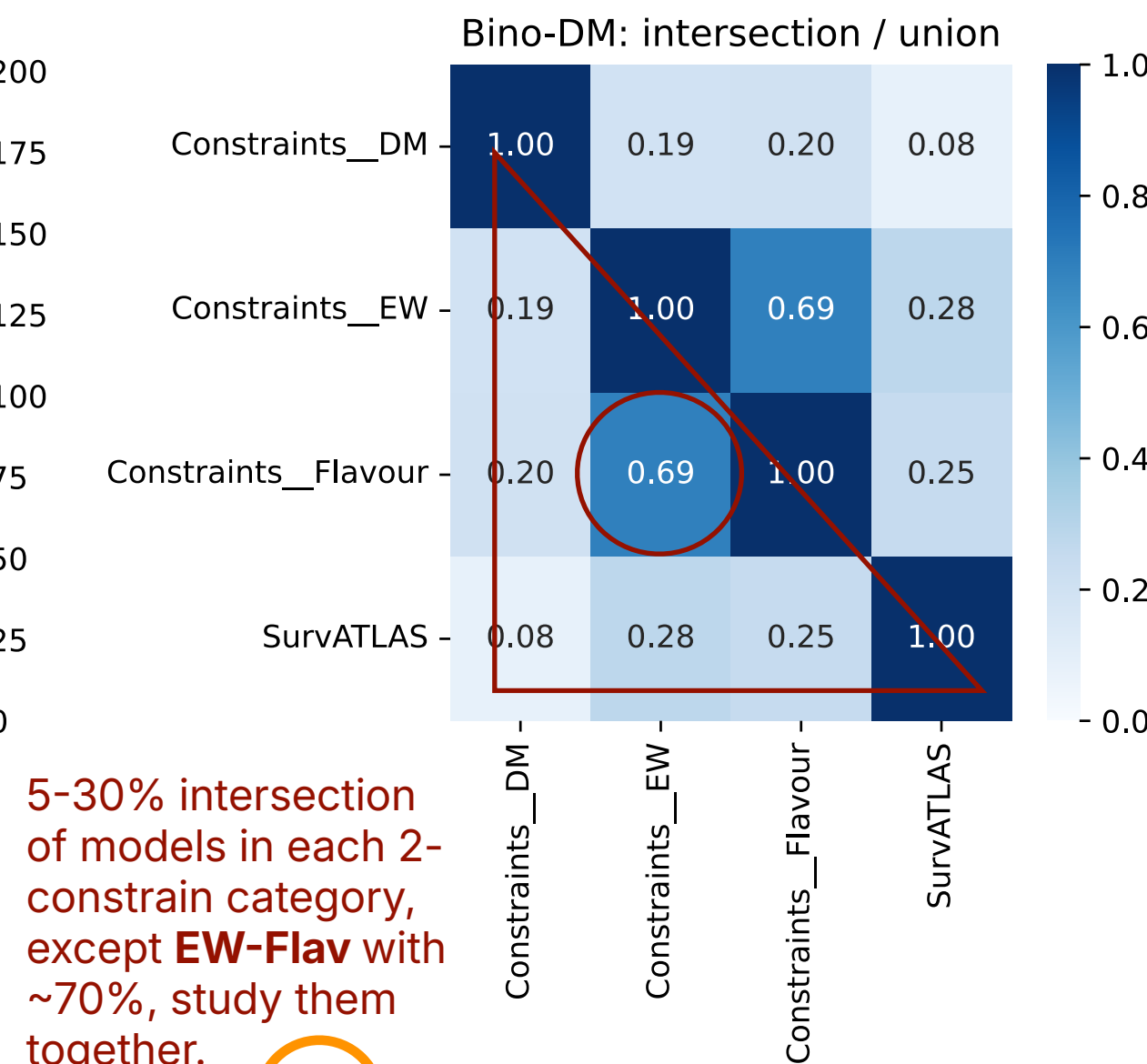
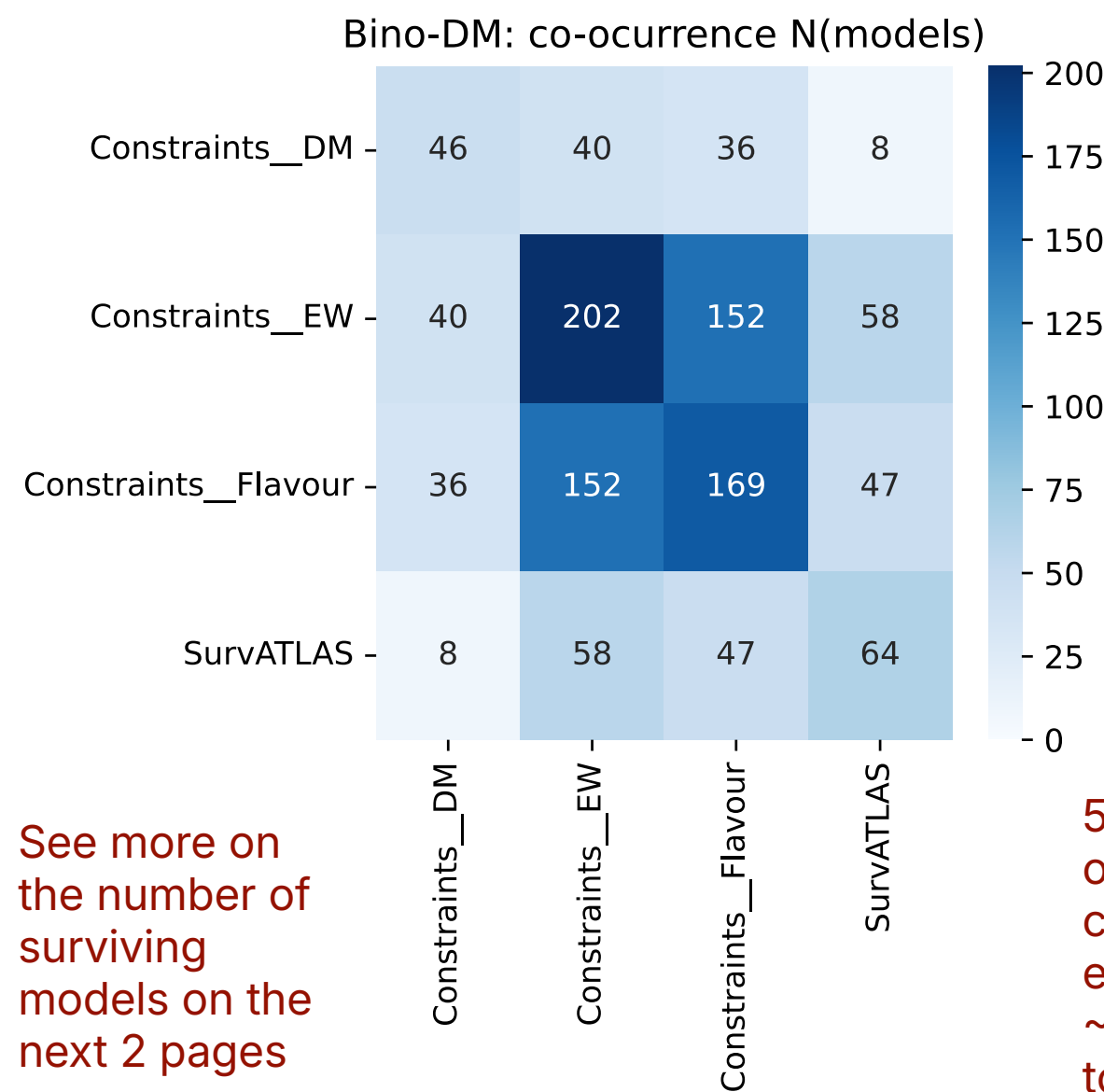
Look into  
these models  
in relation to  
external  
constraints



