Guiding rational vaccine design using random matrix theory

Matthew McKay
Department of Electronic and Computer Engineering
Department of Chemical and Biological Engineering

December 10, 2019
Random Matrices and Complex Data Analysis Workshop
Shanghai University of Finance and Economics

香港科技大學
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

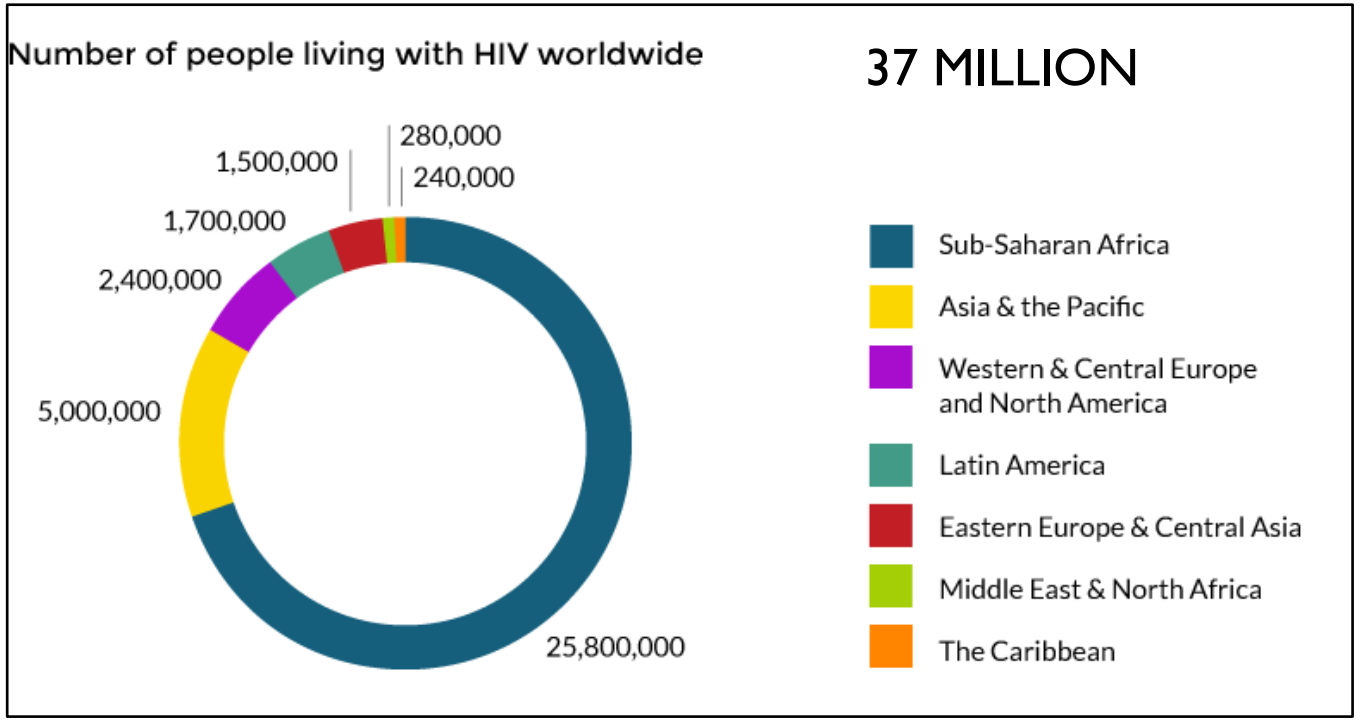
Vaccination

- Eradication or near-eradication of diseases such as smallpox and polio
- Still no effective vaccine against many pathogens
- Main focus of today's talk: **HCV and HIV (and a bit about polio)**

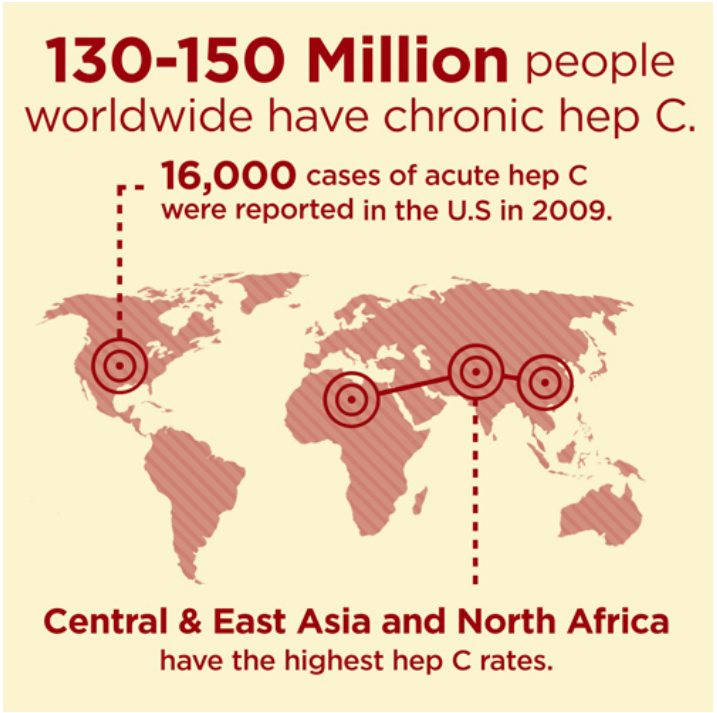
…. and …

How is it that RMT and high-dimensional statistics can potentially help?

HIV

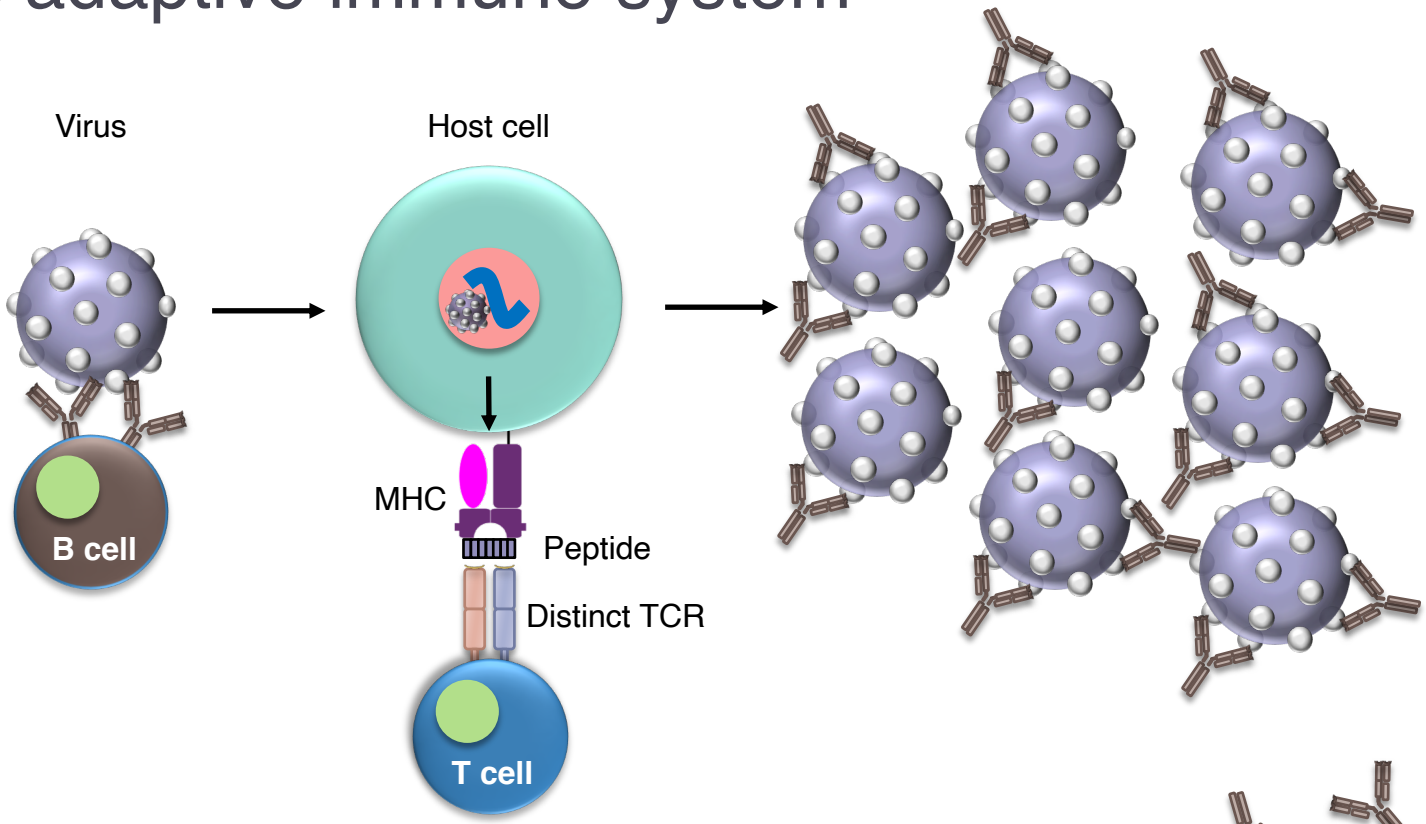


HCV

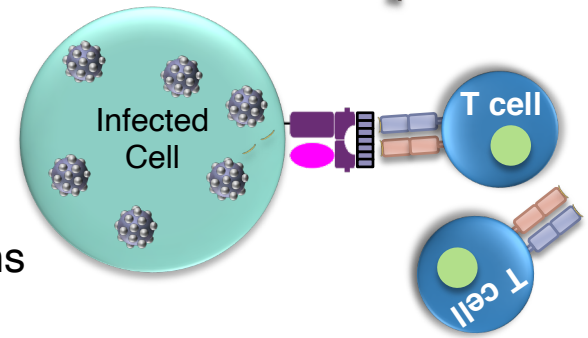


Requirement: Better understanding of these viruses to combat them ...

