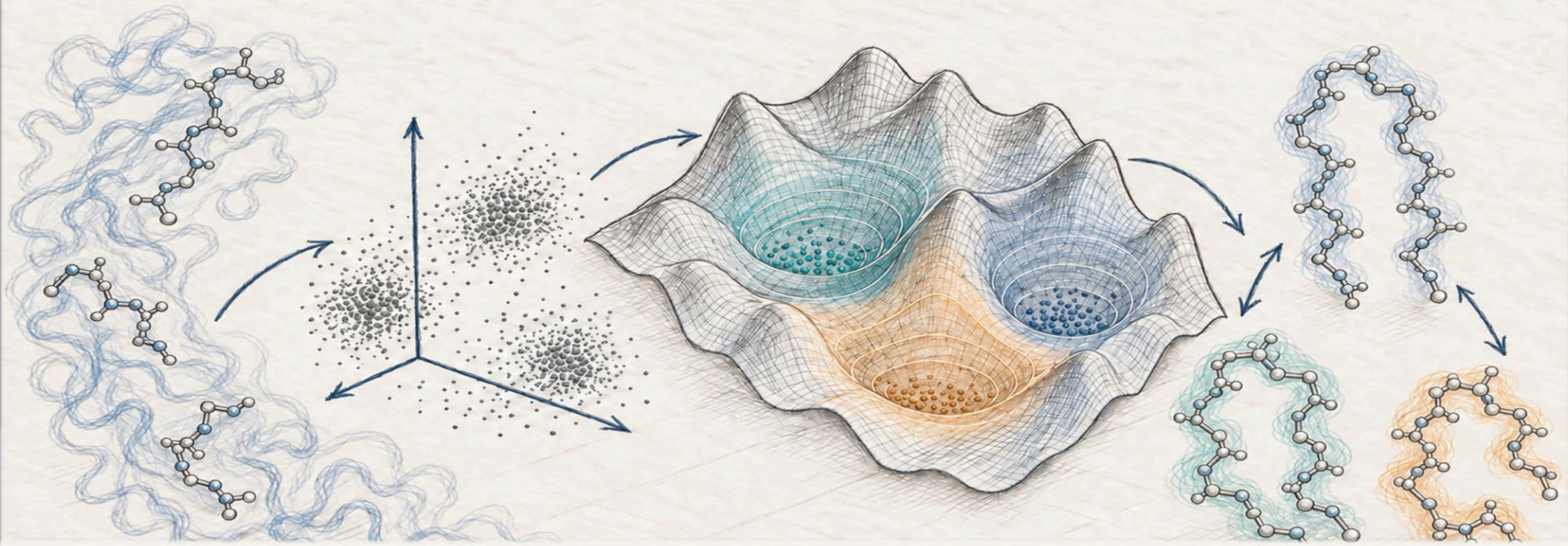




## Automating the multidimensional clustering of data derived from a Principal Component Analysis of molecular dynamics trajectories using HDBSCAN

**Μελάς Αναστάσιος**

Επιβλέπων Καθηγητής: Νικόλαος Γλυκός

**Μεταπτυχιακό στην «Εφαρμοσμένη Βιοπληροφορική και Ανάλυση Δεδομένων»**

Ακαδημαϊκό Έτος: 2025 - 2026

# Το πρόβλημα της ανάλυσης

Οι **προσομοιώσεις μοριακής δυναμικής** (molecular dynamics (MD) simulations) καταγράφουν την εξέλιξη ενός μοριακού συστήματος στον χρόνο, παράγοντας διαδοχικά στιγμιότυπα (frames) των ατομικών του θέσεων. Κάθε στιγμιότυπο περιγράφει μία συγκεκριμένη διαμόρφωση του μορίου.

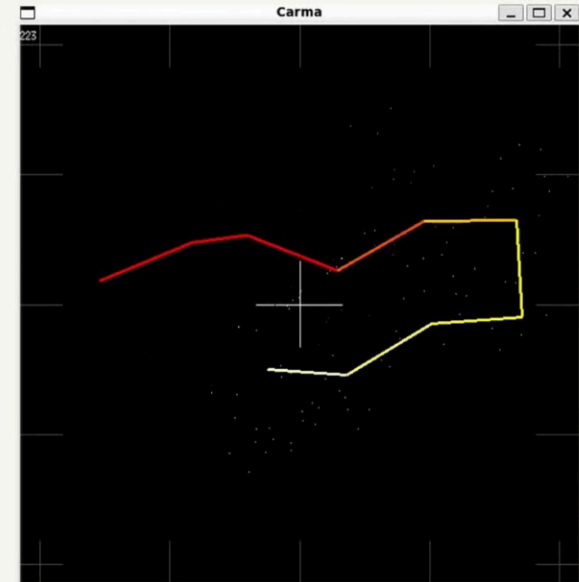
- Η άμεση επιθεώρηση κάθε στιγμιότυπου δεν είναι πρακτικά εφικτή.

Μία τροχιά μοριακής δυναμικής με  $T$  αποθηκευμένα στιγμιότυπα και  $N$  άτομα μπορεί να αναπαρασταθεί ως πίνακας συντεταγμένων σε χώρο  $\mathbb{R}^{3N \times T}$ .

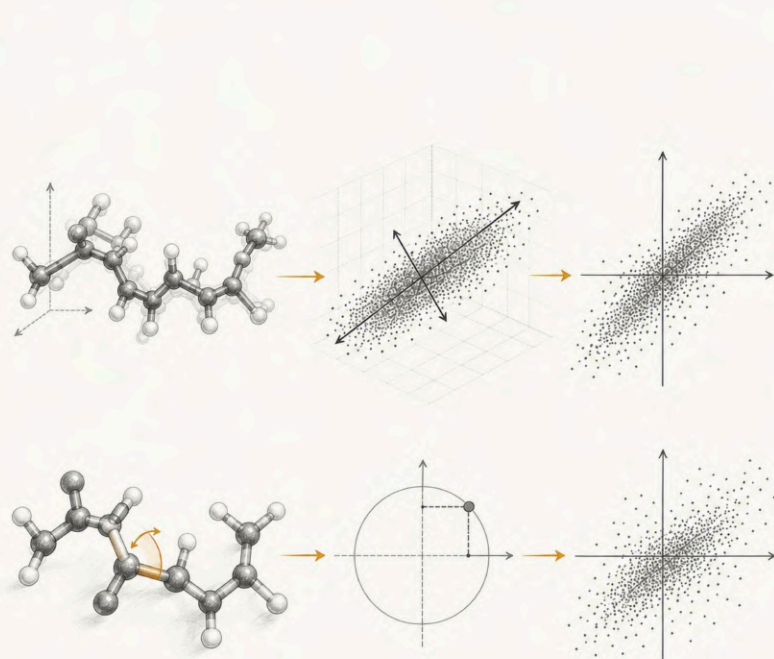
Χρειαζόμαστε έναν τρόπο να οργανώσουμε παρόμοιες διαμορφώσεις και να εξετάσουμε τη δομική τους ποικιλότητα, μεταβαίνοντας από ένα σύνολο με εκατομμύρια στιγμιότυπα σε ερμηνεύσιμες διαμορφώσεις.