# SE + EXT.CONSTR. CORRELATIONS

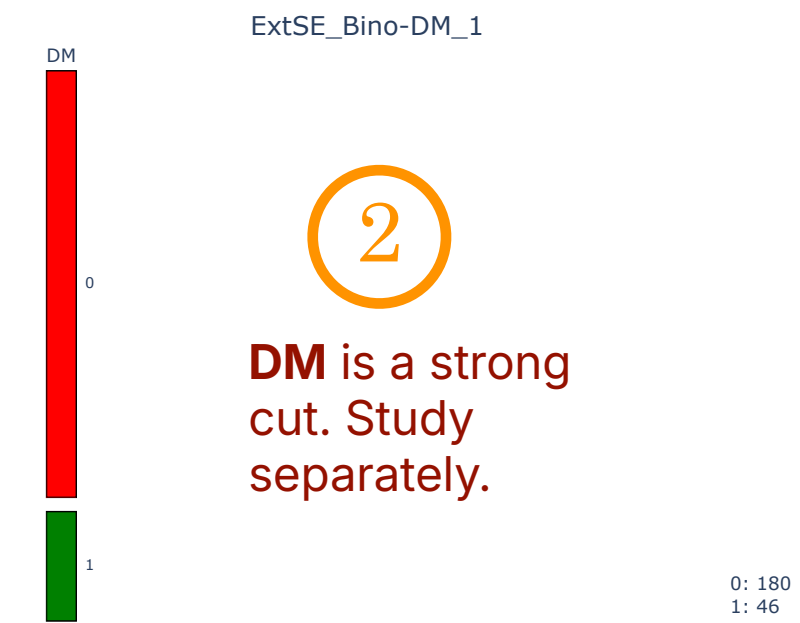
SE = True

EWKino, 149 SEM Bino-DM, 226 SEM



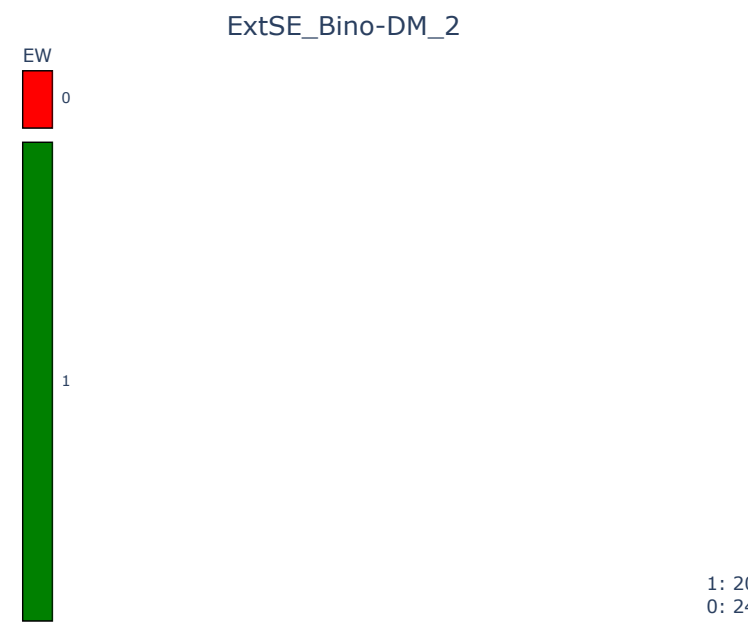
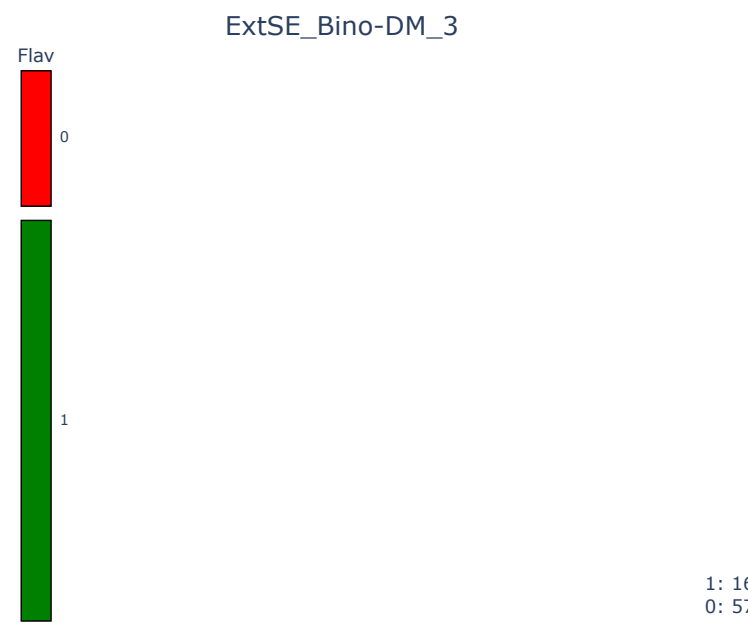
# PARALLEL SETS

## SE: BINO-DM SE = True



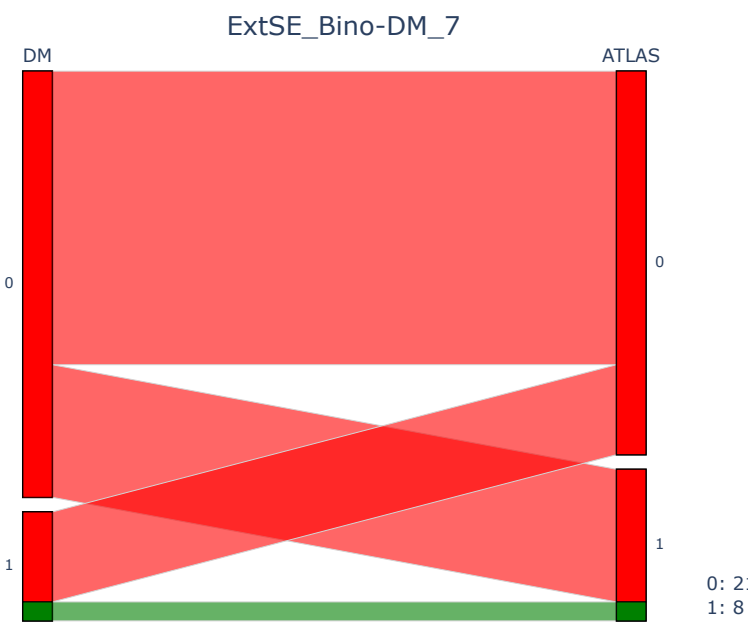
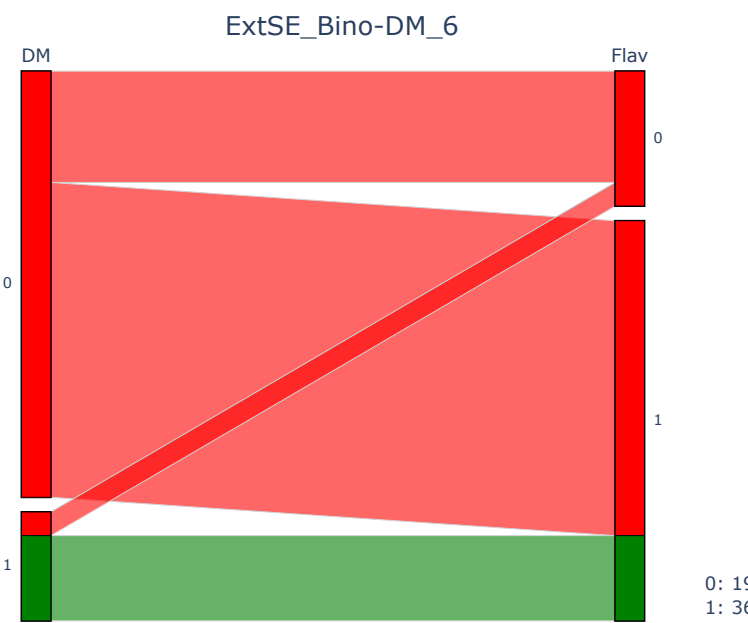
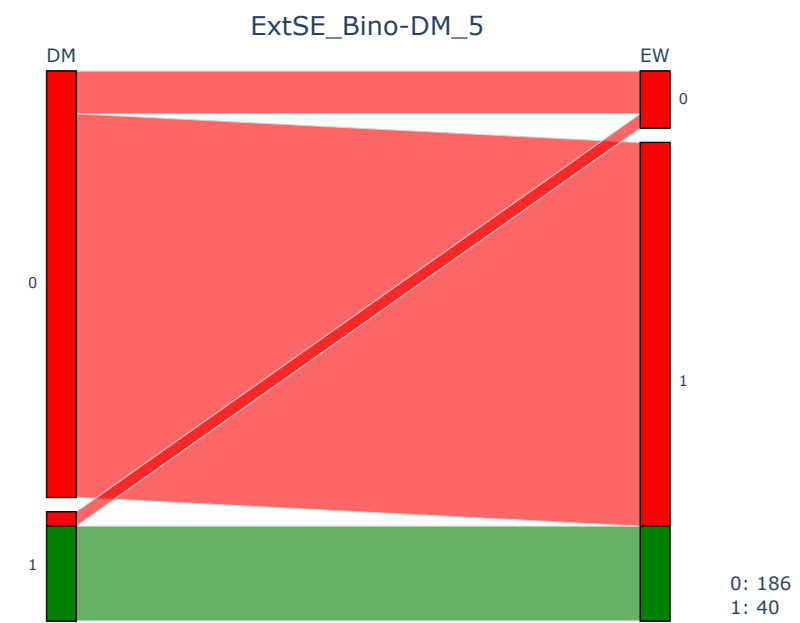
2

DM is a strong cut. Study separately.



3

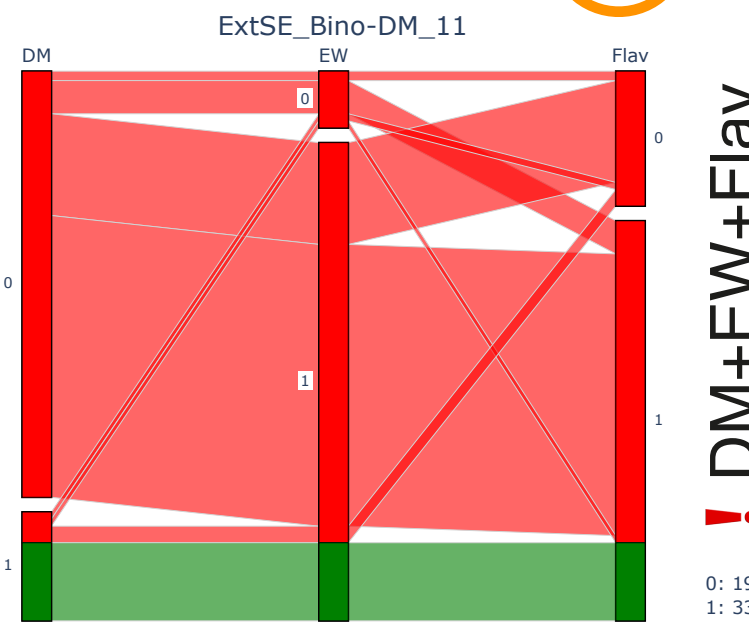
ATLAS is a strong cut. Study separately.