Pathogen specific adaptive immune system



Basis for vaccination

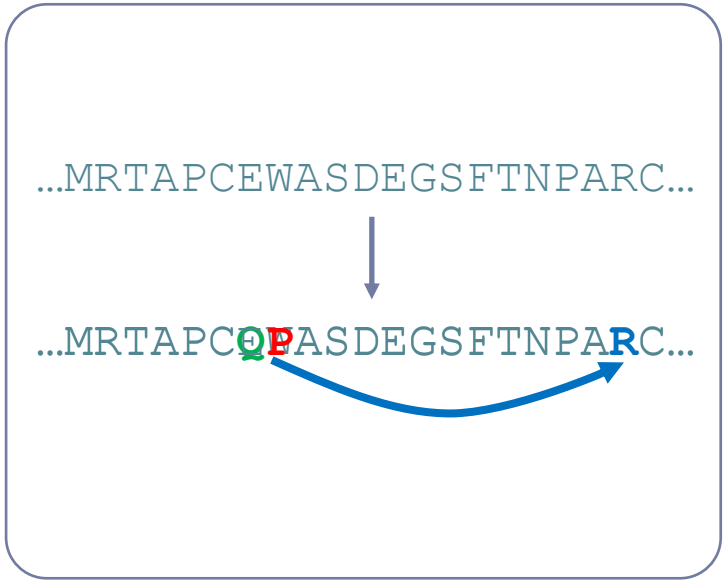


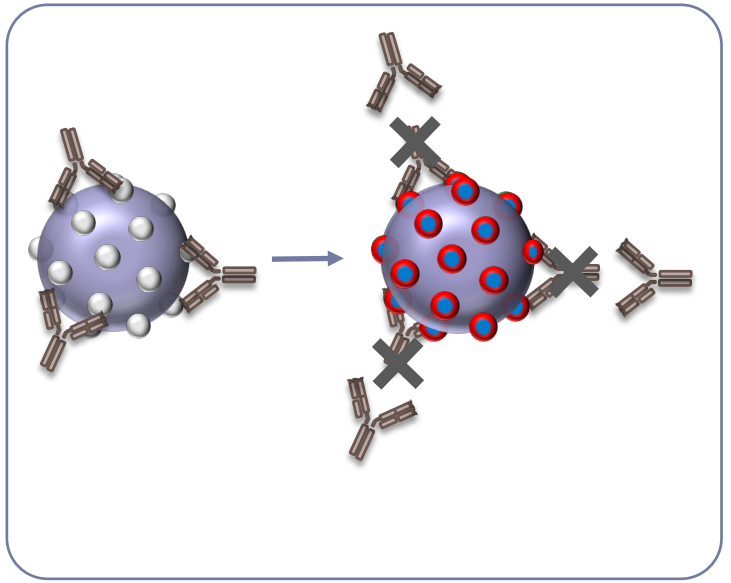
Memory of past infections

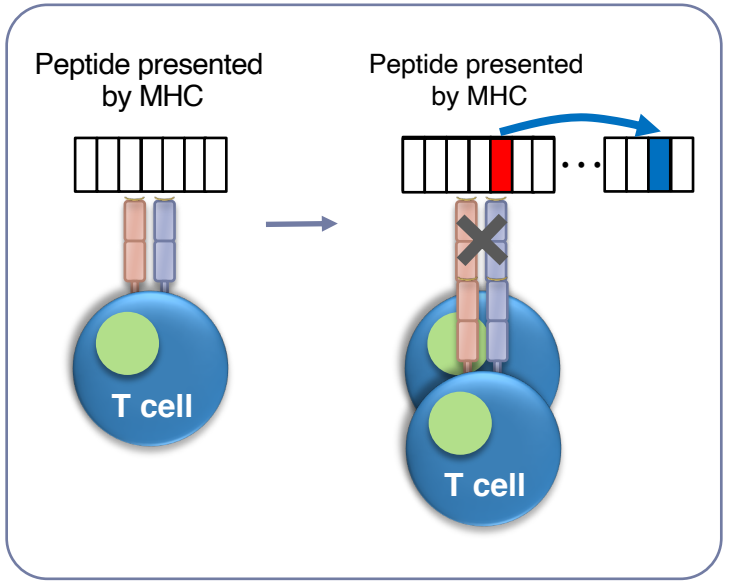
Immune system evasion by HIV and HCV

Principle of vaccination

Train the immune system (B cells and T cells) to recognize viral particles prior to natural infection







Mutations during replication

High mutation rate results in immune escape if the mutant virus has a **high fitness**

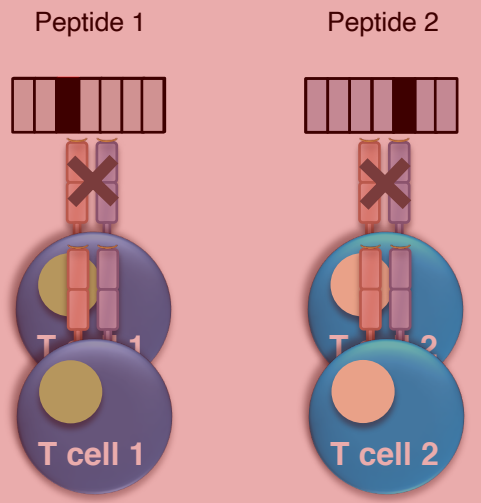
Additional complication

Compensation of **deleterious** effect of individual mutations

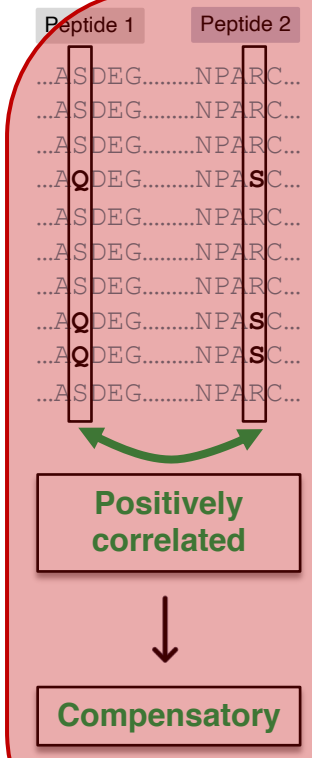
T-cell vaccine design using sparse PCA

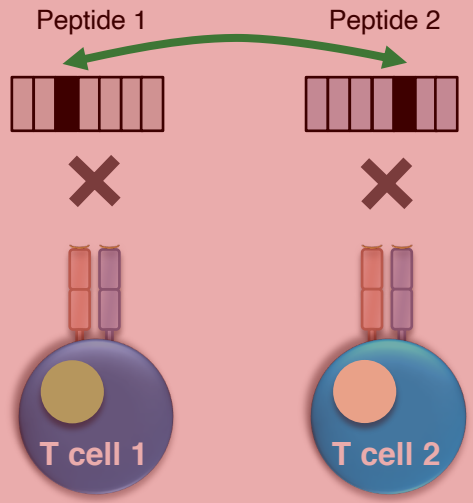
T-cell based vaccine design – Moving beyond conservation

Conventional vaccine design strategy:
Target **conserved** regions of proteins

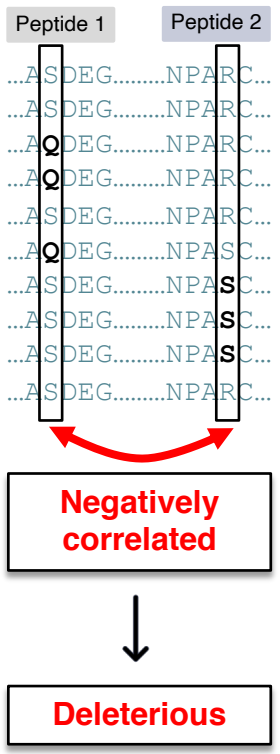


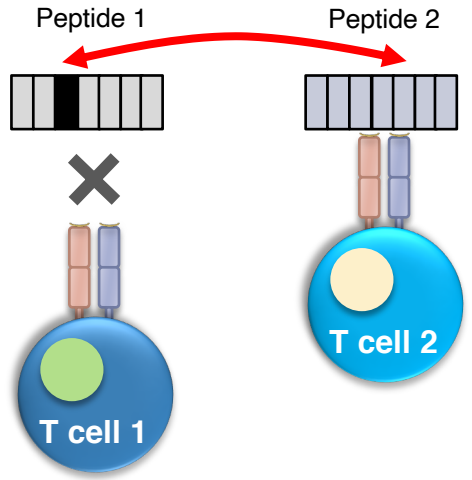
Escapes from **both** T cells



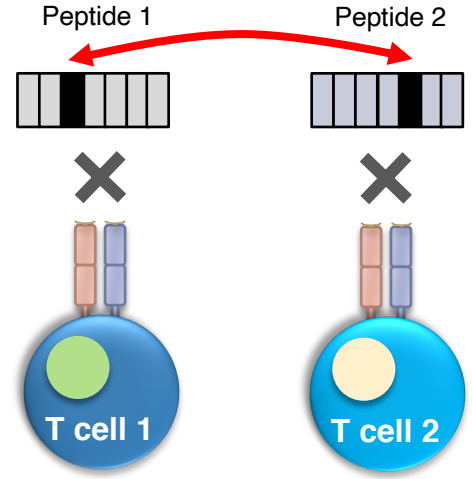


Escapes from **both** T cells





Escape **blocked** from **<u>one</u>** T cell

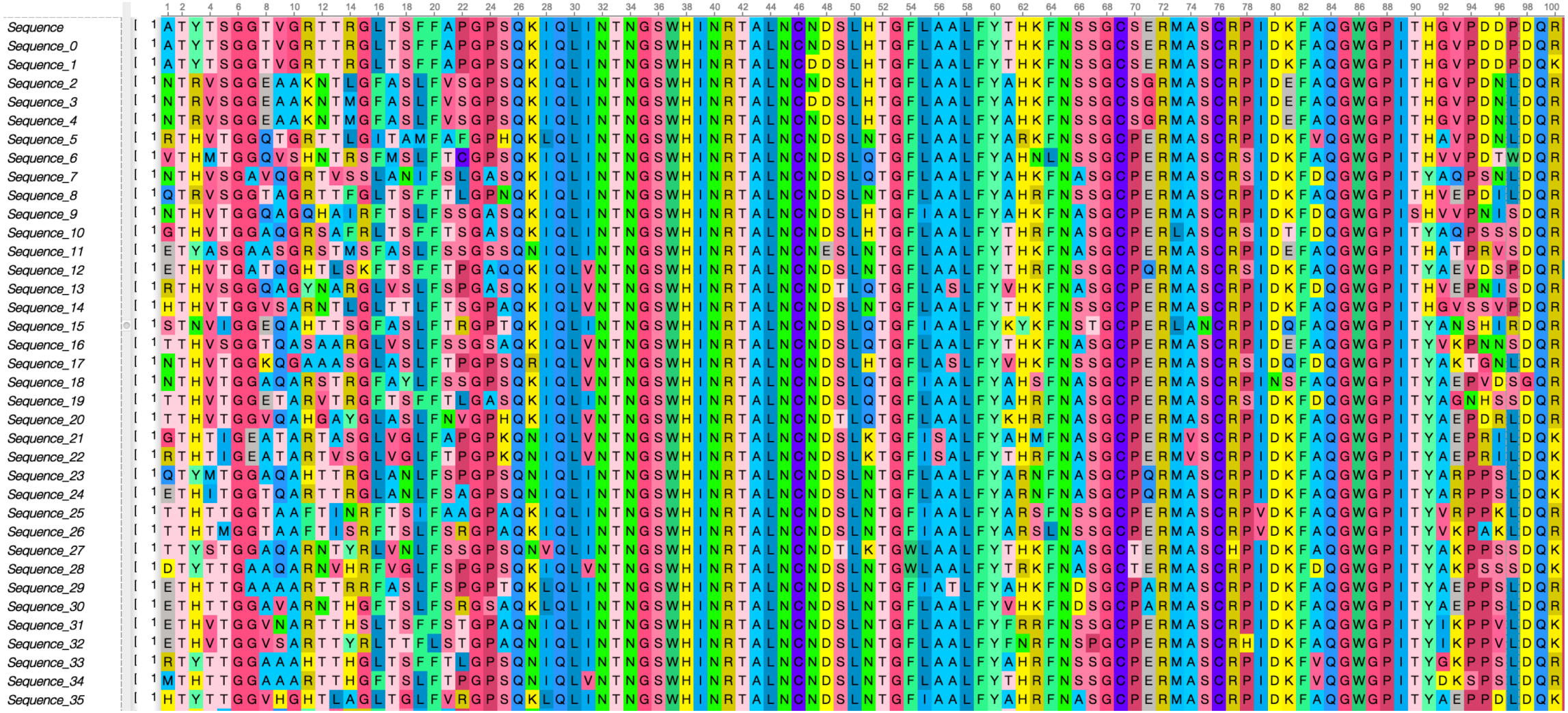


Escapes from **both** T cells
but mutant strain is potentially **<u>unfit</u>**

Strategy: Select peptides that comprise negatively correlated site pairs and avoid positively correlated ones