Κεντρικό ερώτημα:

Πώς συνδέουμε την πολυδιάστατη ανάλυση με τις πραγματικές μοριακές δομές;



# Από τις δομές στις συντεταγμένες προβολής



Για να συγκρίνουμε τα στιγμιότυπα μιας τροχιάς, χρειαζόμαστε μια αναπαράσταση που συνοψίζει τη διακύμανση των μοριακών διαμορφώσεων. Η **Ανάλυση Κύριων Συνιστωσών** (PCA) δημιουργεί νέους άξονες, οι οποίοι περιγράφουν διαδοχικά τη μεγαλύτερη δυνατή διακύμανση των δεδομένων.

Στην **καρτεσιανή PCA** χρησιμοποιούνται οι ατομικές συντεταγμένες, αφού προηγηθεί δομική ευθυγράμμιση για την αφαίρεση της συνολικής μετατόπισης και περιστροφής. Η ερμηνεία των συνιστωσών μπορεί, ωστόσο, να επηρεάζεται από την ευθυγράμμιση όταν εμφανίζονται μεγάλες διαμορφωτικές αλλαγές.

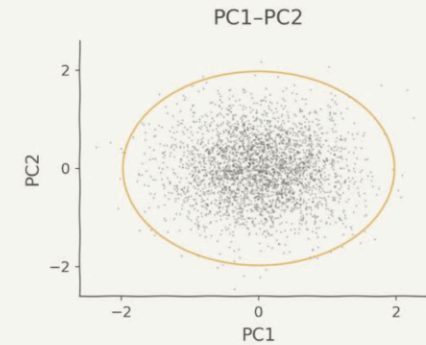
Η **dPCA** (dihedral PCA - διέδρων γωνιών) χρησιμοποιεί εσωτερικές συντεταγμένες, όπως τις διέδρες γωνίες του πεπτιδικού σκελετού. Κάθε γωνία αναπαρίσταται με το ημίτονο και το συνημίτονό της, ώστε να αποφεύγεται η τεχνητή ασυνέχεια μεταξύ  $-180^\circ$  και  $+180^\circ$  (Altis et al., 2008).

Κάθε στιγμιότυπο περιγράφεται πλέον ως **σημείο στον χώρο των διατηρούμενων κύριων συνιστωσών**, στον οποίο πραγματοποιείται η επόμενη φάση της ομαδοποίησης.

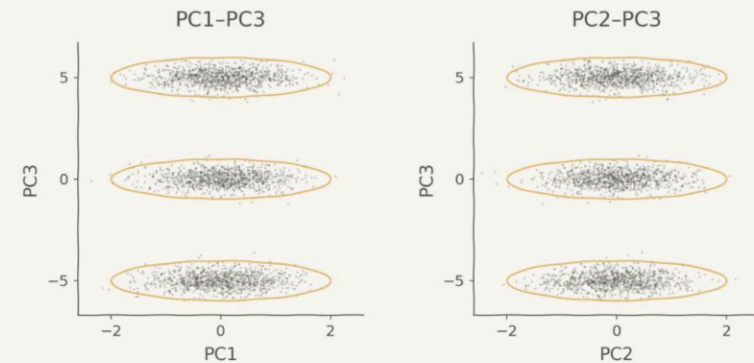
# Γιατί περισσότερες από δύο διαστάσεις;

Η δισδιάστατη εικόνα δεν περιέχει όλη την πληροφορία.

- Η προβολή στις δύο πρώτες κύριες συνιστώσες μπορεί να συγχωνεύσει οπτικά διακριτές διαμορφωτικές περιοχές, αποκρύπτοντας διαφορές που περιγράφονται από πρόσθετες συνιστώσες.
- Το *clusterMD* εφαρμόζει την ομαδοποίηση στον επιλεγμένο πολυδιάστατο χώρο PCA/dPCA και χρησιμοποιεί τις δισδιάστατες προβολές για οπτικοποίηση, με την επάρκεια της διατηρούμενης διάστασης να απαιτεί αξιολόγηση.



Η PC3 αποκαλύπτει τον διαχωρισμό



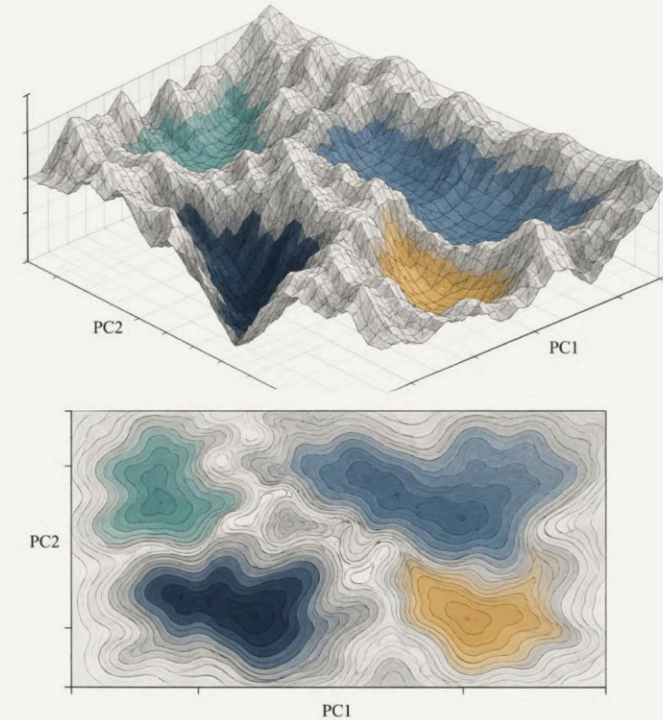
Σχηματικό παράδειγμα με τα ίδια συνθετικά σημεία

# HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise)

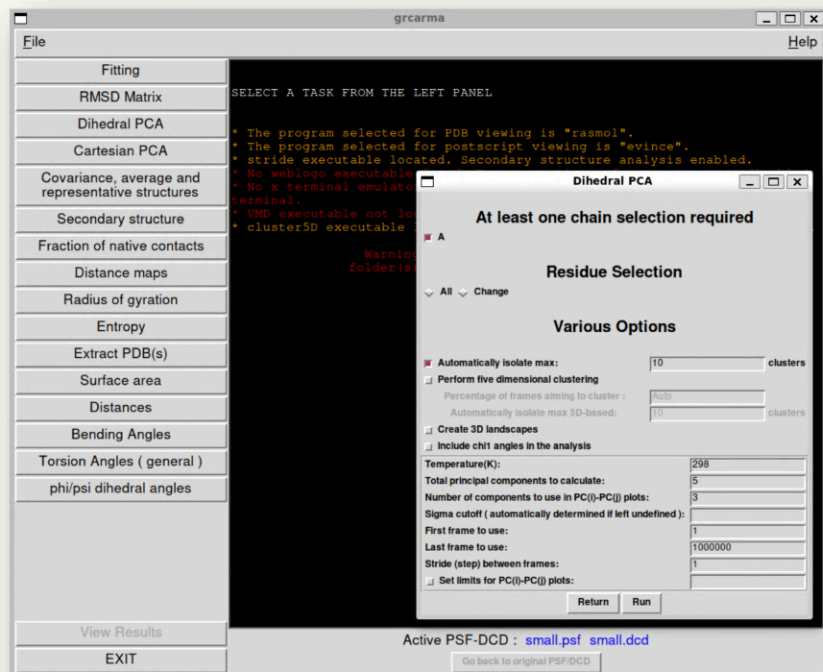
**HDBSCAN:** ιεραρχική ομαδοποίηση με βάση την πυκνότητα για εφαρμογές με θόρυβο.

Ο HDBSCAN κατασκευάζει μια ιεραρχία ομάδων εξετάζοντας τη δομή πυκνότητας των δεδομένων και επιλέγει ομάδες χωρίς να απαιτεί προκαθορισμένο αριθμό ή σφαιρικό σχήμα.

Στο *clusterMD* εφαρμόζεται στον διατηρούμενο χώρο PCA/dPCA, οργανώνοντας τις διαμορφώσεις σε ομάδες και διατηρώντας ως μη ανατεθειμένα τα σημεία με ετικέτα -1. Η βιολογική σημασία των ομάδων απαιτεί πρόσθετη δομική και κινητική αξιολόγηση.



# carma · grcarma



Το **carma**, που αναπτύχθηκε από τον Νικόλαο Γλύκο, είναι λογισμικό ανάλυσης τροχιών μοριακής δυναμικής. Στην παρούσα εργασία, η ανάλυση PCA/dPCA μέσω carma/grcarma **παράγει τις συντεταγμένες προβολής** που χρησιμοποιούνται ως είσοδος για την ομαδοποίηση.

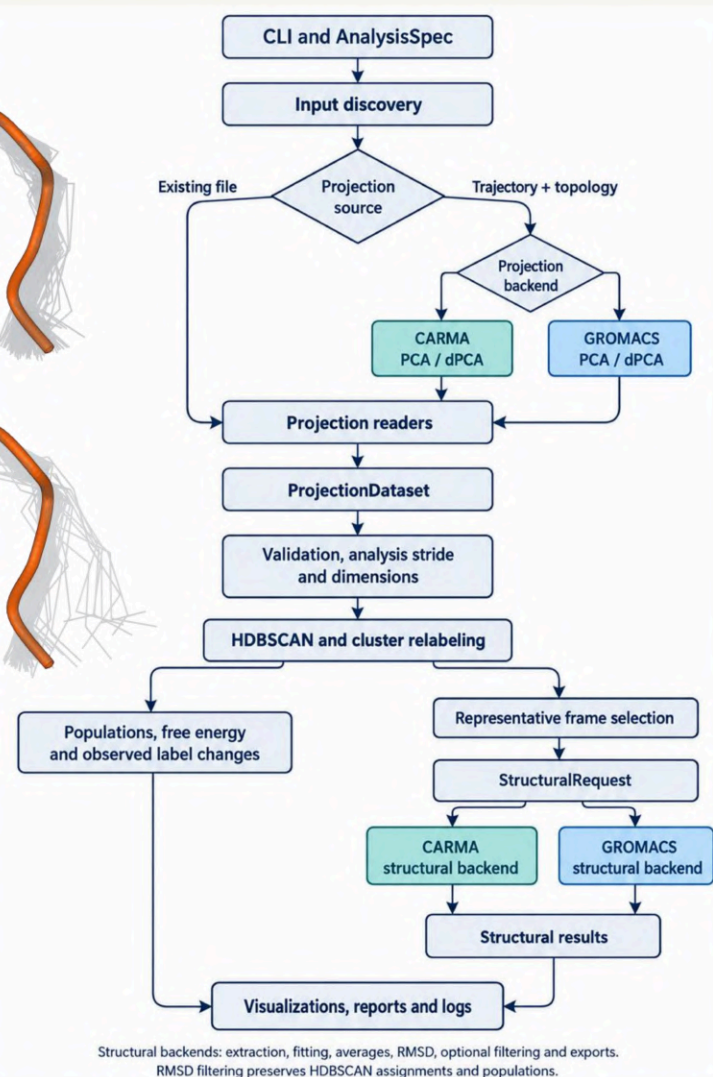
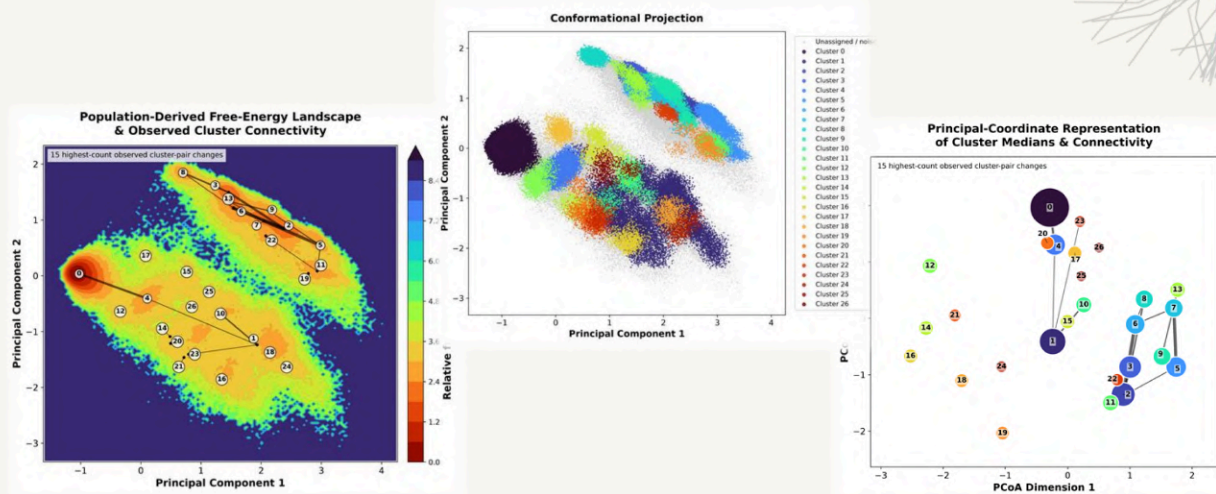
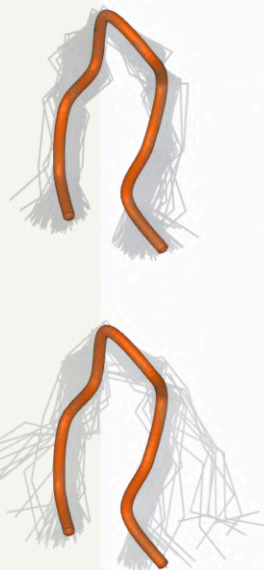
Μετά την ανάθεση των στιγμιotypών σε ομάδες από το clusterMD, το **carma** χρησιμοποιείται για την **εξαγωγή των αντίστοιχων διαμορφώσεων**, τη δομική υπέρθεση και τον υπολογισμό μέσων δομών και RMSD, συνδέοντας τα αποτελέσματα της ομαδοποίησης με την αρχική μοριακή τροχιά.