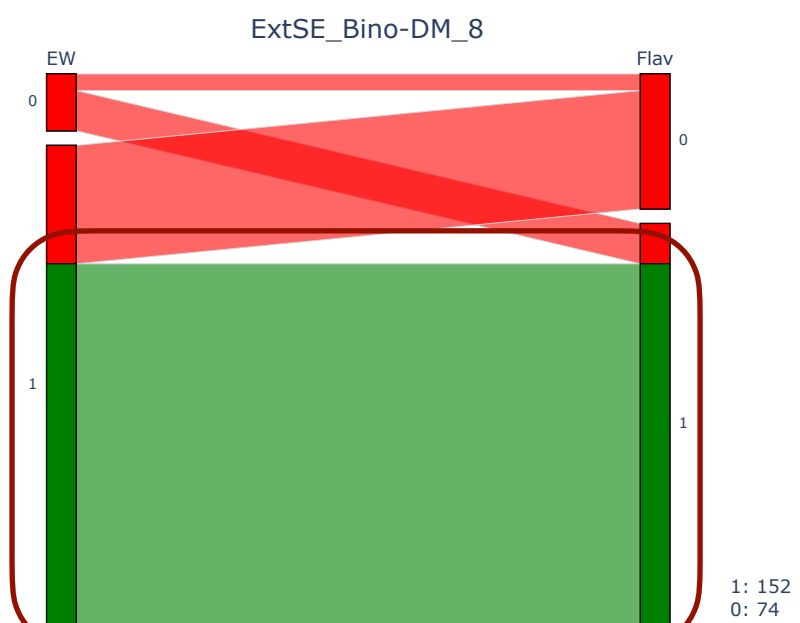
- SE models pass DM, EW and Flav constraints, but are killed by other ATLAS analyses.
- DM somewhat orthogonal to ATLAS

5

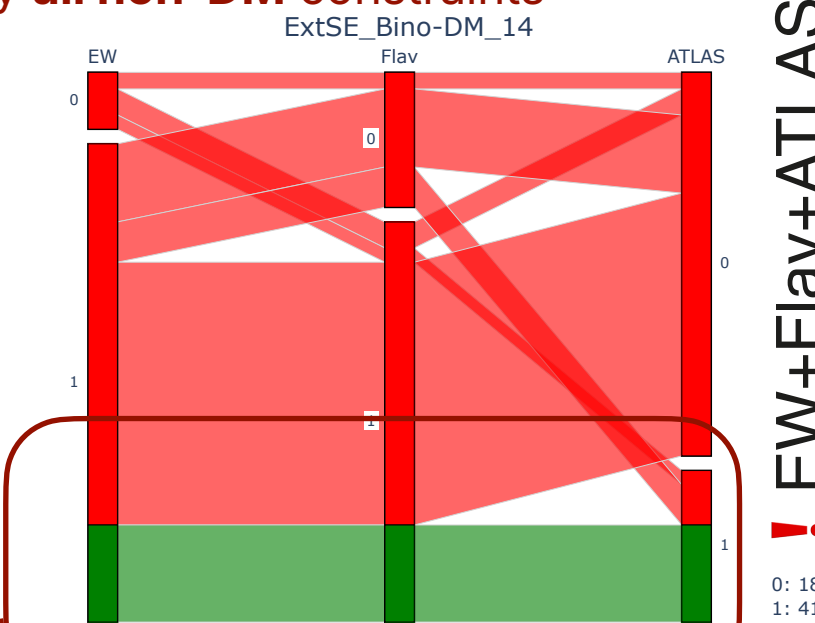
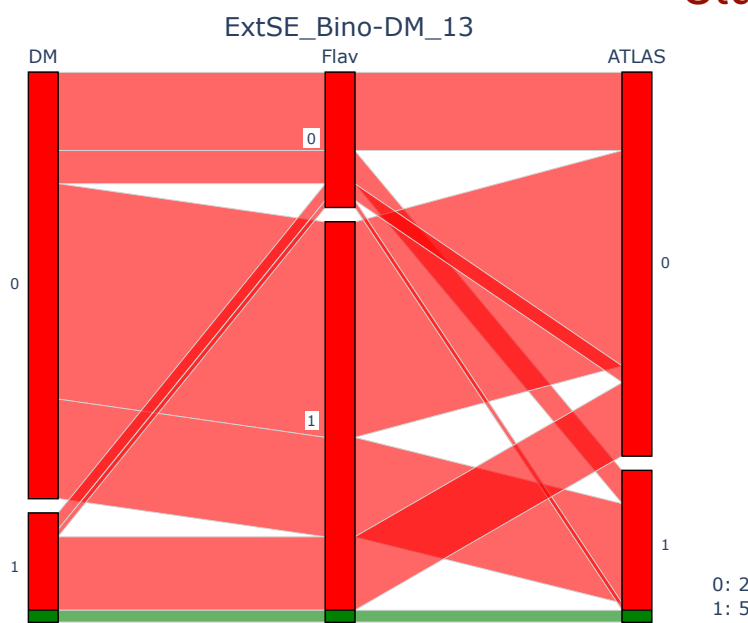
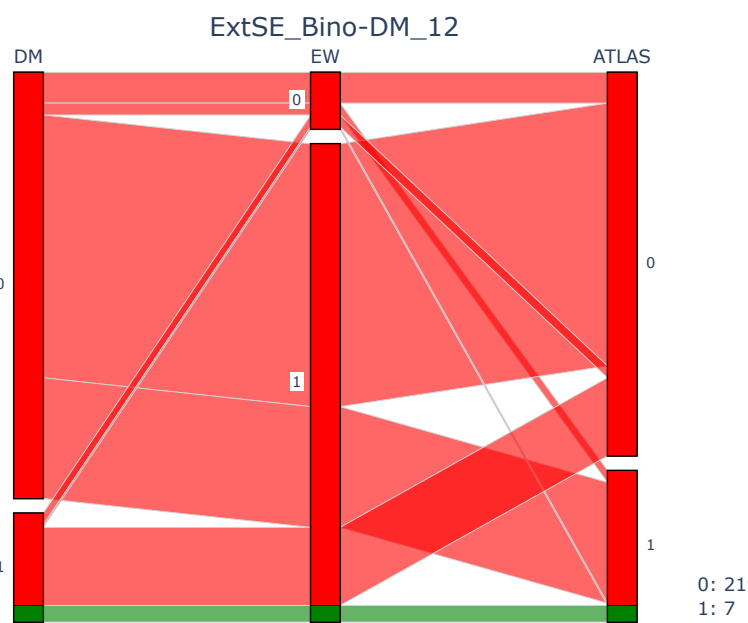
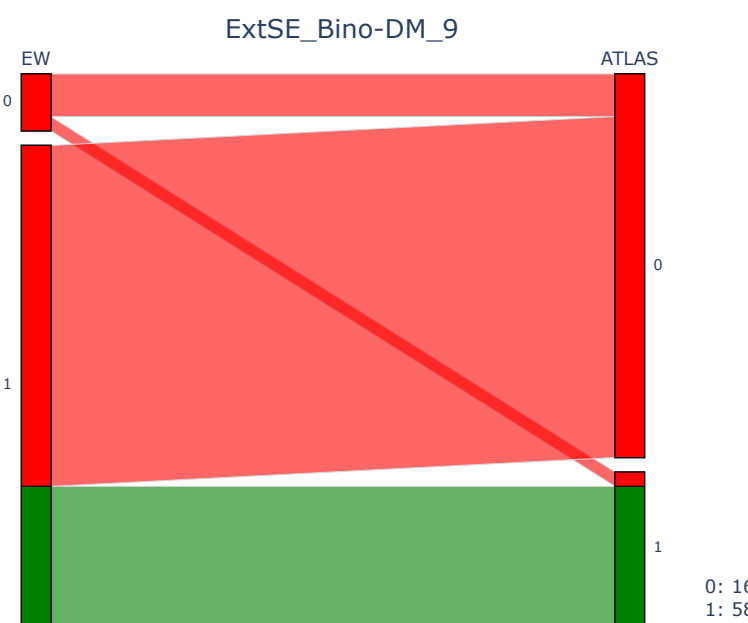
Study all non-ATLAS constraints



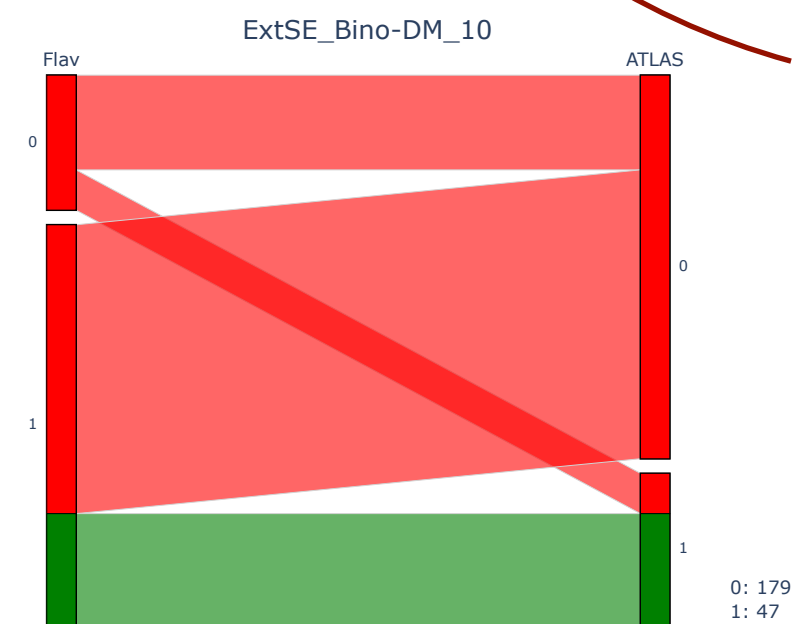
! DM+EW+Flav



EW and Flav not strong cuts, can study surviving models together, separate from ATLAS.