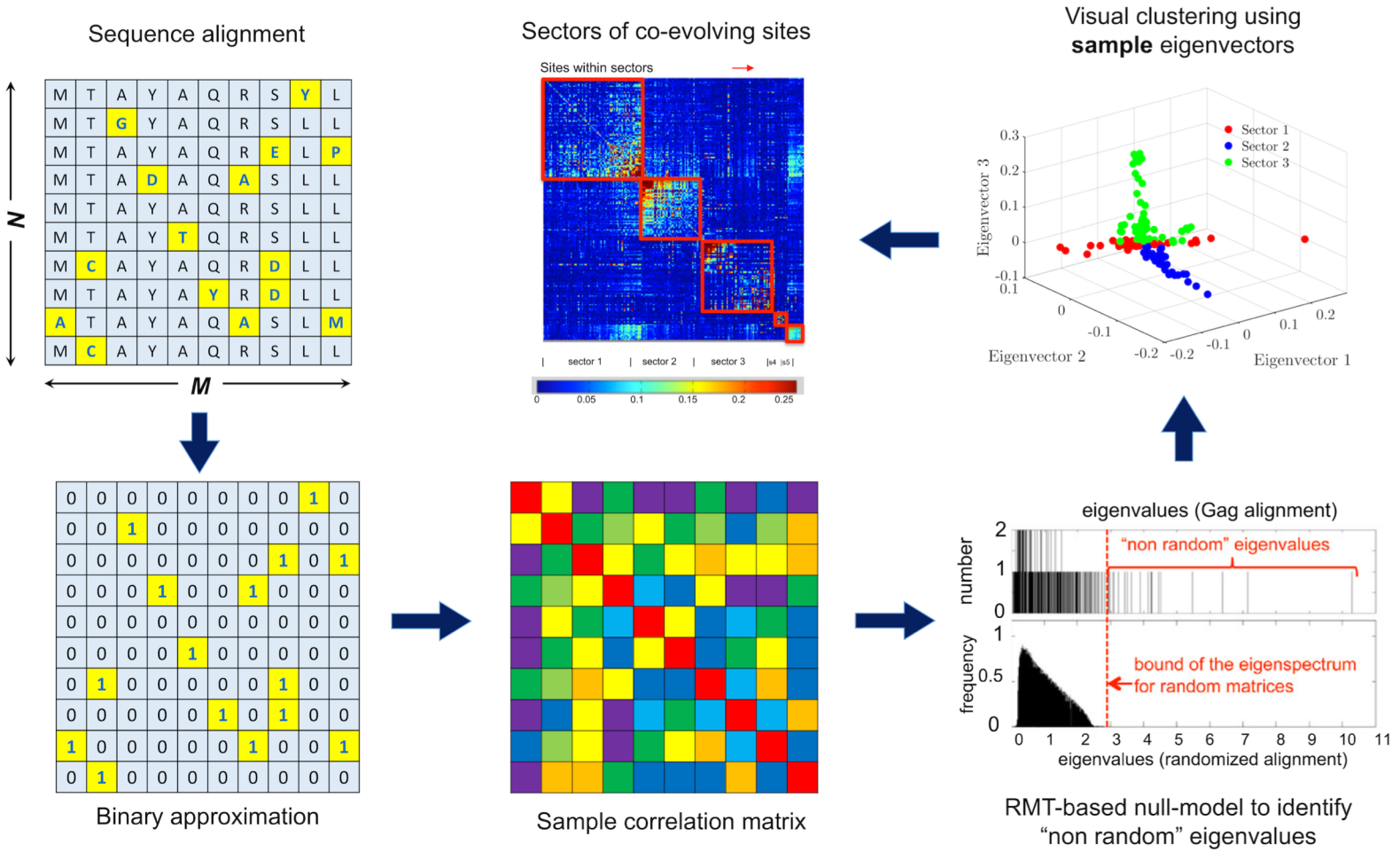
A. A. Quadeer, et al. "Statistical linkage analysis of substitutions in patient-derived sequences of genotype 1a hepatitis C virus nonstructural protein 3 exposes targets for immunogen design." *Journal of Virology* 88 (**2014**).

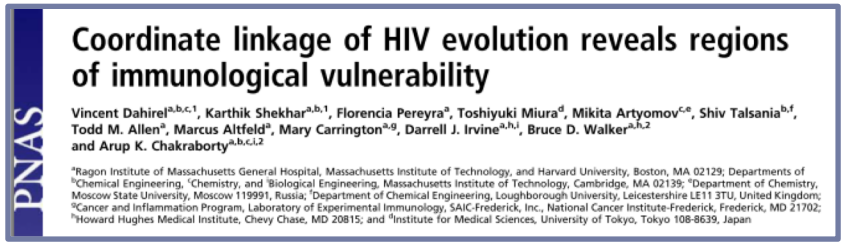
Example multiple sequence alignment (MSA) of a HCV protein



Motivation:
Work on HIV Gag by Arup Chakraborty's group at MIT

- Inspired by **RMT-based noise cleaning** in **finance** (Bouchaud and Stanley's work)





Coordinate linkage of HIV evolution reveals regions of immunological vulnerability

Vincent Dahirel[a,b,c,1], Karthik Shekhar[a,b,1], Florencia Pereyra[a], Toshiyuki Miura[d], Mikita Artyomov[c,e], Shiv Talsania[b,f], Todd M. Allen[a], Marcus Altfeld[a], Mary Carrington[a,g], Darrell J. Irvine[a,h,i], Bruce D. Walker[a,h,2] and Arup K. Chakraborty[a,b,c,i,2]

[a]Ragon Institute of Massachusetts General Hospital, Massachusetts Institute of Technology, and Harvard University, Boston, MA 02129; Departments of [b]Chemical Engineering, [c]Chemistry, and [i]Biological Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139; [e]Department of Chemistry, Moscow State University, Moscow 119991, Russia; [f]Department of Chemical Engineering, Loughborough University, Leicestershire LE11 3TU, United Kingdom; [g]Cancer and Inflammation Program, Laboratory of Experimental Immunology, SAIC-Frederick, Inc., National Cancer Institute-Frederick, Frederick, MD 21702; [h]Howard Hughes Medical Institute, Chevy Chase, MD 20815; and [d]Institute for Medical Sciences, University of Tokyo, Tokyo 108-8639, Japan

- Identified a sector of co-evolving sites enriched in **negative correlations**
- Proposed a vaccine to attack the sites in this sector

THE WALL STREET JOURNAL.

Home World U.S. Politics Economy Business Tech Markets Opinion Arts Life Real Estate

Toshiba Profit Markdown May Double Earlier Estimates

Southwest Airlines and Flight Attendants in Tentative Pact

Greek Importers Begin to Feel the Squeeze

HEALTH

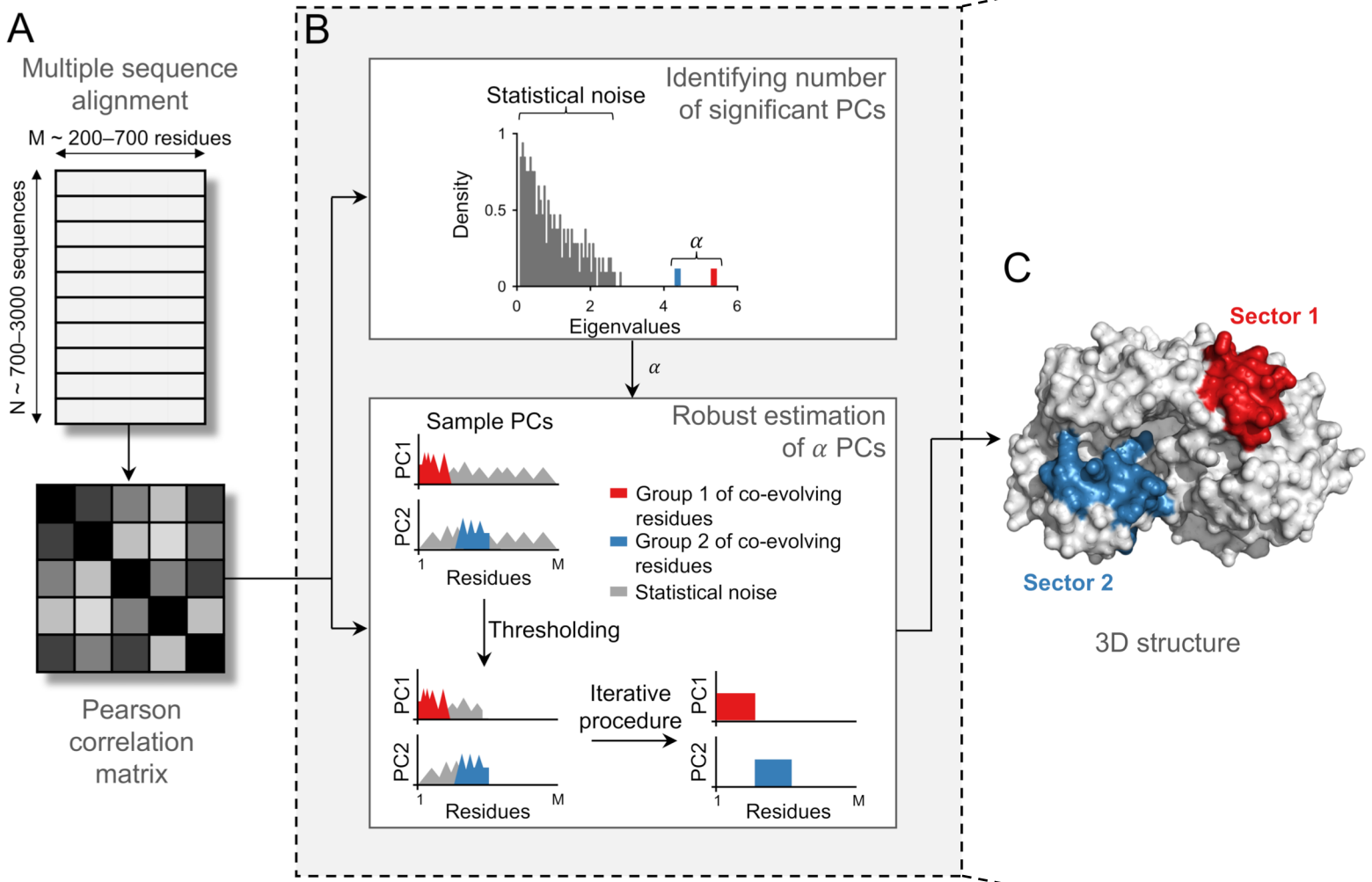
New Math in HIV Fight

Statistical Method Evolves From Physics to Wall Street to Battle Against AIDS

- Methods were based on variant of standard PCA (with eigenvalue clipping)
- With HCV being quite similar to HIV,
  - **Do such groups exist in HCV?**
  - **Is any group enriched with deleterious mutations?**
  - If so, **can this information guide a vaccine design against HCV?**

Dahirel, Vincent. et al. "Coordinate linkage of HIV evolution reveals regions of immunological vulnerability." *Proceedings of the National Academy of Sciences (PNAS)* 108 (**2011**).

Robust co-evolutionary analysis (RoCA) of proteins



- When N and M are both large and comparable, the sample principal components (PCs) are also not **consistent** estimators

- RMT-based sparse PCA
  - Diagonal thresholding [Johnstone 2009]
  - Augmented SPCA [Paul 2012]
  - **Iterative thresholding SPCA (ITSPCA)** [Ma 2013]

Summary

- Based on "orthogonal iteration method"
- Introduces a "**thresholding** step" to introduce sparsity
- Estimates sparse subspace of the leading eigenvectors

Robust Co-evolutionary Analysis (RoCA)

- Adapted to work on correlation matrices
- A data-driven **thresholding** parameter designed using ideas from random matrix theory

A. A. Quadeer, D. Morales-Jimenez, M. R. M., "Co-evolution networks of HIV/HCV are modular with direct association to structure and function." *PLoS Computational Biology* 14 (**2018**).