*Glykos, 2006 · Koukos & Glykos, 2013.*

# clusterMD

Το **clusterMD** αυτοματοποιεί την ανάλυση συντεταγμένων PCA/dPCA, συνδέοντας την πολυδιάστατη ομαδοποίηση με τις πληθυσμιακές κατανομές, την εξαγωγή μοριακών δομών και την παραγωγή γραφημάτων και αναφορών.

Με αυτόν τον τρόπο, οι αριθμητικές αναθέσεις μετατρέπονται σε οργανωμένα αποτελέσματα που μπορούν να εξεταστούν και να ερμηνευτούν δομικά.



# Υλοποίηση και εκτέλεση



```
usage: clusterMD [-h] [--size SIZE] [--samples SAMPLES] [--auto] [--cluster-only]
                [--method {eom,leaf}] [--dim DIM] [--rmsd RMSD] [--temp TEMP]
                [--step STEP] [--segid SEGID] [--top TOP]
                [--transitions [TRANSITIONS ...]] [--events EVENTS]

Automated HDBSCAN clustering, observed change mapping, and structural extraction
pipeline.

options:
  -h, --help            show this help message and exit
  --size SIZE            HDBSCAN min_cluster_size (Default: 500)
  --samples SAMPLES     HDBSCAN min_samples (Default: 50)
  --auto                Set min_cluster_size and min_samples using the frame-
                        count heuristic
  --cluster-only        Run ingestion and HDBSCAN clustering, export assignments,
                        then exit before downstream analysis.
  --method {eom,leaf}   HDBSCAN cluster-selection method (Default: eom)
  --dim DIM             Number of projected dimensions to analyze; 0 uses all
                        available dimensions
  --rmsd RMSD           Optional RMSD threshold for downstream structural
                        filtering in Angstrom; 0 disables filtering
  --temp TEMP           Temperature in Kelvin used for population-derived free-
                        energy estimates (Default: 298.0)
  --step STEP           Frame stride applied during projection-file ingestion
                        (Default: 1)
  --segid SEGID         Restrict structural extraction to the specified segment
                        identifier
  --top TOP             Number of observed changes to display in the diagnostic
                        table and plots (Default: 15. 0 disables displayed edges)
  --transitions [TRANSITIONS ...], --changes [TRANSITIONS ...]
                        Pass 0 args to view the most frequent observed changes,
                        or 2 cluster IDs to inspect a specific cluster pair.
  --events EVENTS       Maximum chronological change events to print in
                        diagnostic mode (Default: 3)
```

Το *clusterMD* υλοποιήθηκε ως ενιαίο πρόγραμμα γραμμής εντολών σε **Python**, χρησιμοποιώντας **NumPy** και **SciPy** για την αριθμητική επεξεργασία, **fast\_hdbscan** για την πολυδιάστατη ομαδοποίηση και **Matplotlib** για την οπτικοποίηση. Καλεί το εξωτερικό πρόγραμμα **carma** για την εξαγωγή και τη δομική επεξεργασία των μοριακών στιγμιотύπων.

Η εκτέλεση ξεκινά από τον κατάλογο της ανάλυσης, με **αυτόματη αναζήτηση των αρχείων προβολής και των αντίστοιχων DCD/PSF**. Ο χρήστης καθορίζει μέσω παραμέτρων τις διατηρούμενες διαστάσεις, το βήμα δειγματοληψίας και τις παραμέτρους του HDBSCAN.

```
clusterMD --dim 5 --step 1 \
          --size 500 --samples 50 --method leaf
```

# Παραδείγματα εκτέλεσης (1) - CLN025

```
(.venv) small >>> clusterMD --auto --rmsd 2 --method leaf
=== clusterMD | HDBSCAN Analysis of Molecular-Dynamics PCA Coordinates ===
[*] Locating fluctuations file...
[+] Input file: carma_results_26.05.2026_08.41.46/carma.dPCA.fluctuations.dat
[*] Analysis directory: carma_results_26.05.2026_08.41.46
Command line: clusterMD --auto --rmsd 2 --method leaf
[*] Output directory: clusterMD_results_11.10.2026_16.17.22.241
[*] Pre-allocating memory and loading data (Stride: 1)...
[*] Data ingested successfully: 1,000,000 frames (5D).
[*] Auto-scaling activated: min_cluster_size=500
```