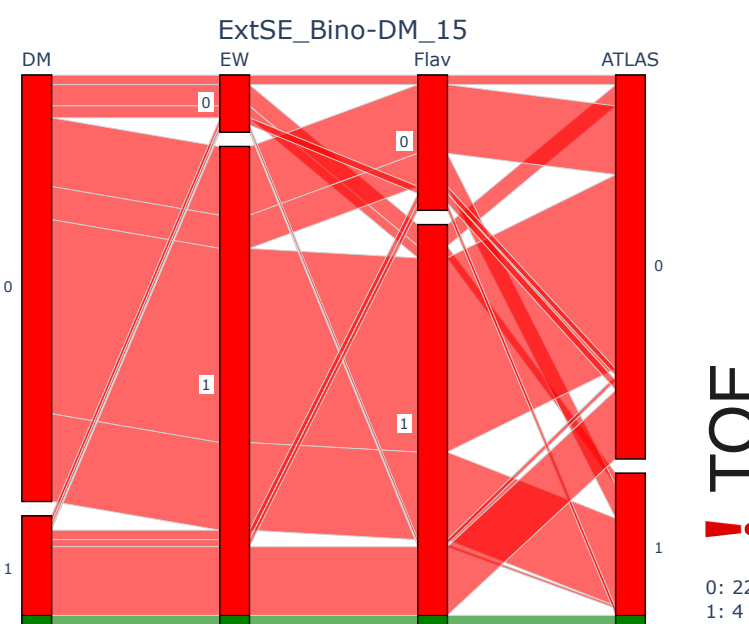
! EW+Flav+ATLAS



1

7

Study all constraints



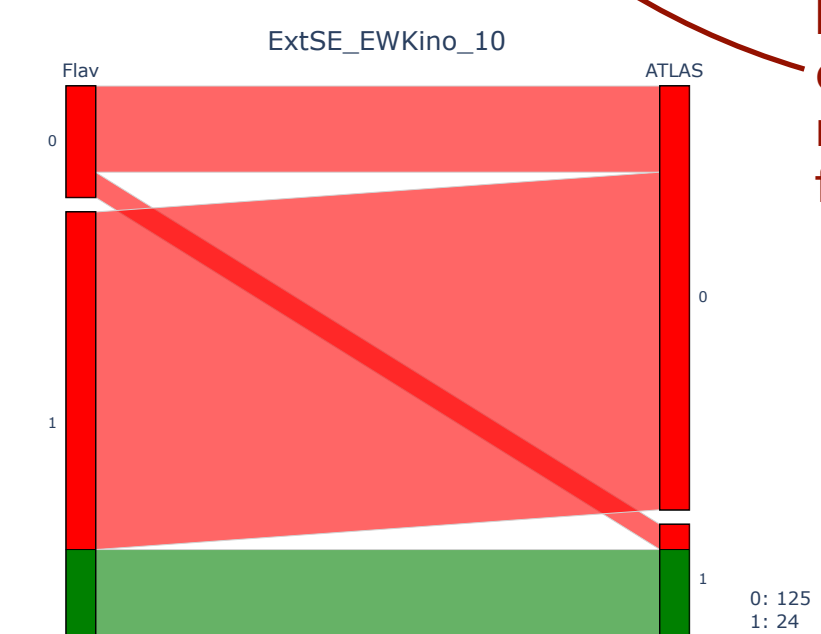
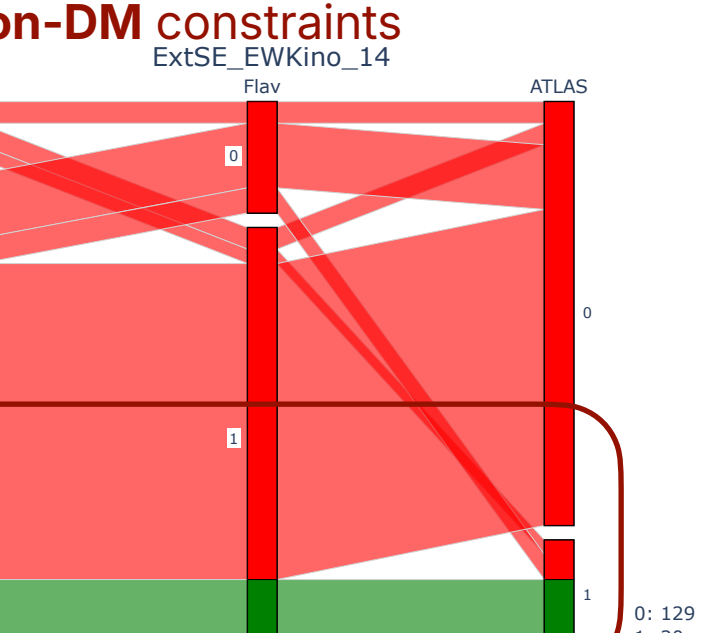
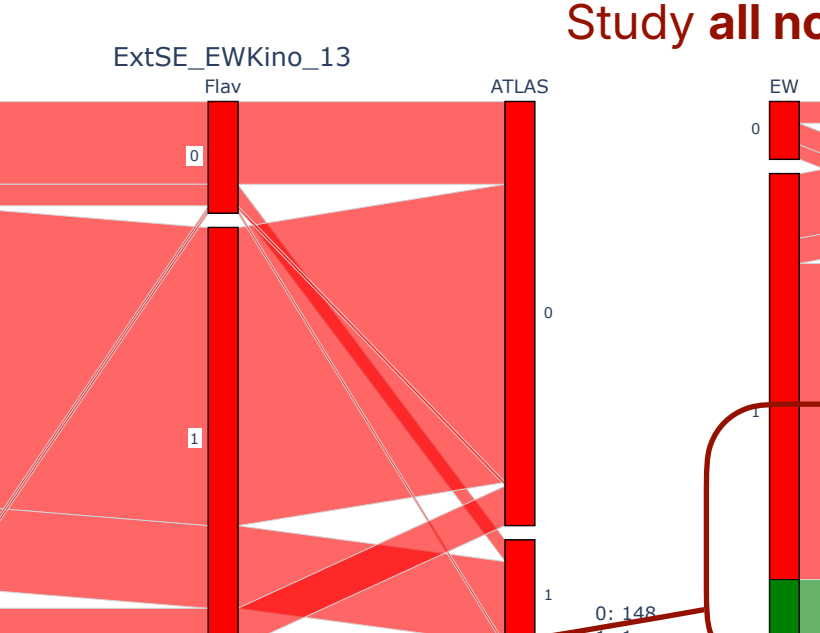
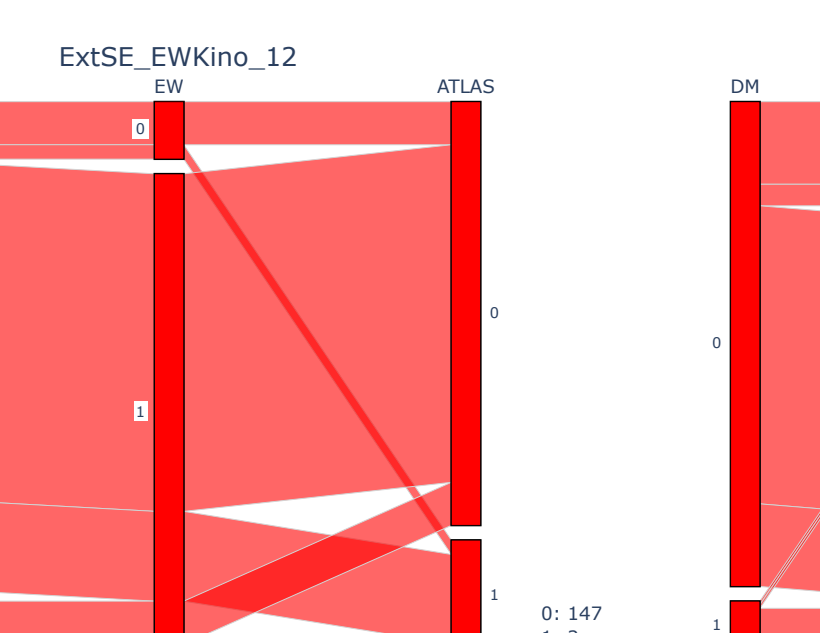
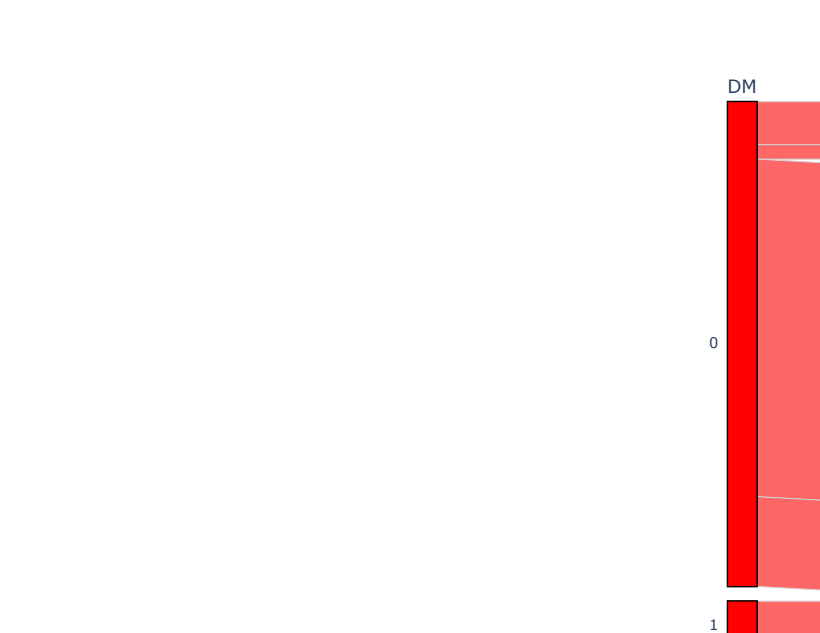