Robust co-evolutionary analysis (RoCA) algorithm

- Noise threshold designed based on **worst-case non-sector coordinate**
- Following [Paul, Stat Sinica 2007], entries of $\mathbf{q}_k^{\mathcal{NS}}$ behave asymptotically, up to scaling, as entries of a uniformly distributed vector on the $(M-\alpha)$-dimensional unit sphere
- Define $\eta = \frac{M}{N}$. For N and M large, $M \gg \alpha$,

$$q_{k,\max}^{\mathcal{NS}} \triangleq \max_{i\in\{1,\ldots,M_{\text{ns}}\}} |q_k^{\mathcal{NS}}(i)| \overset{d}{\sim} \sqrt{\frac{1-c_k^2}{M}} \max_{i\in\{1,\ldots,M_{\text{ns}}\}} |y_{k,i}|$$

  - $y_{k,i}$ are independent Gaussian random variables

$$c_k = \sqrt{\left(1-\frac{\eta}{(\ell_k-1)^2}\right)\Big/\left(1+\frac{\eta}{\ell_k-1}\right)} \qquad \ell_k = [(\lambda_k+1-\eta)+\sqrt{(\lambda_k+1-\eta)^2-4\lambda_k}]/2$$

- CDF $F_{k,\max}(x)$ follows from standard order statistics
- Threshold selection: $\gamma_k = \lambda_k \tilde{\gamma}_k$ where $\tilde{\gamma}_k = F_{k,\max}^{-1}(0.95)$

Inputs:
1. Correlation matrix of size $M \times M$, $\mathbf{C}$;
2. Number of PCs to be estimated, α;
3. Noise threshold, γ_k, $k = 1, 2, \cdots, \alpha$.

Output: Robust estimates of the PCs, $\mathbf{p}_k$, given as columns of the $M \times \alpha$ matrix $\mathbf{P} = \mathbf{Q}^{(\infty)}$, where $\mathbf{Q}^{(\infty)}$ denotes $\mathbf{Q}^{(i)}$ at convergence.

1: Initialization: $i = 1$;
2: Initial orthonormal matrix of size $M \times \alpha$, $\mathbf{Q}^{(0)} = \mathbf{Q}_\alpha$; here $\mathbf{Q}_\alpha$ is a matrix whose columns are the α leading eigenvectors of $\mathbf{C}$, i.e., $\mathbf{Q}_\alpha = [\mathbf{q}_1\ \mathbf{q}_2\ \cdots\ \mathbf{q}_\alpha]$.
3: **repeat**
4: Multiplication: $\mathbf{T}^{(i)} = (T_{\ell k}^{(i)}) = \mathbf{C}\mathbf{Q}^{(i-1)}$;
5: Thresholding: $\hat{\mathbf{T}}^{(i)} = (\hat{T}_{\ell k}^{(i)})$, with $\hat{T}_{\ell k}^{(i)} = T_{\ell k}^{(i)} \mathbf{1}_{\{|T_{\ell k}^{(i)}|>\gamma_k\}}$, where $\mathbf{1}_{\{E\}}$ is the indicator function of an event E;
6: QR Factorization: $\mathbf{Q}^{(i)}\mathbf{R}^{(i)} = \hat{\mathbf{T}}^{(i)}$;
7: $i = i + 1$;
8: **until** *convergence*

$$\mathbf{C} = \sum_{i=1}^{M} \lambda_i \mathbf{q}_i \mathbf{q}_i^T$$

$$[\mathbf{t}_1, \mathbf{t}_2, \ldots, \mathbf{t}_\alpha] = [\lambda_1\mathbf{q}_1, \lambda_2\mathbf{q}_2, \ldots, \lambda_\alpha\mathbf{q}_\alpha]$$

$$\alpha = \max\{k \in \{1,\ldots,M\} : \lambda_k > \lambda_{\max}^{\text{rnd}}\}$$

$$\mathbf{q}_k = \begin{bmatrix} \mathbf{q}_k^{\mathcal{S}} \\ \mathbf{q}_k^{\mathcal{NS}} \end{bmatrix}$$

Sector coordinates

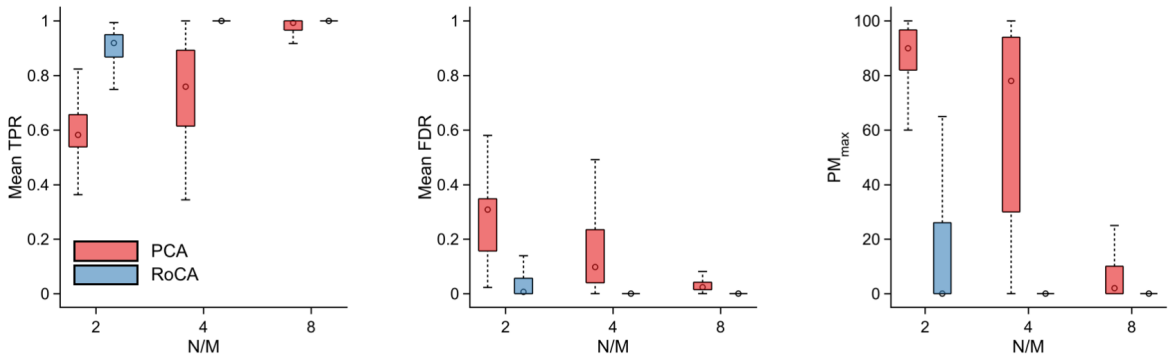
Non-Sector coordinates

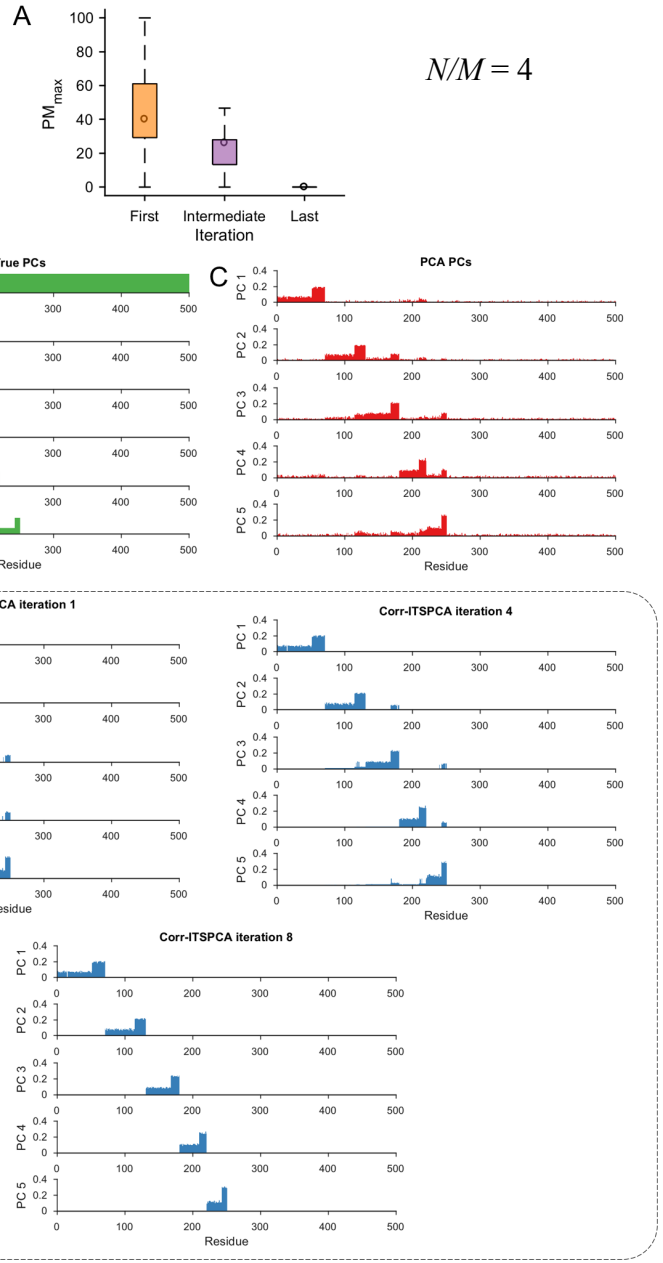
Model based simulation example

- M = 500 variables, various number of samples N
- 5 non-overlapping sectors
- Correlation matrix:

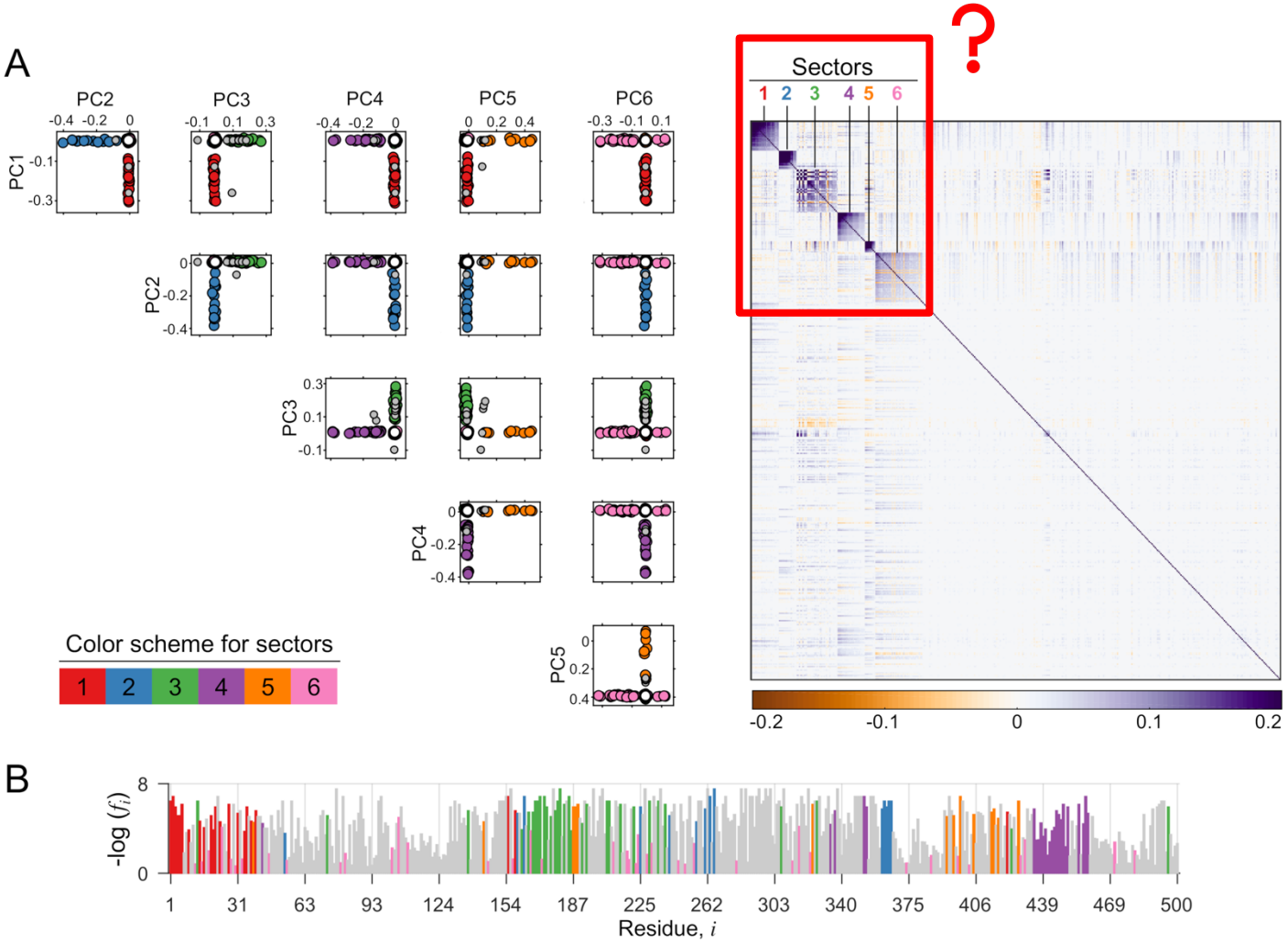
$$\mathbf{\Gamma} = (\mathbf{I}_M - \mathbf{Z}) + \underbrace{\sum_{k=1}^{r} \ell_k \mathbf{u}_k \mathbf{u}_k^{\mathsf{T}}}_{r \text{ units}} + \underbrace{\ell_0 \mathbf{u}_0 \mathbf{u}_0^{\mathsf{T}}}_{\text{Phylogeny}}$$