## ANALYSIS SETUP

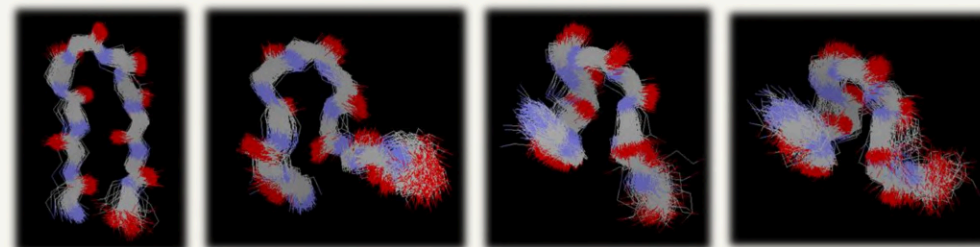
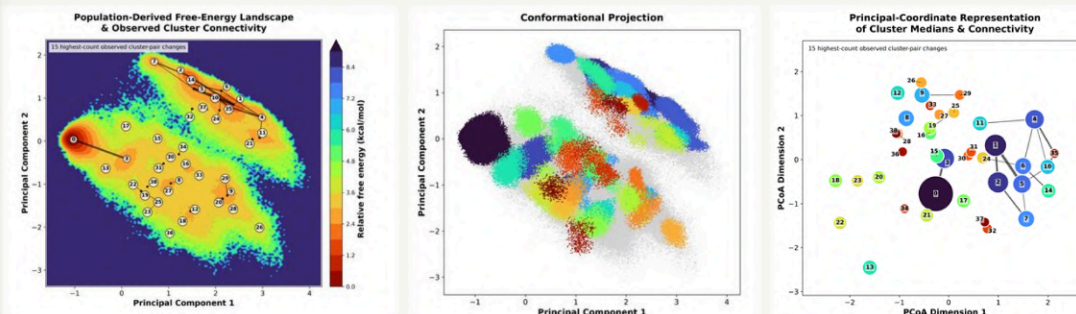
```
Projection file      : carma.dPCA.fluctuations.dat
Frames analyzed     : 1,000,000
Frame stride        : 1
Dimensions used      : 5
Temperature          : 298.0 K
HDBSCAN method      : leaf
min_cluster_size    : 500
min_samples          : 50
Auto-scaling         : Active
```

```
[*] Running HDBSCAN clustering...
[+] Identified 39 clusters in 37.12 seconds.
[*] Sorting cluster assignments by descending population size...
```

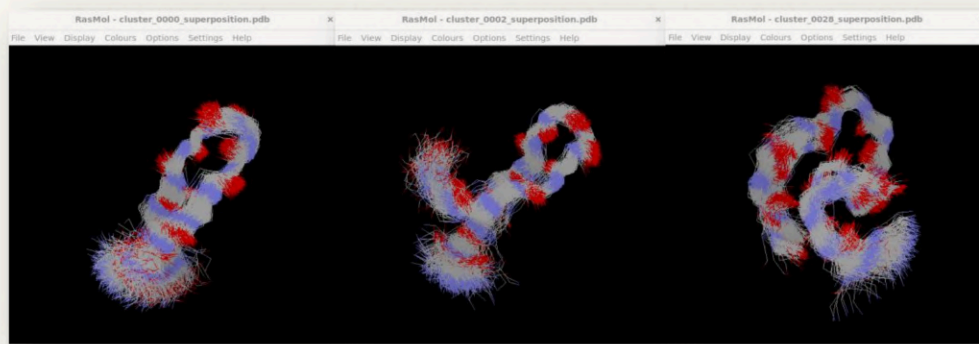
## CLUSTER POPULATIONS

| Cluster | Frames  | Population |
|---------|---------|------------|
| 0       | 613,098 | 61.3%      |
| 1       | 25,771  | 2.6%       |
| 2       | 19,504  | 2.0%       |
| 3       | 14,578  | 1.5%       |
| 4       | 11,678  | 1.2%       |
| 5       | 9,145   | 0.9%       |
| 6       | 5,908   | 0.6%       |

**CLN025:** συνθετικό πεπτίδιο **10 αμινοξέων**, με αλληλουχία **YYDPETGTWY**, που αποτελεί παραλλαγή του chignolin και σχηματίζει **β-φουρκέτα**, δηλαδή δύο αντιπαράλληλους β-κλώνους συνδεδεμένους μέσω μιας στροφής. Το μικρό του μέγεθος το καθιστά κατάλληλο για τη μελέτη της αναδίπλωσης και των αλλαγών διαμόρφωσης με μοριακές προσομοιώσεις. Στην παρούσα εργασία χρησιμοποιείται τροχιά **1.000.000 στιγμιотύπων**, με προβολή dPCA, για την παρουσίαση της πολυδιάστατης ομαδοποίησης και της σύνδεσης των ομάδων με τις αντίστοιχες μοριακές δομές.



# Παραδείγματα εκτέλεσης (2) - CDP-1



Το **CDP-1 (Cysteine Deleted Protegrin-1)** είναι ένα πεπτίδιο **14 αμινοξέων**, με αλληλουχία **RGGRLYRRRFVVGR**, που προκύπτει από την αφαίρεση των τεσσάρων κυστεϊνών της protegrin-1. Είναι πλούσιο σε αργινίνη και διατηρεί αντιβακτηριακή δράση παρά την απουσία των δύο δισουλφιδικών δεσμών του αρχικού πεπτιδίου. Πειραματικά, υιοθετεί συμπαγή δομή τύπου **β-φουρκέτας** **παρουσία λιποπολυσακχαρίτη (LPS)**, ενώ η διαμόρφωσή του εξαρτάται από το περιβάλλον. Στην παρούσα εργασία χρησιμοποιείται τροχιά περίπου **7,5 εκατομμυρίων στιγμιοτύπων**, για την εξέταση της εφαρμογής του *clusterMD* σε μεγάλο σύνολο δεδομένων και της αυτοματοποιημένης εξαγωγής και δομικής ανάλυσης των ομάδων.

```
(.venv) star >>> clusterMD --auto --rmsd 2 --method eom
=== clusterMD | HDBSCAN Analysis of Molecular-Dynamics PCA Coordinates ===
[*] Locating fluctuations file...
[WARN] Multiple fluctuation files found. Selecting the most recently modified.
[+] Input file: carma_results_03.10.2026_18.04.13/carma.dPCA.fluctuations.dat
[*] Analysis directory: carma_results_03.10.2026_18.04.13
Command line: clusterMD --auto --rmsd 2 --method eom
[*] Output directory: clusterMD_results_11.10.2026_16.33.08.761
[*] Pre-allocating memory and loading data (Stride: 1)...
[+] Data ingested successfully: 2,510,667 frames (5D).
[*] Auto-scaling activated: min_cluster_size=792
```