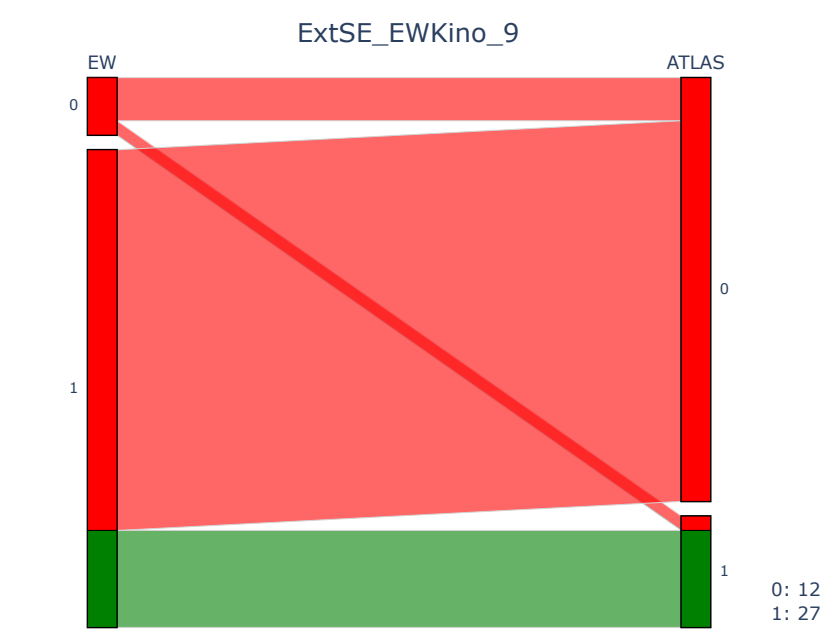
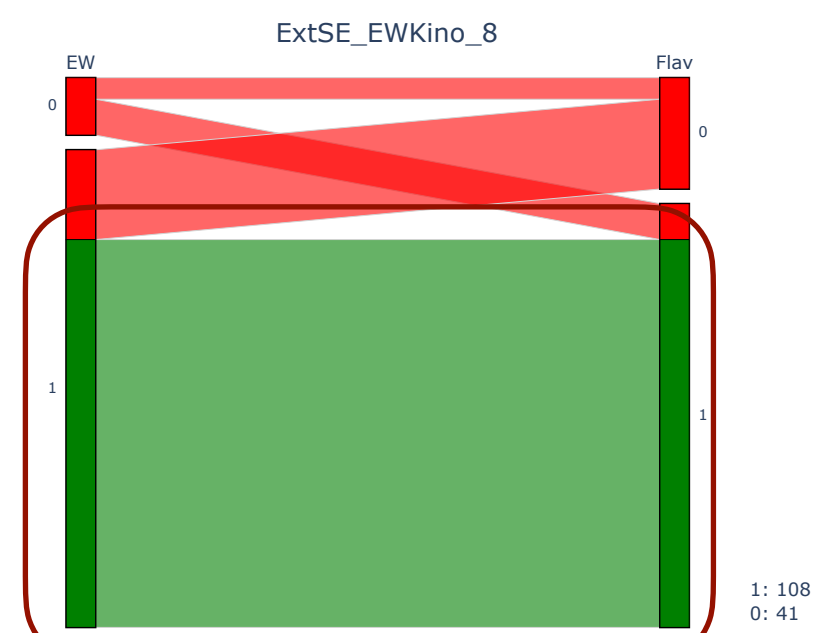
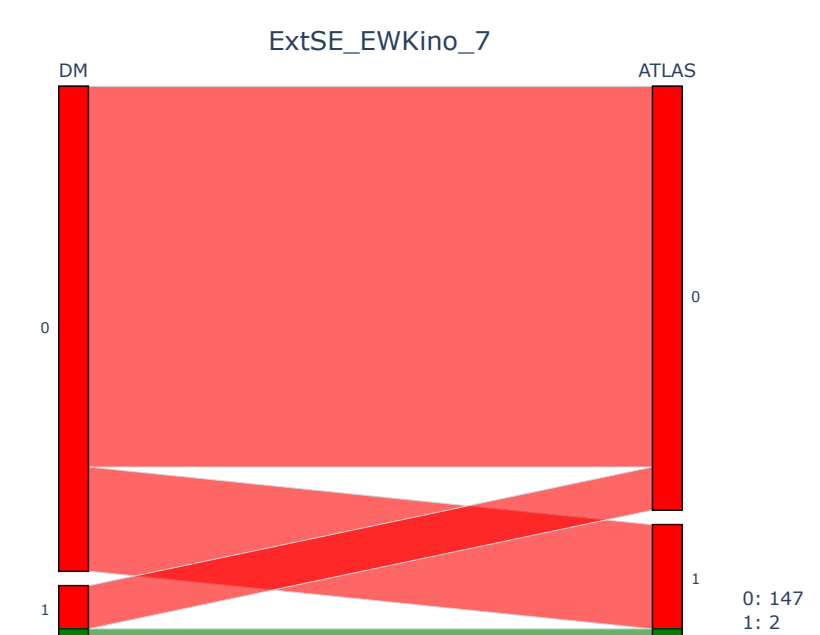
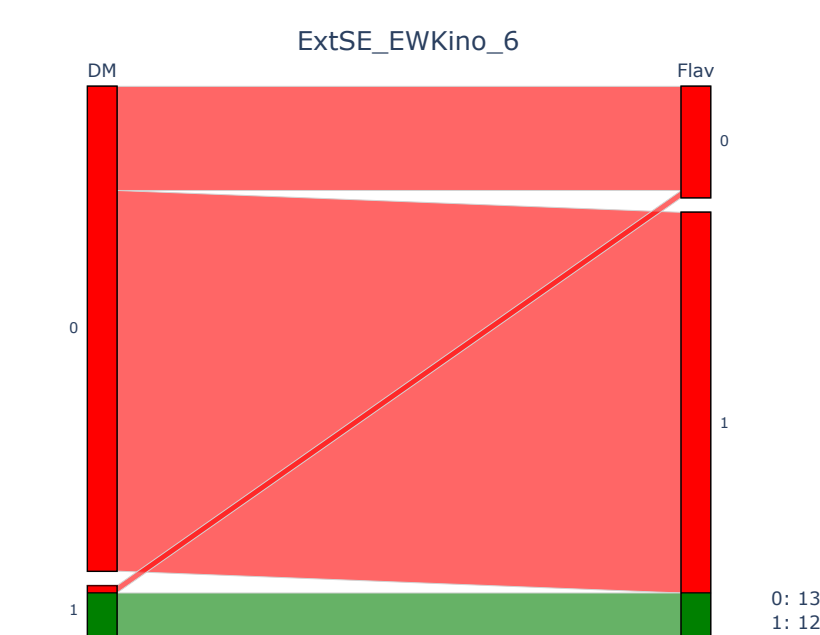
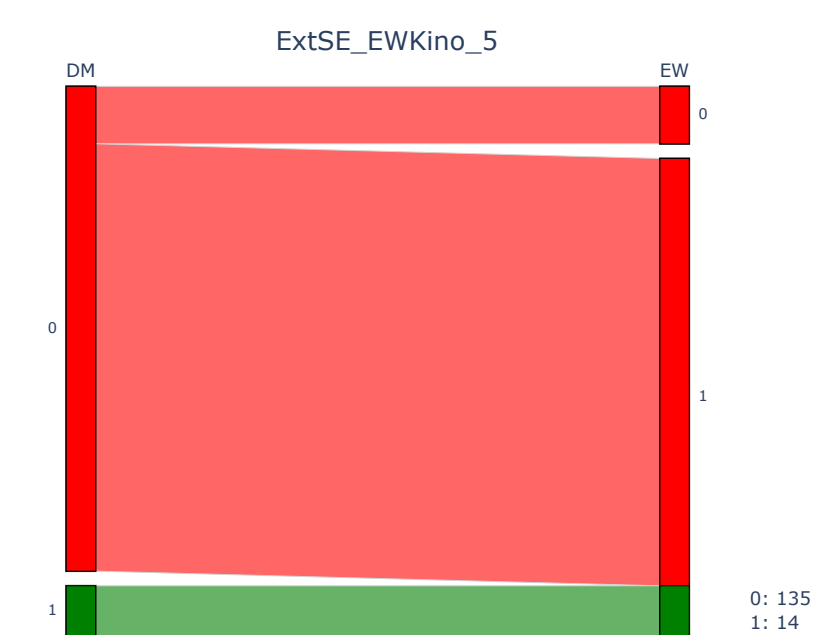
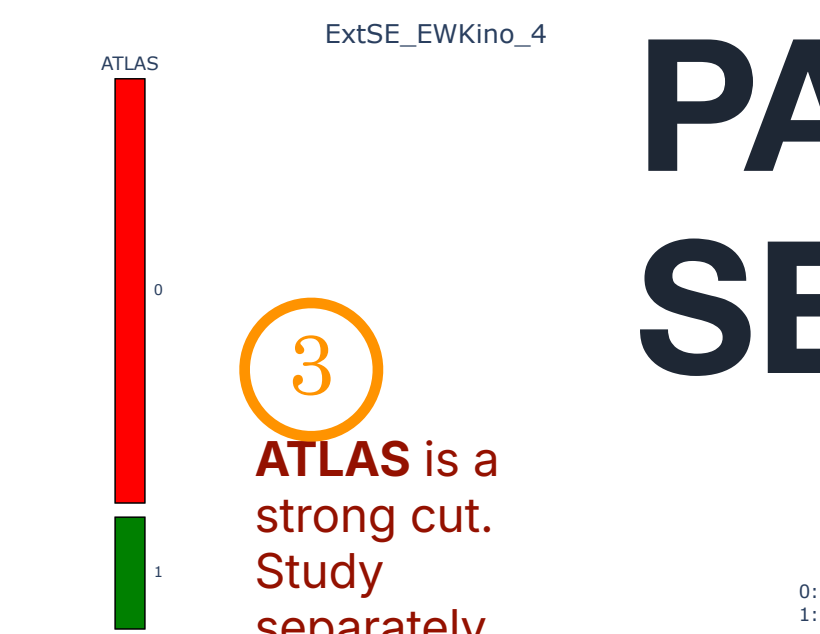
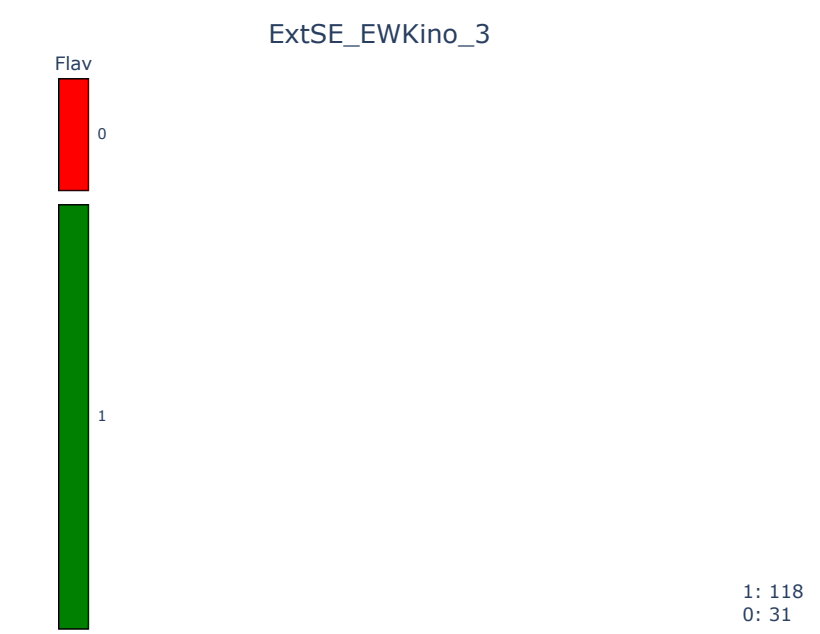
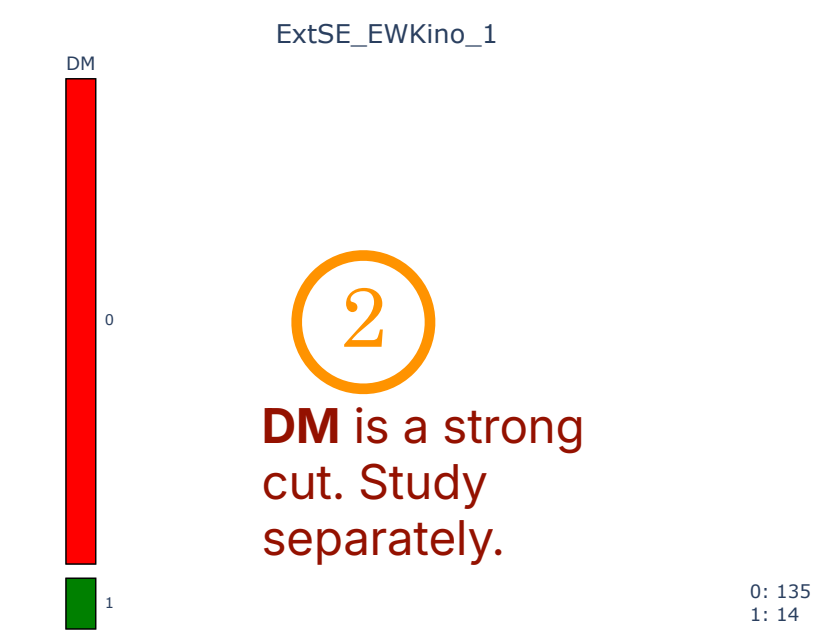
! TOE