- Eigenvalues equally spaced in range [4, 6]

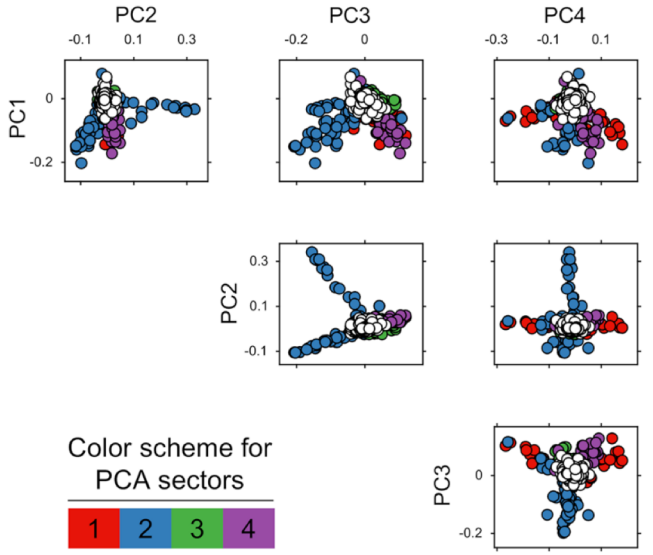




RoCA sectors for HIV Gag

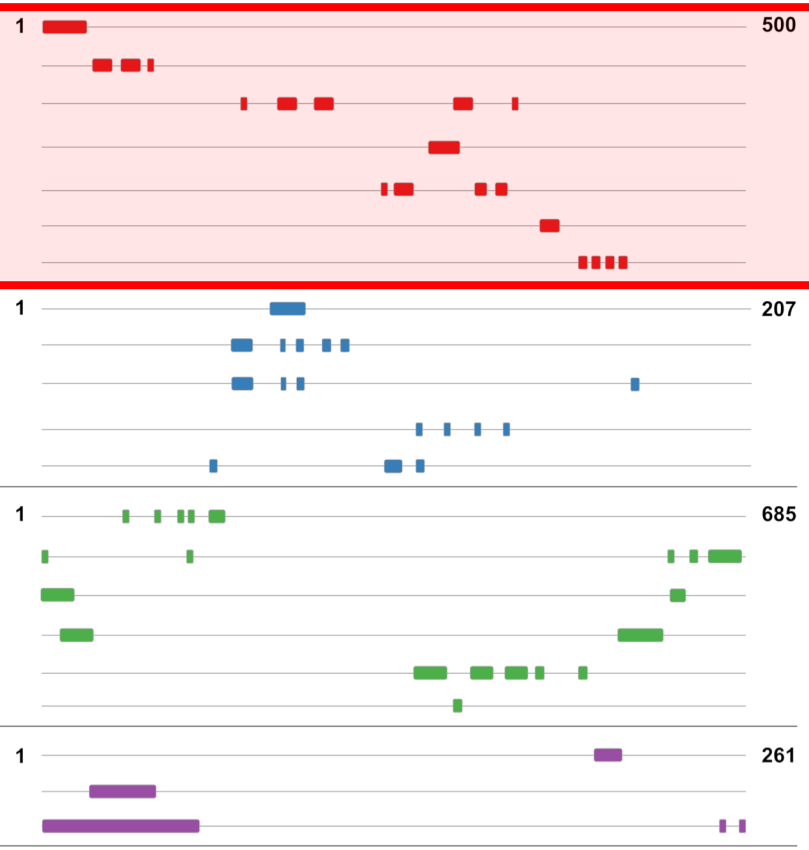


Sectors obtained using standard PCA



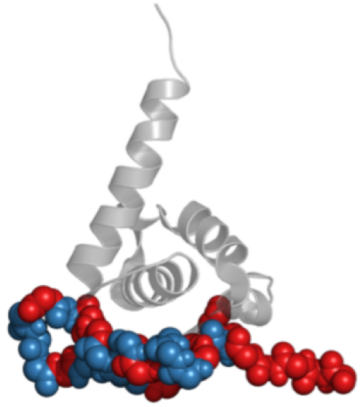
Experimentally identified biochemical domains

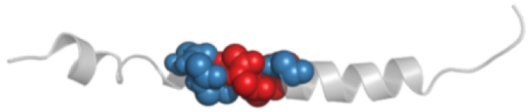
Protein	Biochemical domains[a]	Acronyms	Approximate location in protein sequence[b]
HIV Gag	Membrane-binding domain in p17	P17-Mem-Bin-Dom	
	Intra-trimer interface in p17	P17-Intra-Tri-Int	
	Intra-hexamer interface in p24	P24-Intra-Hex-Int	
	Major homology region in p24	P24-MHR	
	Inter-hexamer interface in p24	P24-Inter-Hex-Int	
	p24-SP1 interface	P24-SP1-Int	
	Zinc finger structures in p7	P7-Zinc-Finger	
HIV Nef	Enhancement of viral infectivity	Enh-Viral-Inf	
	HLA1 down-regulation	HLA1-Down-Reg	
	HLA2 down-regulation	HLA2-Down-Reg	
	Intra-dimer interface	Intra-Dimer-Int	
	CD4 down-regulation	CD4-Down-Reg	
HCV NS3-4A	NS3 substrate binding groove of protease activity	NS3-Sub-Bin-Pro	
	NS5A hyper-phosphorylation	NS5A-Hyper-Phos	
	NS3-NS4A membrane association	NS3-NS4A-Mem-Asso	
	NS3-NS4A interface for protease activation	NS3-NS4A-Pro-Act	
	Intra-dimer interface in NS3 helicase	NS3-Intra-Dimer-Int	
	Motif important for enzymatic and helicase activities in NS3	NS3-Motif-Enz-Heli	
HCV NS4B	Viral replication and assembly	Viral-Rep-Assm	
	Interaction with human protein ATF6beta	NS4B-ATF6beta	
	NS4B oligomerization	Oligomerization	