## ANALYSIS SETUP

```
Projection file      : carma.dPCA.fluctuations.dat
Frames analyzed     : 2,510,667
Frame stride        : 1
Dimensions used     : 5
Temperature         : 298.0 K
HDBSCAN method      : eom
min_cluster_size    : 792
min_samples         : 79
Auto-scaling        : Active
```

```
[*] Running HDBSCAN clustering...
[+] Identified 115 clusters in 149.88 seconds.
[*] Sorting cluster assignments by descending population size...
```

## CLUSTER POPULATIONS

| Cluster | Frames  | Population |
|---------|---------|------------|
| 0       | 252,841 | 10.1%      |
| 1       | 182,466 | 7.3%       |
| 2       | 79,108  | 3.2%       |
| 3       | 76,025  | 3.0%       |
| 4       | 41,966  | 1.7%       |
| 5       | 26,441  | 1.1%       |

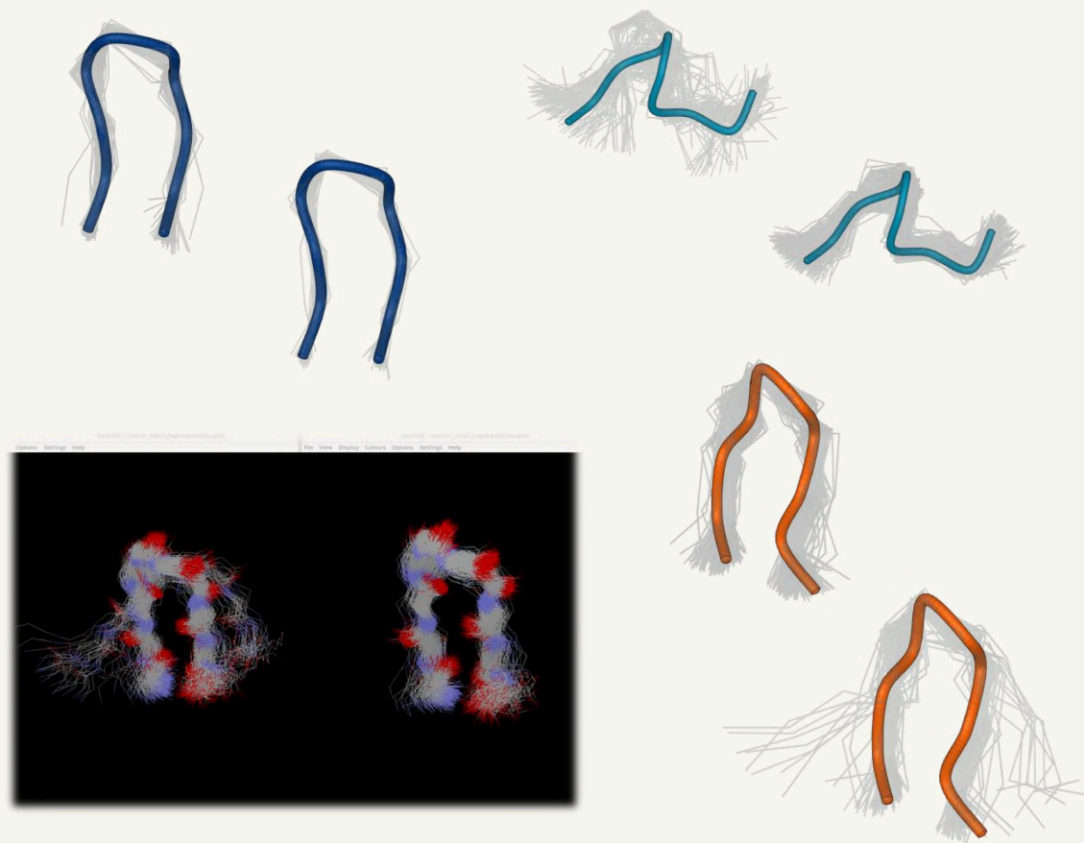
# Αντιπροσωπευτικές δομές και φιλτράρισμα

Το *clusterMD* χρησιμοποιεί τα αναγνωριστικά των στιγμιότυπων κάθε ομάδας για να καλέσει το **carma** και να **εξαγάγει τις αντίστοιχες διαμορφώσεις από την αρχική τροχιά**. Οι δομές υπερτίθενται με προσαρμογή των επιλεγμένων ατόμων, αφαιρώντας τις συνολικές μετατοπίσεις και περιστροφές, και υπολογίζεται η μέση δομή του συνόλου.

Ως **αντιπροσωπευτικό** επιλέγεται το στιγμιότυπο που βρίσκεται πλησιέστερα στο αριθμητικό κέντρο της ομάδας στον χώρο ομαδοποίησης.

Το **μέγεθος RMSD** εκφράζει την απόκλιση των αντίστοιχων ατομικών θέσεων κάθε υπερτεθειμένης διαμόρφωσης από τη μέση δομή, σε Å. Εφόσον οριστεί όριο με `--rmsd`, διατηρούνται οι διαμορφώσεις εντός του ορίου και επανυπολογίζεται η μέση δομή του υποσυνόλου, **χωρίς αλλαγή των ετικετών ή των πληθυσμών του HDBSCAN**.

```
clusterMD --dim 5 --size 500 --samples 50 --method eom --rmsd 2
```



# Ειδικές επιλογές εκτέλεσης

Η επιλογή **--cluster-only** εκτελεί την ανάγνωση των συντεταγμένων και την ομαδοποίηση HDBSCAN, αποθηκεύει τις αναθέσεις και τερματίζει πριν από την παραγωγή γραφημάτων και τη δομική ανάλυση. Επιτρέπει έτσι τη διερεύνηση των παραμέτρων ομαδοποίησης με περιορισμένη επεξεργασία.