# PARALLEL SETS

## SE: EWKINO

SE = True



EW and Flav not strong cuts, can study surviving models together, separate from ATLAS.

- By eye:
- 1 EW & Flav
  - 2 DM
  - 3 SurvATLAS
  - 4 EW
  - 5 DM & EW & Flav
  - 6 EW & Flav & SurvATLAS
  - 7 All

7 Study all constraints

5 Study all non-ATLAS constraints

6 Study all non-DM constraints

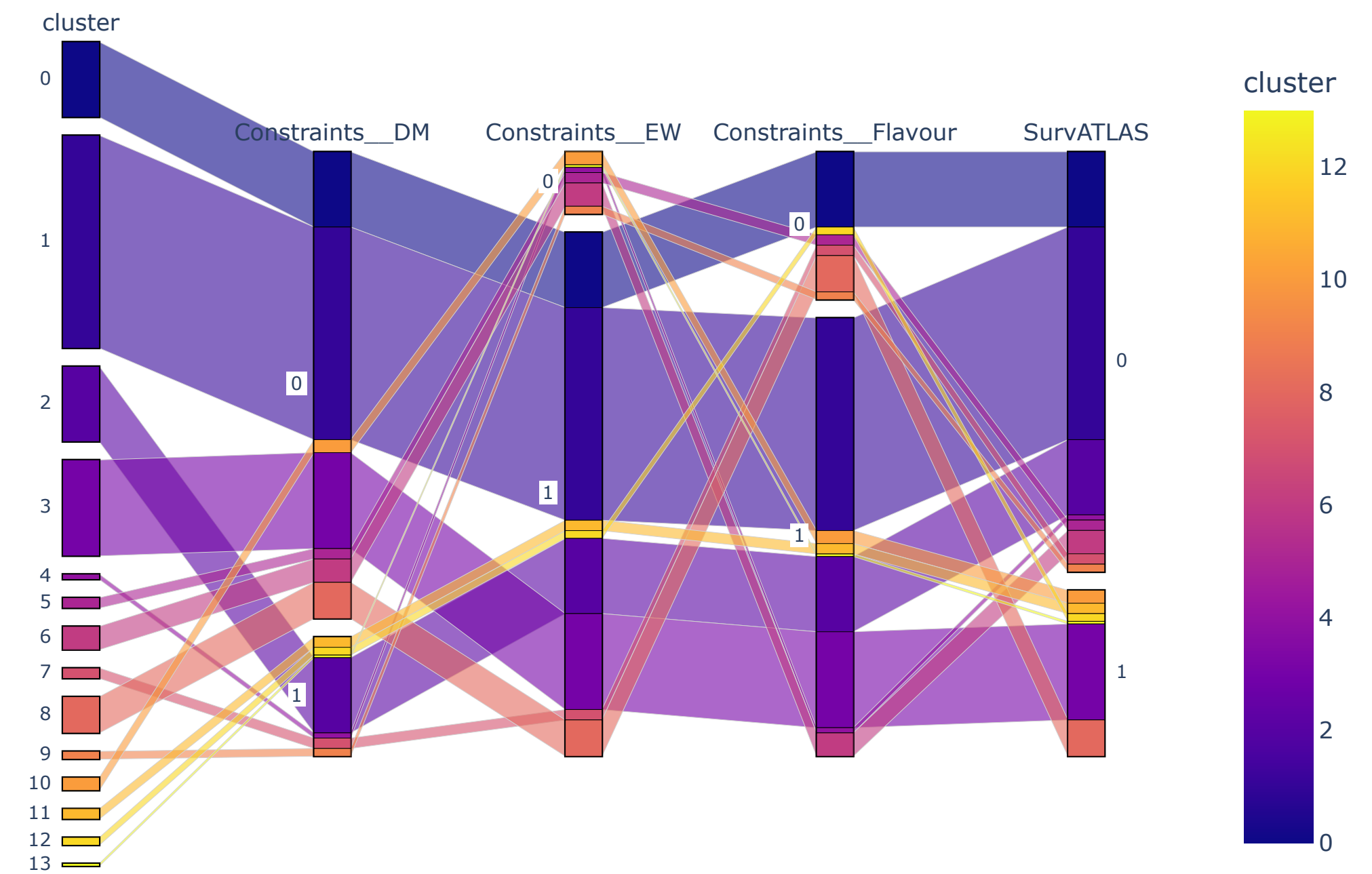
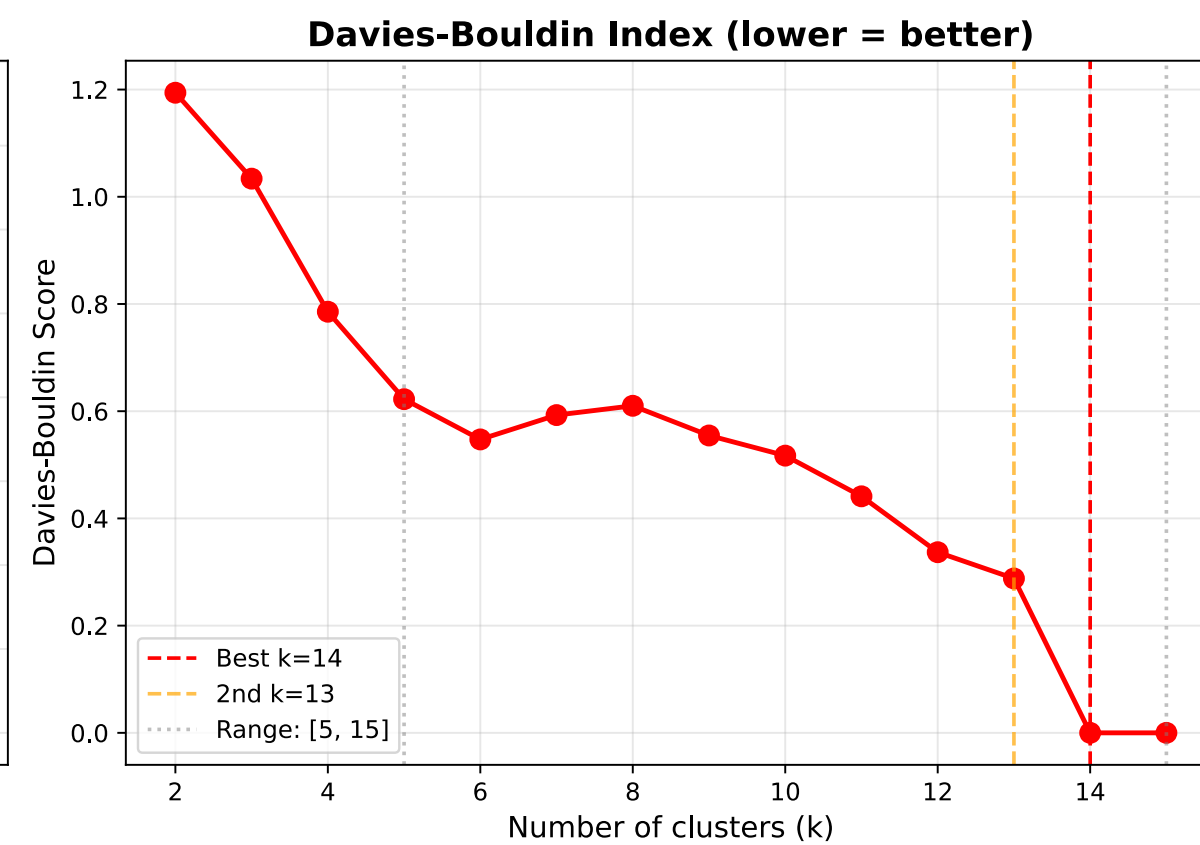
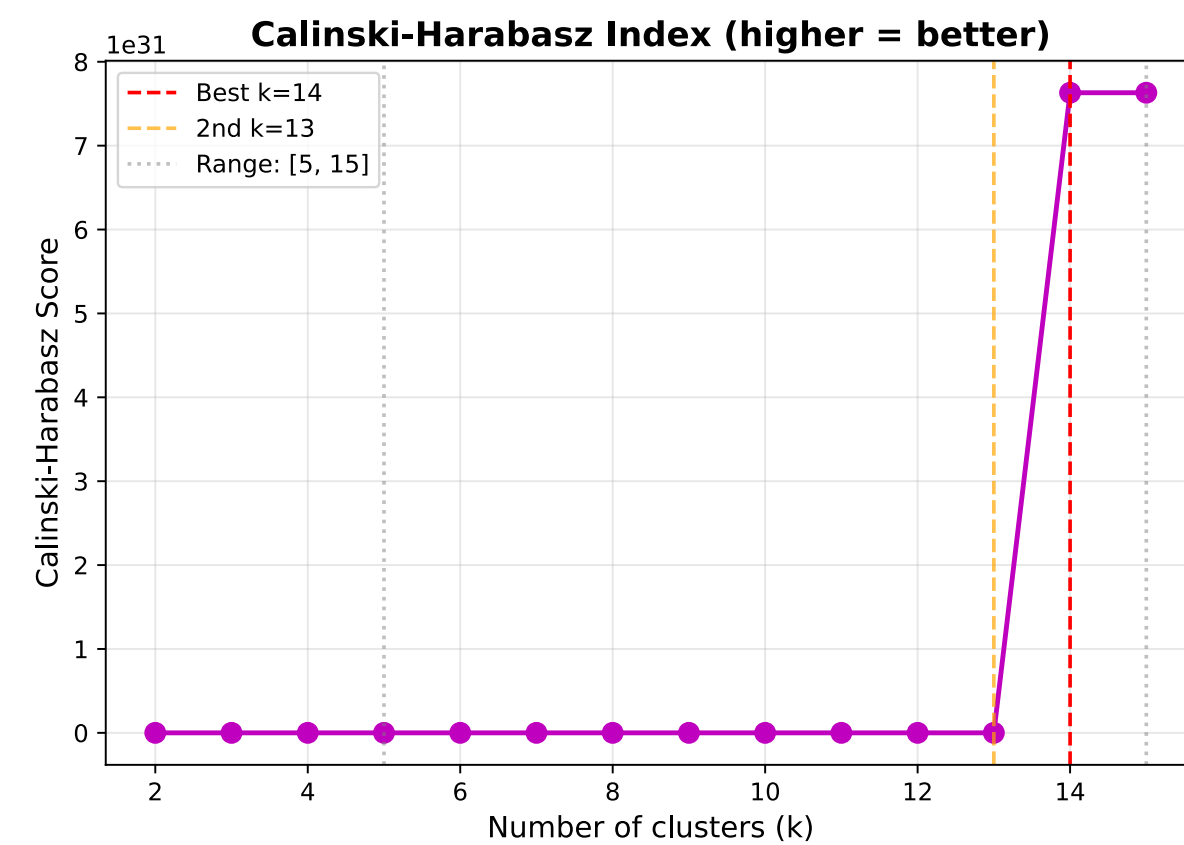
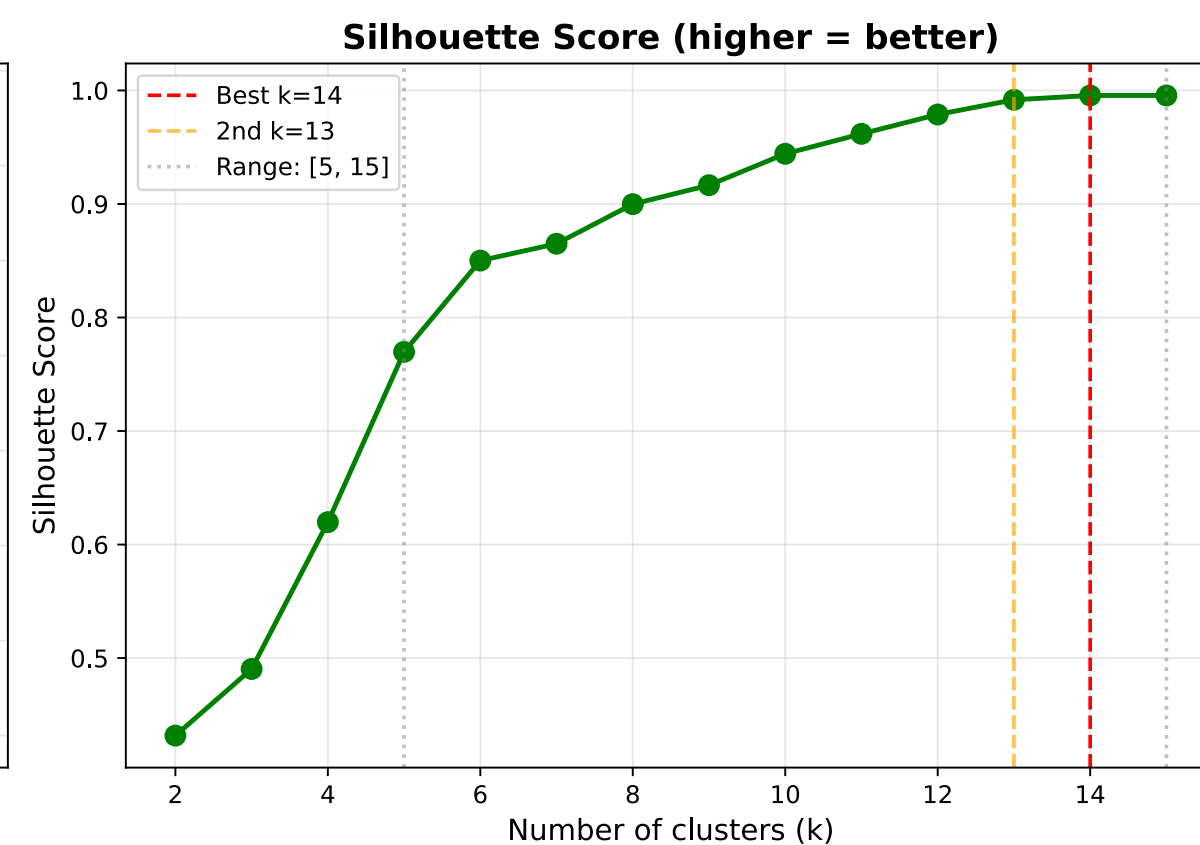
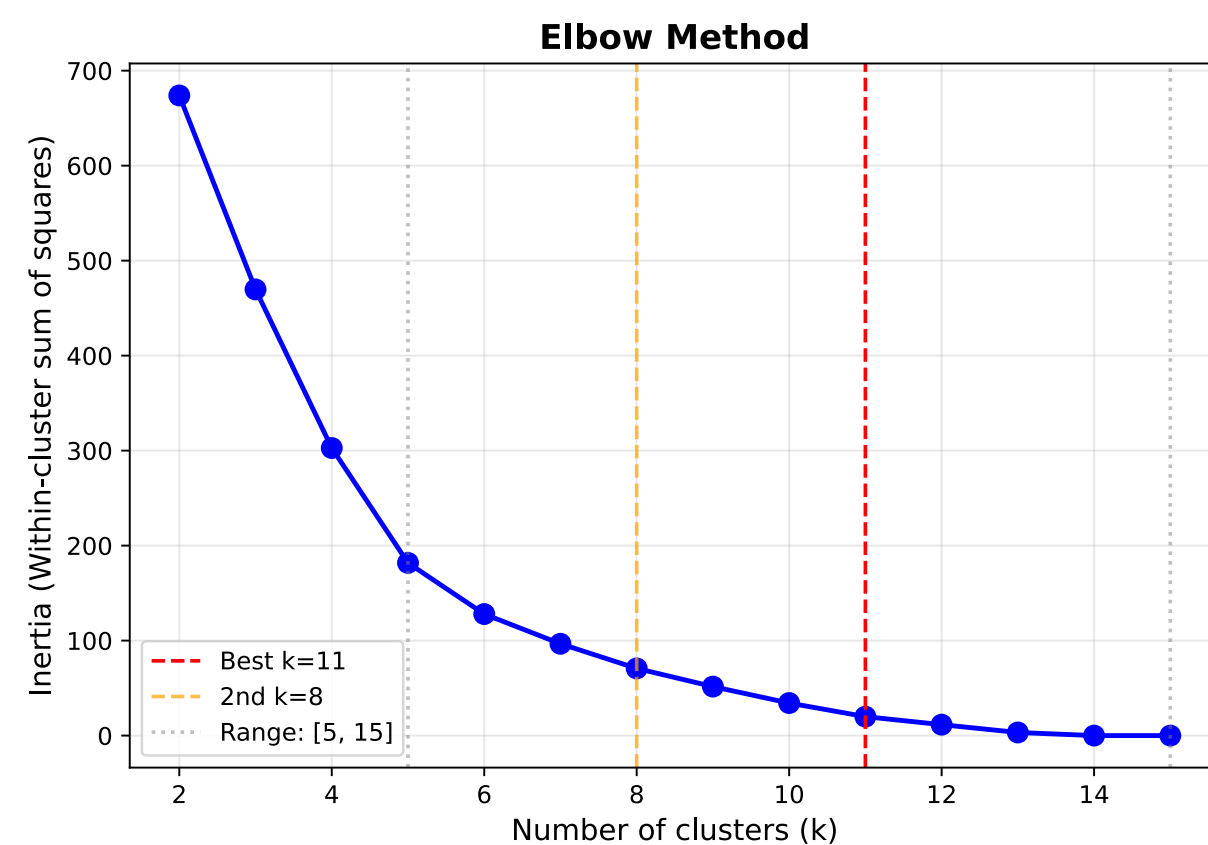
# CLUSTERING OPTIMAL, 14