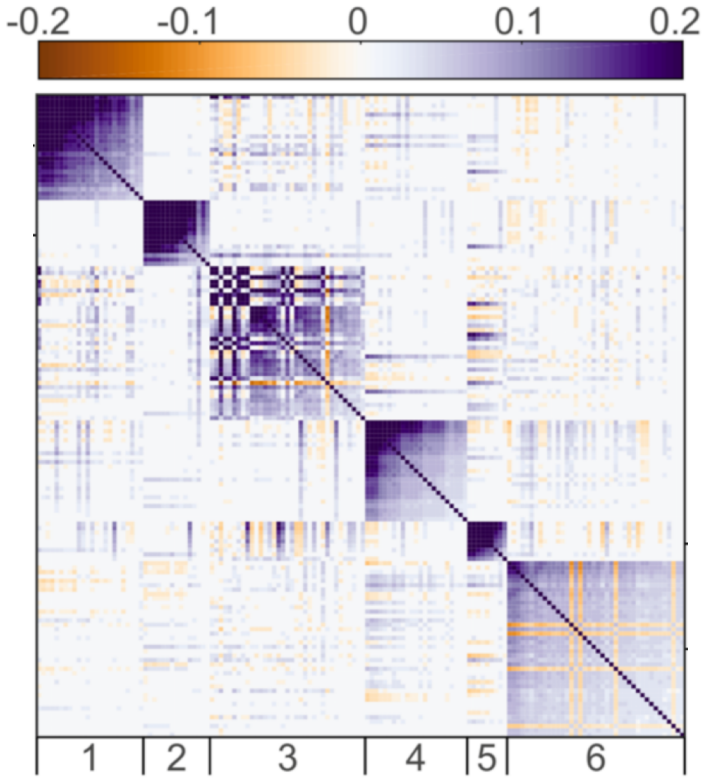
HIV Gag

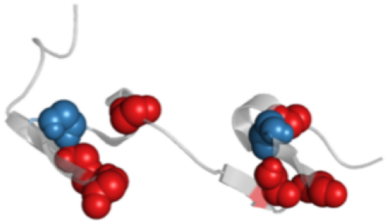
P17-Mem-Bin-Dom



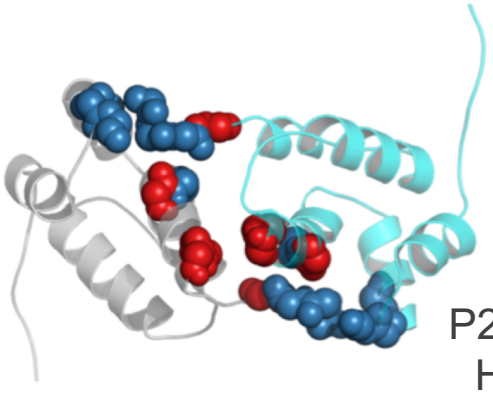


P24-SP1-Int

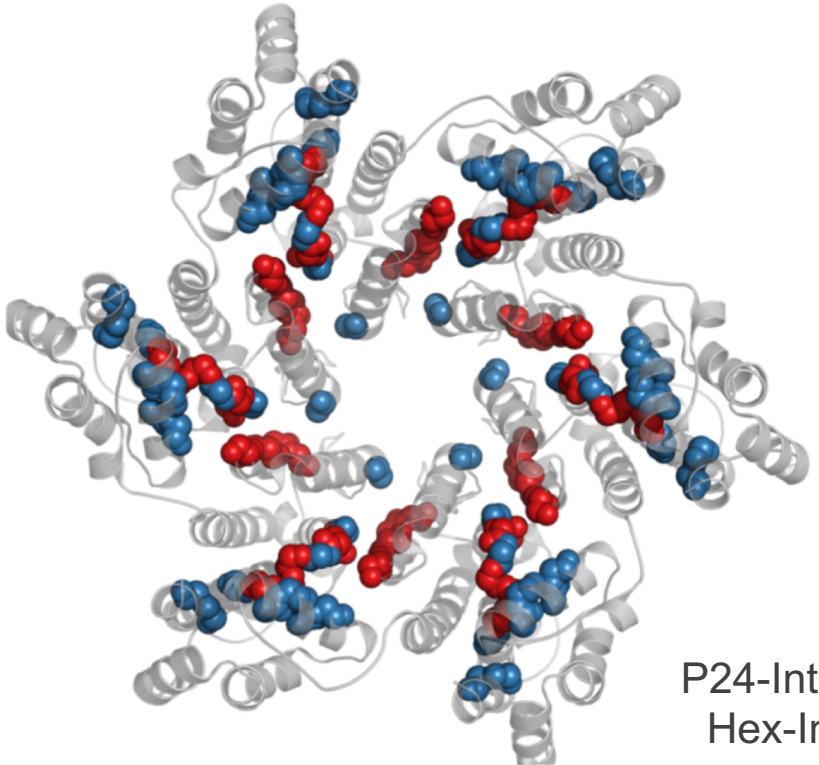




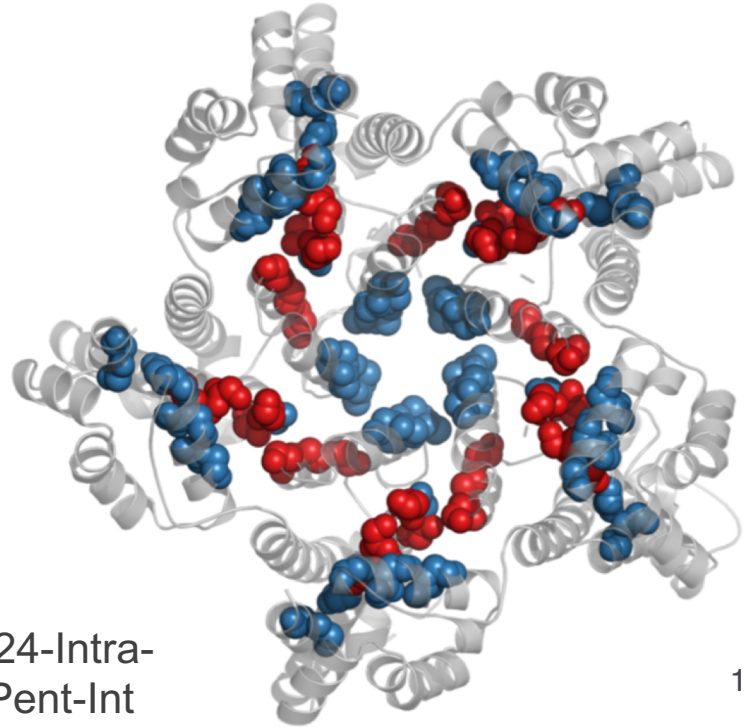
P7-Zinc-Finger



P24-Inter-Hex-Int



P24-Intra-Hex-Int



P24-Intra-Pent-Int

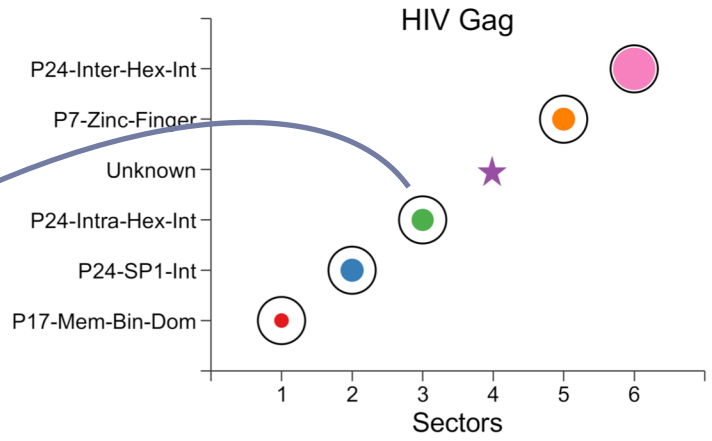
Results for other HIV and HCV proteins

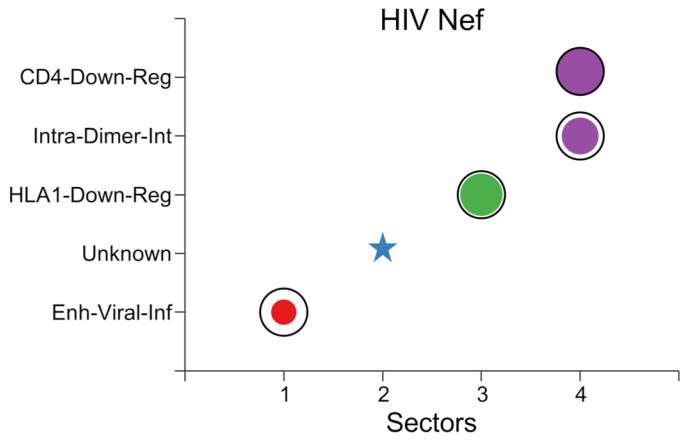
Immunological significance

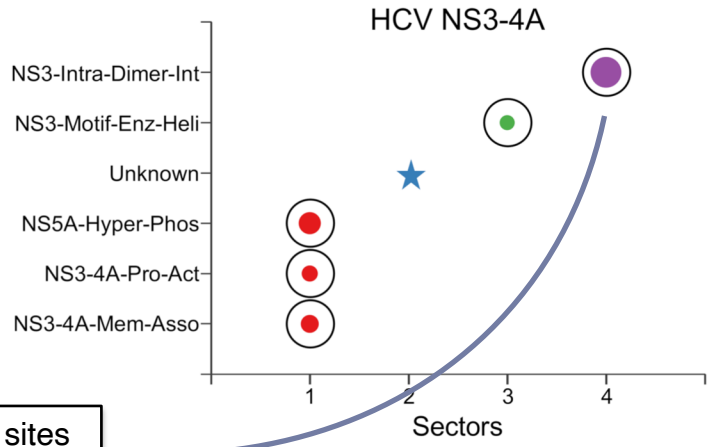
Association with sites targeted by **HIV LTNPs**

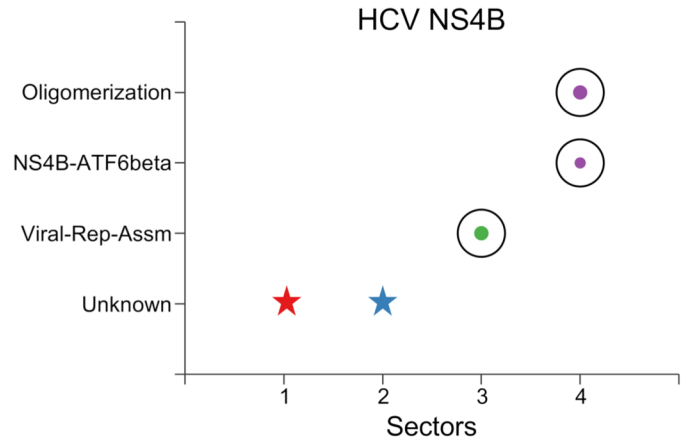
Enriched in **negative correlations**

Association with sites related to **HCV spontaneous clearance**



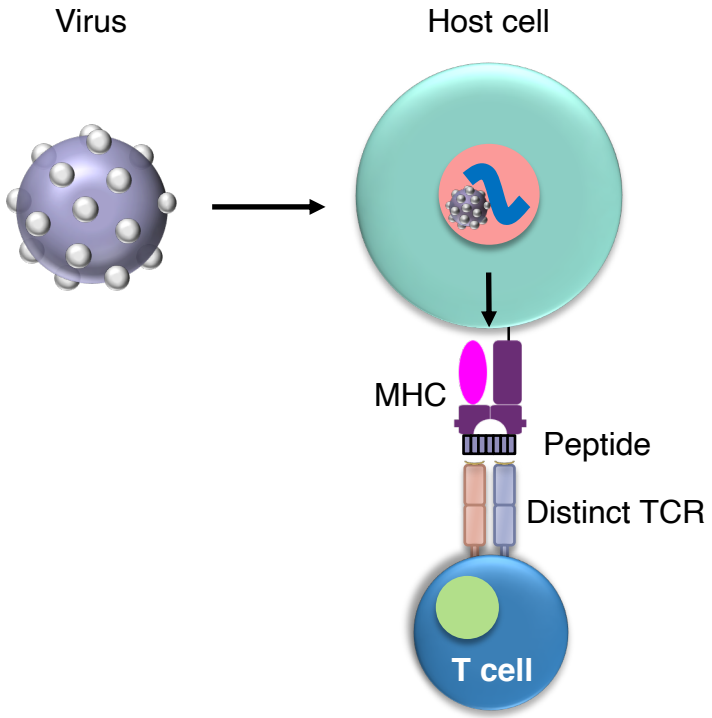


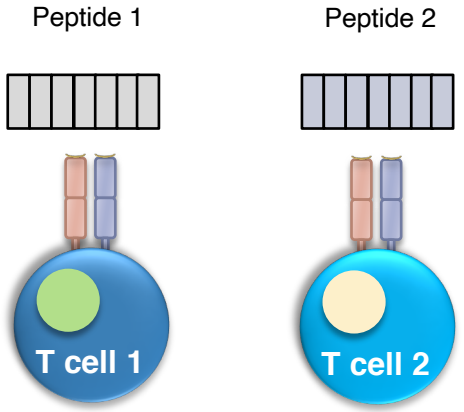




A. A. Quadeer, D. Morales-Jimenez, and M. R. M. “Co-evolution networks of HIV/HCV are modular with direct association to structure and function.” *PLoS Computational Biology* 14 (**2018**).

Back to the T-cell vaccine design problem…





Vaccine design goal: To **find sets of viral peptides** that:

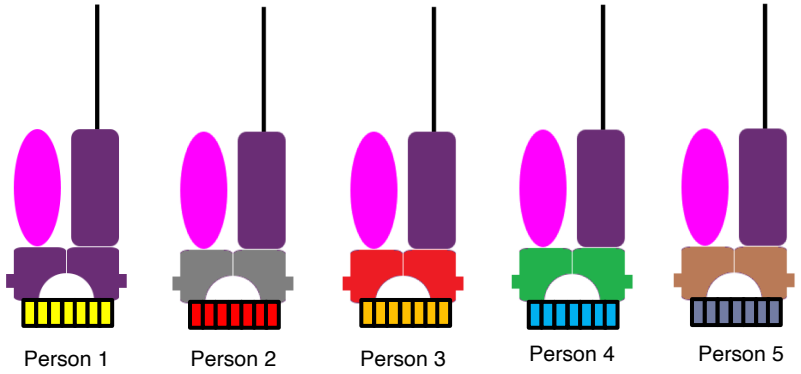
1. Can be “seen” by T cells in many people
2. Make it difficult for the virus to escape through mutation

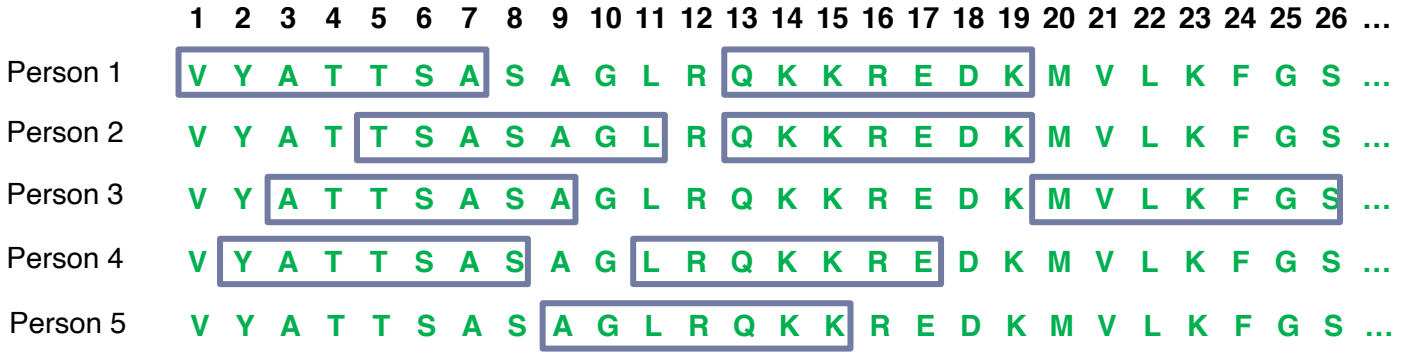
Proposed T-cell based vaccine design candidates

Peptide-selection problem

1. Maximize **population coverage**

Different people have different types of MHC molecules
Different MHC molecules may present different peptides





2. Maximize **L = PCP + PNCP – PPCP**

PCP = Percentage of 100% conserved pairs
PNCP = Percentage of negatively correlated pairs
PPCP = Percentage of positively correlated pairs

HCV NS3

Combination	Peptide 1	Peptide 2	Peptide 3	Peptide 4	Peptide 5	L	Dcov
1	1251-1259	1292-1300	1436-1444	1585-1594	1585-1595	63.6	**0.50**
2	1123-1131	1169-1177	1251-1259	1292-1300	1436-1444	61.6	**0.44**
3	1123-1131	1175-1183	1251-1259	1292-1300	1436-1444	65.4	**0.37**
4	1123-1131	1175-1183	1251-1259	1359-1367	1436-1444	61.6	**0.37**
5	1169-1177	1175-1183	1251-1259	1292-1300	1436-1444	64.5	**0.34**

A. A. Quadeer et al. “Statistical linkage analysis of substitutions in patient-derived sequences of genotype 1a hepatitis C virus nonstructural protein 3 exposes targets for immunogen design.” *Journal of Virology* 88 **(2014)**.