Η επιλογή **--auto** υπολογίζει αυτόματα τις παραμέτρους `min_cluster_size` και `min_samples` μέσω ευρετικού κανόνα που βασίζεται στον αριθμό των αναλυόμενων στιγμιοτύπων· αποτελεί αρχική επιλογή παραμέτρων και απαιτεί αξιολόγηση της ευαισθησίας των αποτελεσμάτων.

```
clusterMD --cluster-only --auto --dim 5 --method leaf
```

# Παρατηρούμενες αλλαγές μεταξύ ομάδων

Το *clusterMD* εξετάζει τις ετικέτες κατά τη σειρά των στιγμιοτύπων στην τροχιά, παραλείπει τα μη ανατεθειμένα σημεία (-1) και καταμετρά τις αλλαγές μεταξύ διαδοχικών ανατεθειμένων ομάδων. Τα ζεύγη καταγράφονται χωρίς κατεύθυνση: η ακολουθία **A** → **-1** → **B** προσθέτει μία αλλαγή στο ζεύγος A-B, ενώ η **A** → **-1** → **A** δεν προσθέτει αλλαγή.

Η επιλογή **--transitions**, μέσα σε υπάρχοντα κατάλογο αποτελεσμάτων, εμφανίζει τις συχνότερες αλλαγές ή εξετάζει συγκεκριμένο ζεύγος ομάδων, χωρίς επανάληψη της ομαδοποίησης. Οι μετρήσεις περιγράφουν τη συνδεσιμότητα της ακολουθίας αναθέσεων και **δεν αποτελούν εκτιμήσεις κινητικών ρυθμών**.

```
=== clusterMD | Post-Run Observed-Change Inspection ===
[*] Target data located: /home/tmelas/thesis/examples/small/clusterMD_results_05.
[*] Pre-allocating memory and loading data (Stride: 1)...

=====
[*] DIAGNOSTIC MODE: OBSERVED CHANGE OVERVIEW
=====

OBSERVED LABEL CHANGES
-----
Cluster pair      Observed changes
-----
2 <-> 6           255
5 <-> 7           224
2 <-> 3           169
0 <-> 4           162
3 <-> 8           112
1 <-> 10          102
6 <-> 7           64
9 <-> 13          63
5 <-> 9           53
5 <-> 11          51
1 <-> 4           46
6 <-> 8           42
1 <-> 14          39
7 <-> 13          38
1 <-> 23          34
-----
Unique cluster pairs      88
[+] Validation complete. Exiting smoothly without triggering extraction pipeline.
```

```
[*] Inspecting chronological change events for cluster pair (1, 5)...

--- Observed Change Event 1 ---
PREVIOUS ASSIGNED : Frame 48907 | Cluster 5 | Coordinates: 2.94123 0.48448 0.07982 0.58009 -0.11484
... [ 4 intervening unassigned/noise frames (-1) ] ...
UNASSIGNED: Frame 48909 | Cluster -1 | Coordinates: 3.06675 0.69396 -0.24928 0.47223 -0.51778
UNASSIGNED: Frame 48911 | Cluster -1 | Coordinates: 2.68307 0.94140 0.39563 0.59435 -0.35896
... (additional unassigned/noise frames hidden to save space)
NEXT ASSIGNED : Frame 48917 | Cluster 1 | Coordinates: 2.16158 -0.55150 -0.15230 0.47289 -0.62759

--- Observed Change Event 2 ---
PREVIOUS ASSIGNED : Frame 700975 | Cluster 5 | Coordinates: 3.42246 -0.09907 0.57995 0.29125 -0.56134
... [ 2 intervening unassigned/noise frames (-1) ] ...
UNASSIGNED: Frame 700977 | Cluster -1 | Coordinates: 2.89804 -0.26646 0.32505 0.47900 -0.85195
UNASSIGNED: Frame 700979 | Cluster -1 | Coordinates: 2.51286 -0.69806 0.26153 0.73453 -1.13089
NEXT ASSIGNED : Frame 700981 | Cluster 1 | Coordinates: 2.21195 -0.54332 -0.17641 0.27457 -0.84378

--- Observed Change Event 3 ---
PREVIOUS ASSIGNED : Frame 700985 | Cluster 1 | Coordinates: 2.24881 -0.46949 -0.28676 0.29145 -0.78170
... [ 12 intervening unassigned/noise frames (-1) ] ...
UNASSIGNED: Frame 700987 | Cluster -1 | Coordinates: 2.45369 -0.57948 0.34107 0.33571 -0.36180
UNASSIGNED: Frame 700989 | Cluster -1 | Coordinates: 2.26856 -0.70184 0.63119 0.60337 -0.14782
... (additional unassigned/noise frames hidden to save space)
NEXT ASSIGNED : Frame 701011 | Cluster 5 | Coordinates: 3.15693 -0.08791 0.66679 0.51002 -0.47390

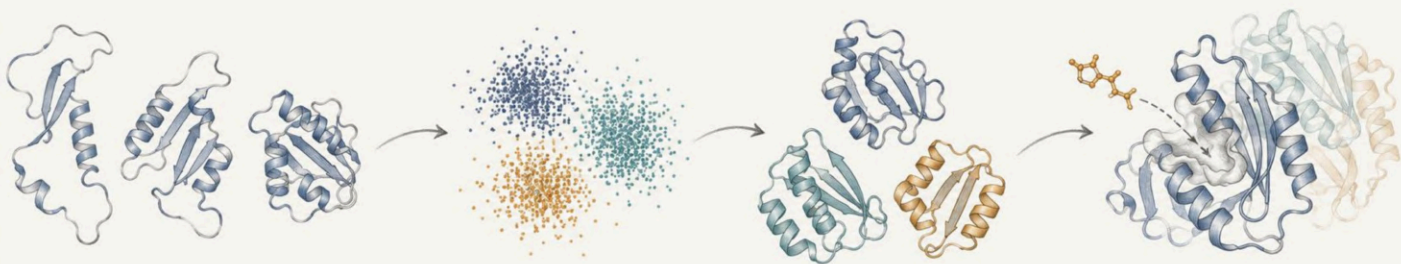
[+] ... and 1 additional events hidden.
=====
```

clusterMD --transitions # Συχνότερες αλλαγές  
clusterMD --transitions 1 5 # Έλεγχος του ζεύγους 1-5

# Πιθανές εφαρμογές στη Δομική Βιοπληροφορική

Το clusterMD μπορεί να υποστηρίξει τη διερεύνηση της **διαμορφωτικής ποικιλότητας πρωτεϊνών και πεπτιδίων**, οργανώνοντας στιγμιότυπα μοριακής δυναμικής σε ομάδες και εξάγοντας δομές για περαιτέρω ανάλυση. Πιθανές εφαρμογές περιλαμβάνουν τη μελέτη αναδίπλωσης, την εξέταση κινήσεων λειτουργικών περιοχών και τη σύγκριση προσομοιώσεων διαφορετικών μεταλλάξεων ή συνθηκών πρόσδεσης.