- Elbow: point of maximum curvature in inertia decrease
- Silhouette: measures how similar objects are to their own cluster vs others
- Calinski-Harabasz: ratio of between-cluster to within-cluster dispersion
- Davies-Bouldin: average similarity between each cluster and its most similar one

kMeans(n\_clusters=k, random\_state=0, n\_init=20)

Starting from total number of combinations and extracting the optimal number of clusters from 4 tests.

Parallel sets: SE clusters with constraint dimensions (red-green gradient)



Max k from combinations: 15

Valid range: [5, 15]

Elbow Method suggests: k = 11

Silhouette Score suggests: k = 14 (score: 0.996)

Calinski-Harabasz suggests: k = 14 (score: 7e31)

Davies-Bouldin suggests: k = 14 (score: 0.000)



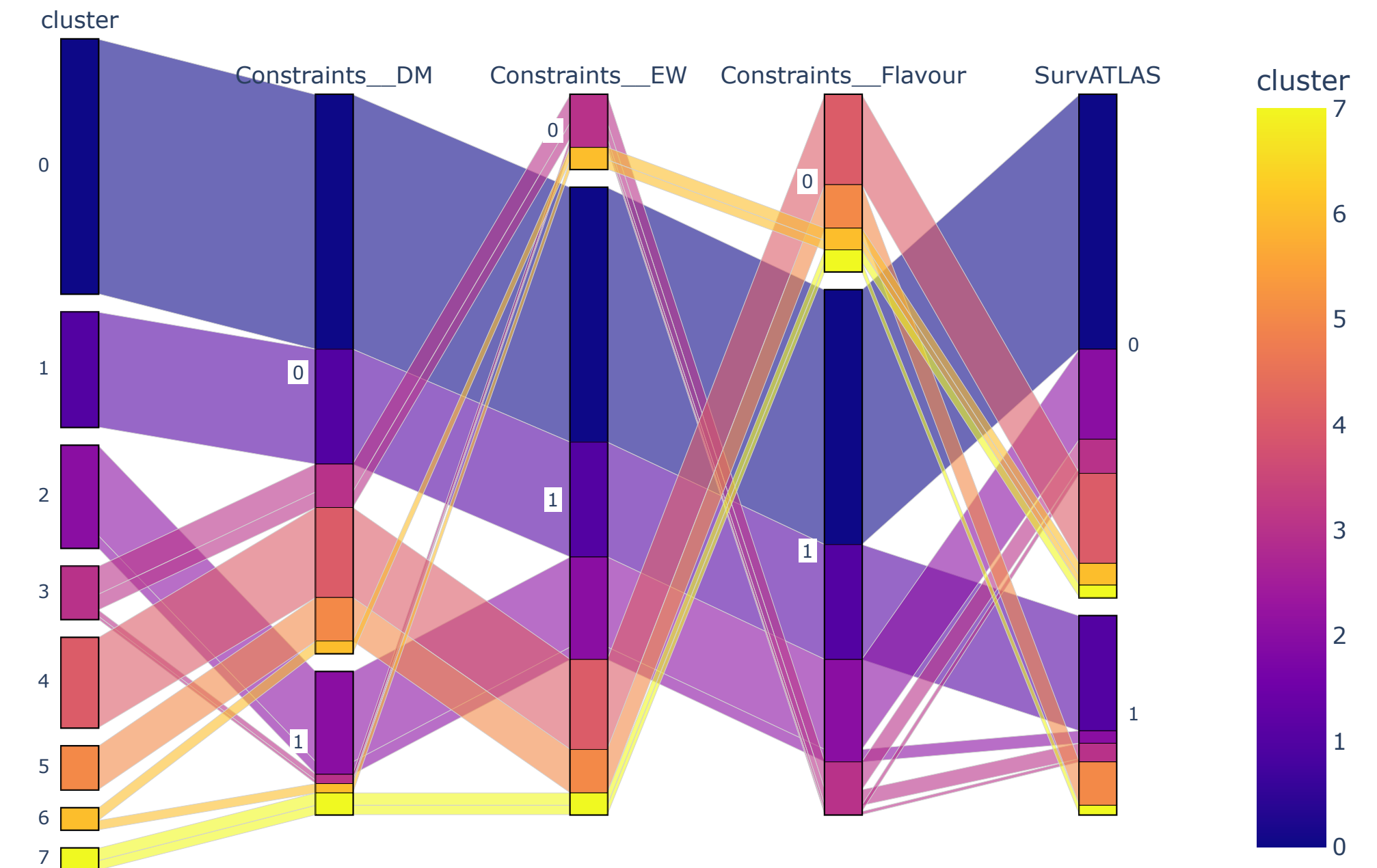
# CLUSTERING REDUCED, 8

kMeans(n\_clusters=k, random\_state=0, n\_init=20)

Scanning (1/3-1) total number of combinations and extracting the second best number of clusters from left using the 4 tests.

Parallel sets: SE clusters with constraint dimensions (red-green gradient)

- Reduce the number of clusters as second best, see if similar to by eye study.
- ==> Clustering takes pass, fail, ignore, while by eye we focused on pass and ignore. Reconsider the fail option in by eye clustering.



Max k from combinations: 15

Valid range: [5, 15]

Elbow Method suggests: k = 8

Silhouette Score suggests: k = 13 (score: 0.992)

Calinski-Harabasz suggests: k = 13 (score: 4868.1)

Davies-Bouldin suggests: k = 13 (score: 0.288)

# RL REDUCED: CLUSTER MERGING 10

Aim: Find me interesting models:

NOW: Use RL to define clusters of models passing external and ALTAS constraints

Next: Use RL to define clusters using in addition mass spectrum, increase of xsec, topology, mixing, survived analyses

Test example:

STATE: Cluster (similarity bin and feature distance bin)

REWARD: Merge similar clusters by survived constraints and ATLAS exclusion

ACTION: with 20% probability it explores and improves understanding

Parallel sets: SE clusters with constraint dimensions (red-green gradient)