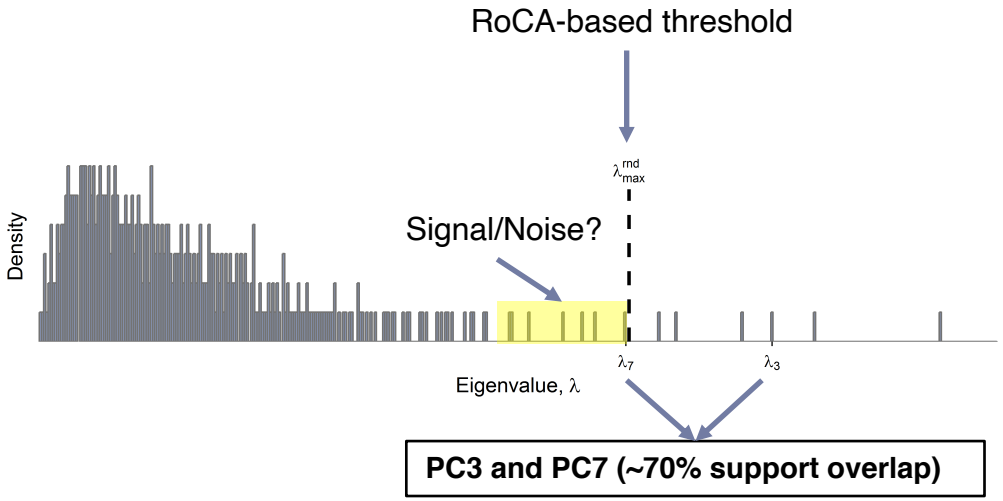
RocaSec: A GUI-based package for robust co-evolutionary analysis of proteins

- Standalone
- Cross-platform
- GUI-based
- Minimum input requirement
- Publication-quality figures output

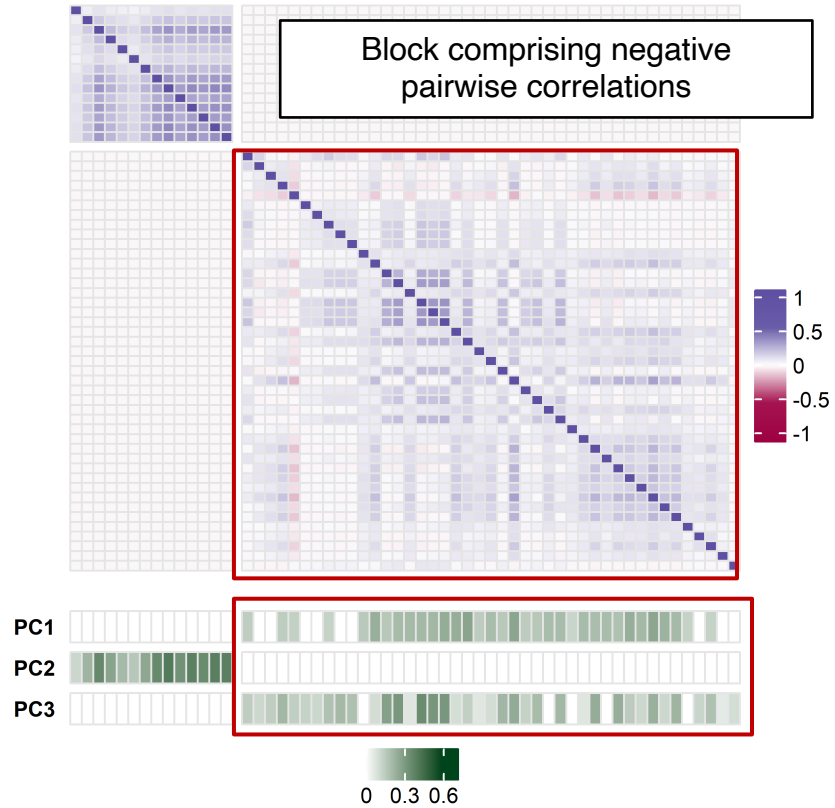
A. A. Quadeer, D. Morales-Jimenez, and M. R. M. "RocaSec: A GUI-based package for robust co-evolutionary analysis of proteins", to appear in *Bioinformatics*

Sub-dominant principal components of sample correlation matrices

Histogram of eigenvalues of HIV Gag sample correlation matrix



Ground-truth model of correlation matrix



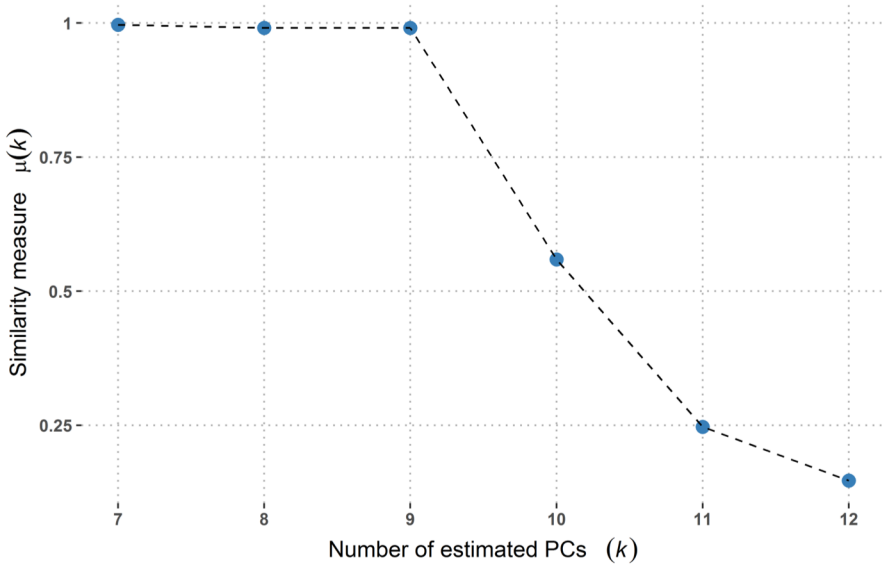
Overlapping PCs indicate a single underlying sector

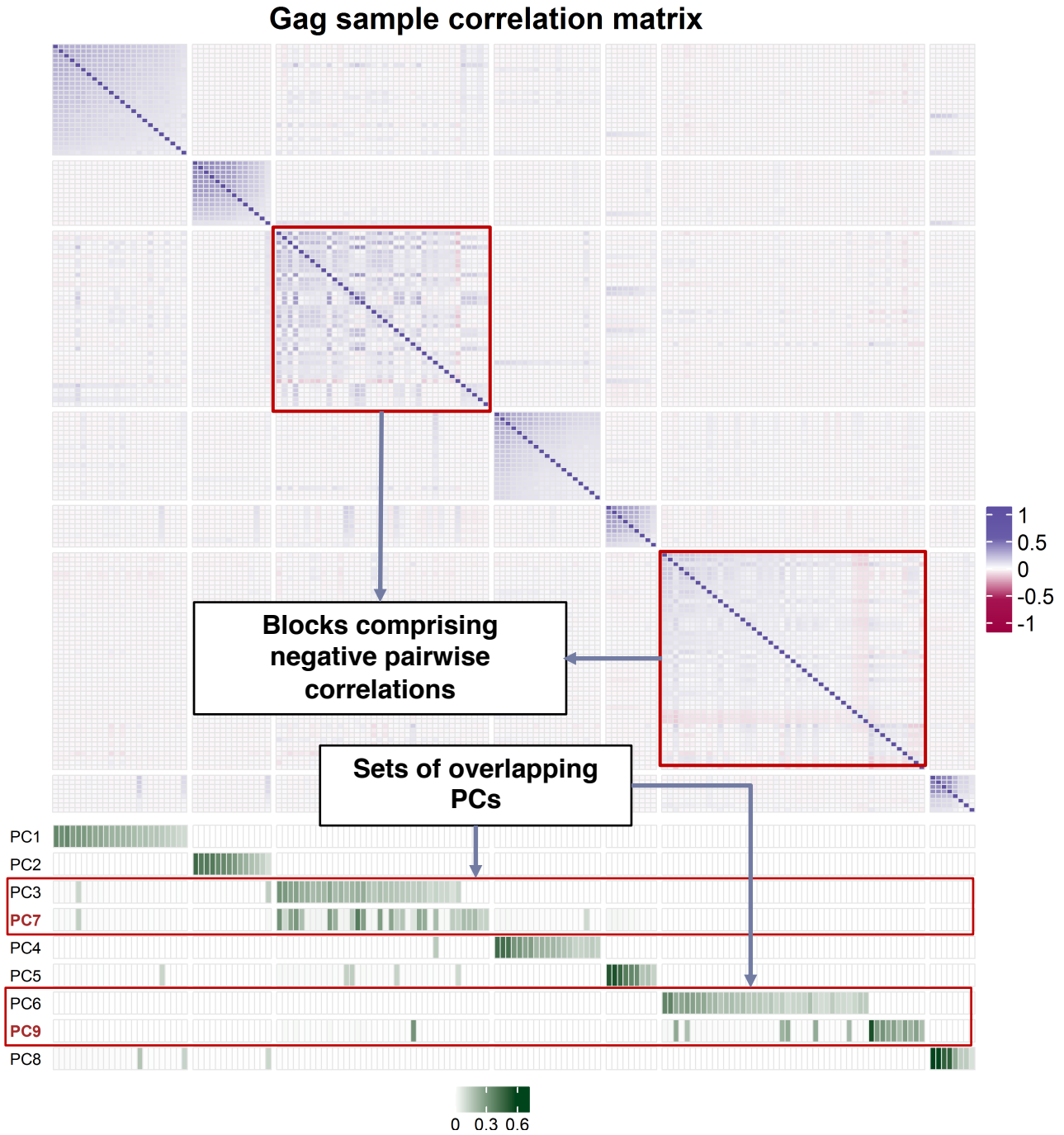
S. F. Ahmed, A. A Quadeer, D. Morales-Jimenez, and M. R M., "Sub-dominant principal components inform new vaccine targets for HIV Gag", Bioinformatics (**2019**)

HIV-Gag sample correlation matrix

- Which sub-dominant PCs to include for HIV-Gag sample correlation matrix?
  - Progressively include sub-dominant PCs and estimate subspace
  - Check similarity of the dominant (six) PCs as dimension of subspace increases

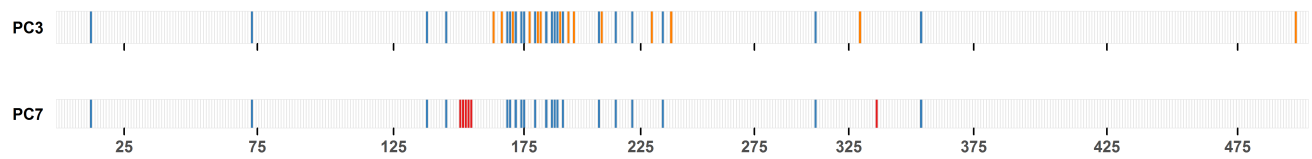
$$\mu(k) = \frac{1}{6}\sum_{i=1}^{6} \left|\langle \mathbf{v}_i^6, \mathbf{v}_i^k \rangle\right|$$