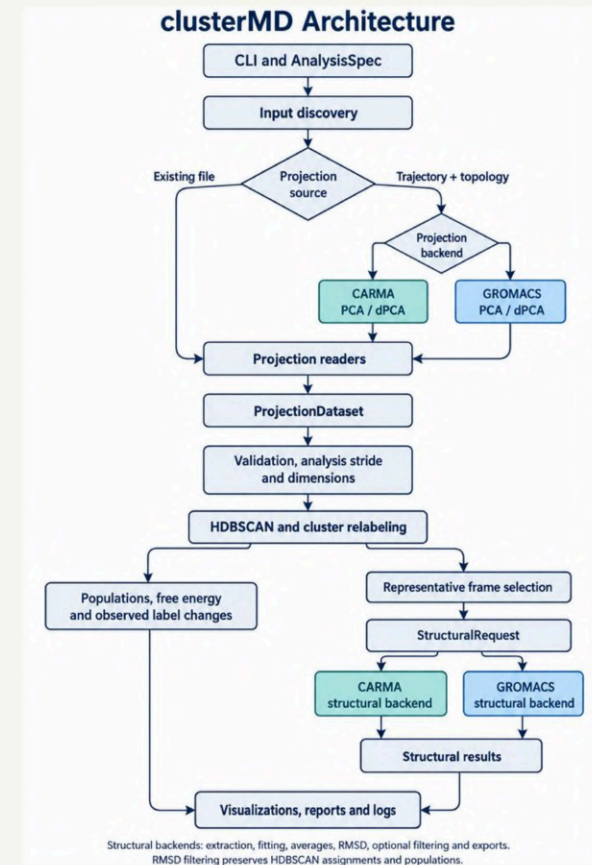
Οι εξαγόμενες διαμορφώσεις μπορούν επίσης να αξιοποιηθούν ως **σύνολα δομών για μοριακή πρόσδεση ή λεπτομερέστερη δομική αξιολόγηση**, με την επιλογή τους να απαιτεί έλεγχο της επάρκειας της δειγματοληψίας και της βιολογικής τους συνάφειας.



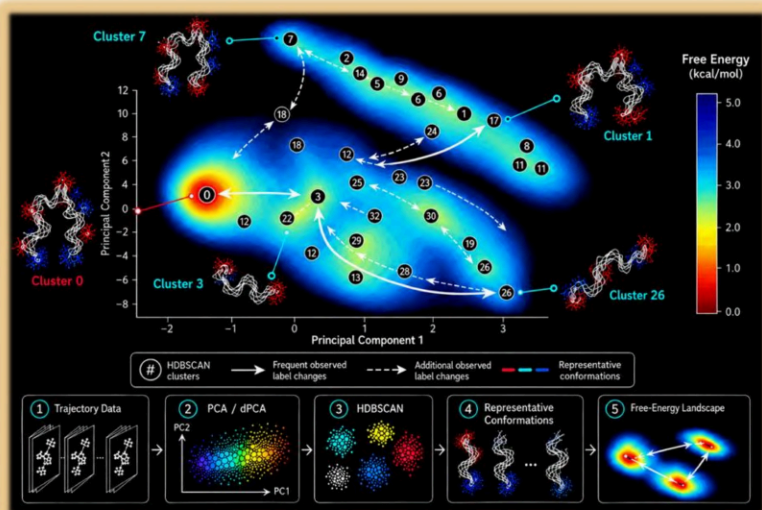
# Επόμενα βήματα: αρθρωτή αρχιτεκτονική και διεύρυνση υποστήριξης

Η εξέλιξη του clusterMD περιλαμβάνει τη μετάβαση από το ενιαίο script σε **αρθρωτή αρχιτεκτονική (modular architecture)**, με διακριτές ενότητες για την ανάγνωση δεδομένων, την ομαδοποίηση, τον χαρακτηρισμό των ομάδων, τη δομική ανάλυση και την παραγωγή αναφορών. Ο διαχωρισμός αυτός διευκολύνει τη συντήρηση, τους ανεξάρτητους ελέγχους και την προσθήκη λειτουργιών, διατηρώντας τη συμπεριφορά της υπάρχουσας έκδοσης.

Παράλληλα, προβλέπεται η ενσωμάτωση **πρόσθετων εργαλείων δομικής ανάλυσης, όπως το GROMACS**, μέσω κοινής διεπαφής εξαγωγής, υπέρθεσης και υπολογισμού RMSD. Η επέκταση θα επιτρέψει την επεξεργασία περισσότερων μορφών τροχιών και τοπολογιών, με επαλήθευση της αντιστοίχισης των στιγμιοτύπων και της συνέπειας των αποτελεσμάτων.



# Ευχαριστώ για την προσοχή σας!



Περισσότερες πληροφορίες και την τρέχουσα έκδοση του λογισμικού μπορείτε να βρείτε στον σύνδεσμο: <https://github.com/tasosmelas/md-hdbscan-clustering>

## Ενδεικτική βιβλιογραφία

- **Glykos, N. M. (2006).** Carma: A molecular dynamics analysis program. *Journal of Computational Chemistry*, **27**, 1765–1768.
- **Altis, A., et al. (2008).** Dihedral angle principal component analysis of molecular dynamics simulations. *The Journal of Chemical Physics*, **128**, 245102.
- **Koukos, P. I., & Glykos, N. M. (2013).** grcarma: A fully automated task-oriented interface for the analysis of molecular dynamics trajectories. *Journal of Computational Chemistry*, **34**, 2310–2312.
- **Campello, R. J. G. B., Moulavi, D., & Sander, J. (2013).** Density-based clustering based on hierarchical density estimates. *Advances in Knowledge Discovery and Data Mining*, 160–172.
- **McInnes, L., Healy, J., & Astels, S. (2017).** hdbscan: Hierarchical density based clustering. *Journal of Open Source Software*, **2**(11), 205.