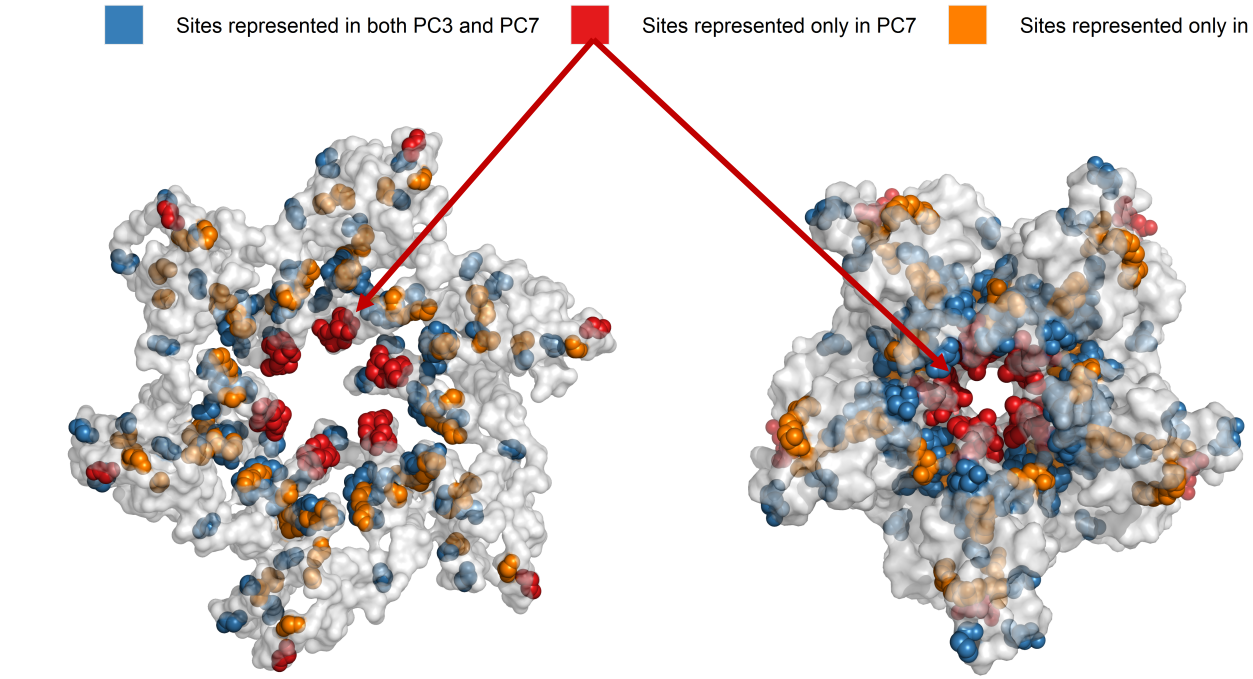


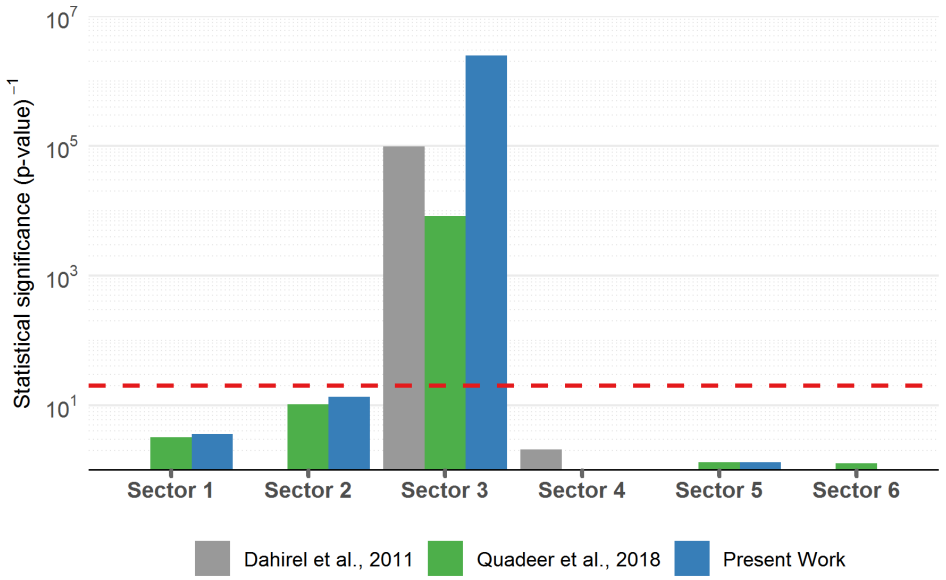


Biological/immunological significance of modified sector inferred in HIV-Gag

- A single sector inferred jointly from (overlapping) PC3 and PC7
  - Enriched in negative pairwise correlations
  - Picks up important protein interface sites
  - Strong association with known protective epitopes







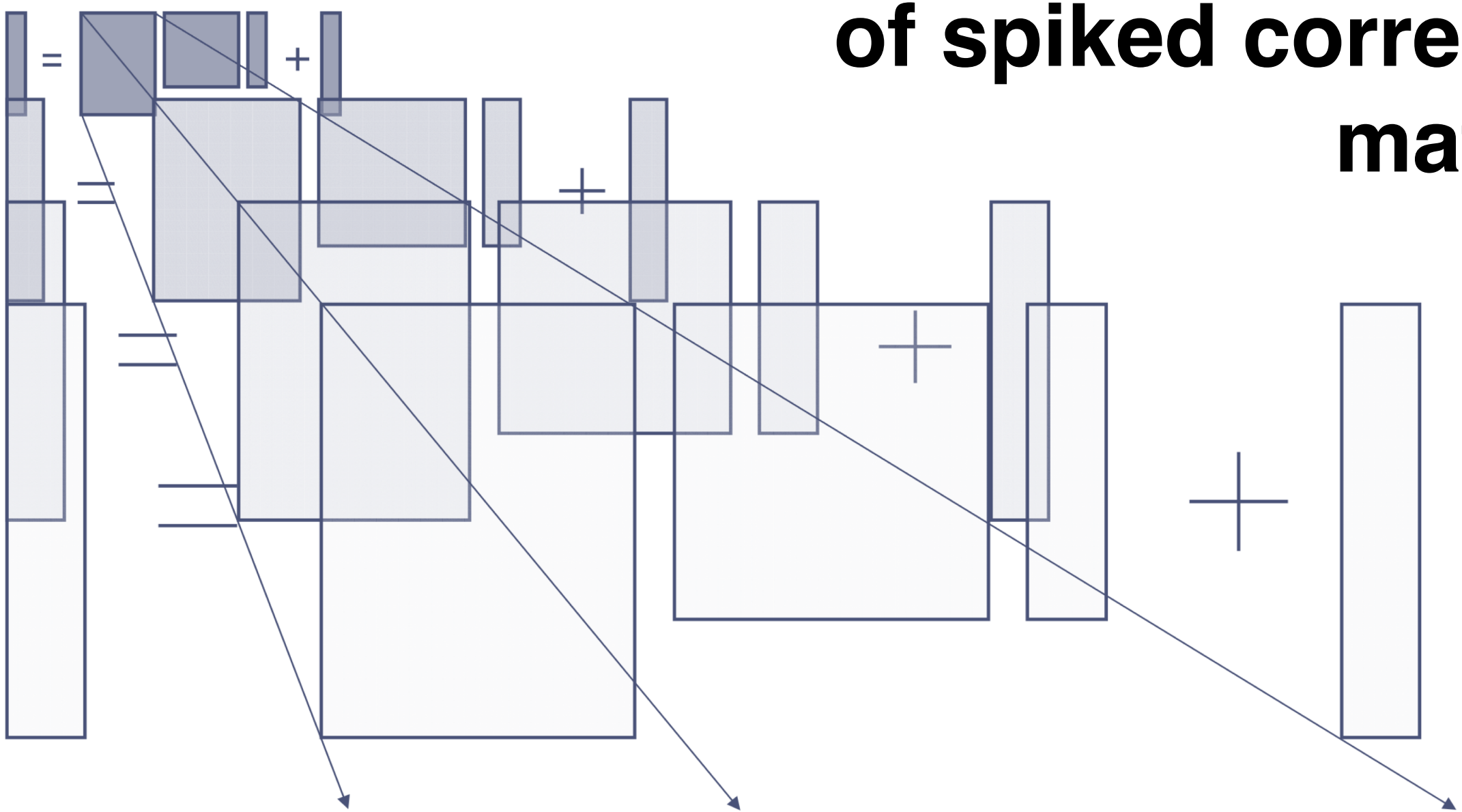
Proposed T cell-based vaccine design candidates

- HIV-Gag epitopes as targets
- Different set of epitopes compared to previous works with increased double coverage

Combination	Epitope 1	Epitope 2	Epitope 3	Epitope 4	Epitope 5	Dcov	L score
1	148-156	269-277	294-304	355-363	433-442	40.4%	0.39
2	148-156	269-277	306-316	355-363	433-442	40.4%	0.38
3	148-156	180-188	269-277	294-304	433-442	40.4%	0.37
4	180-188	269-277	294-304	355-363	433-442	40.4%	0.37
5	180-188	269-277	306-316	355-363	433-442	40.4%	0.34

S. F. Ahmed, A. A Quadeer, D. Morales-Jimenez, and M. R M., “Sub-dominant principal components inform new vaccine targets for HIV Gag”, Bioinformatics (**2019**)

Random matrix analysis of spiked correlation matrices

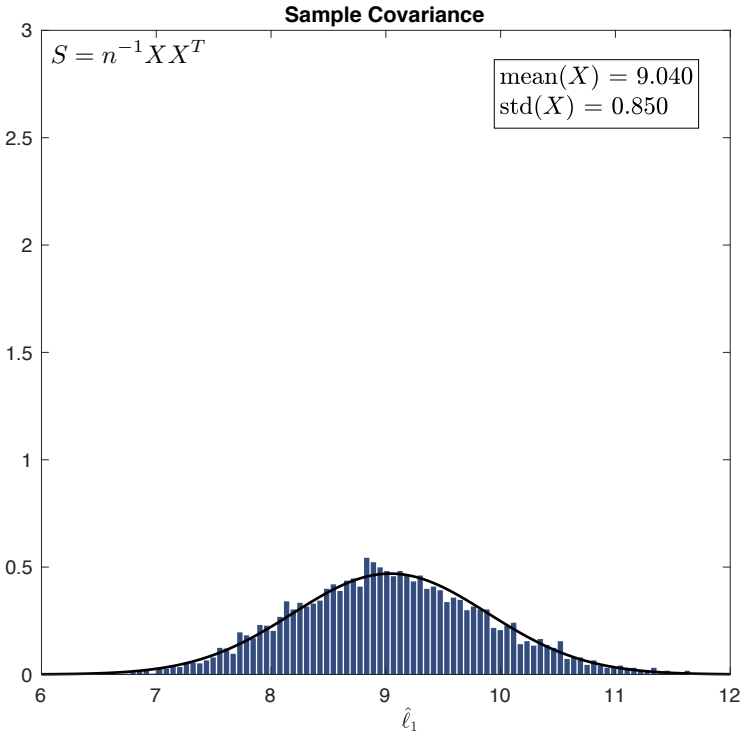


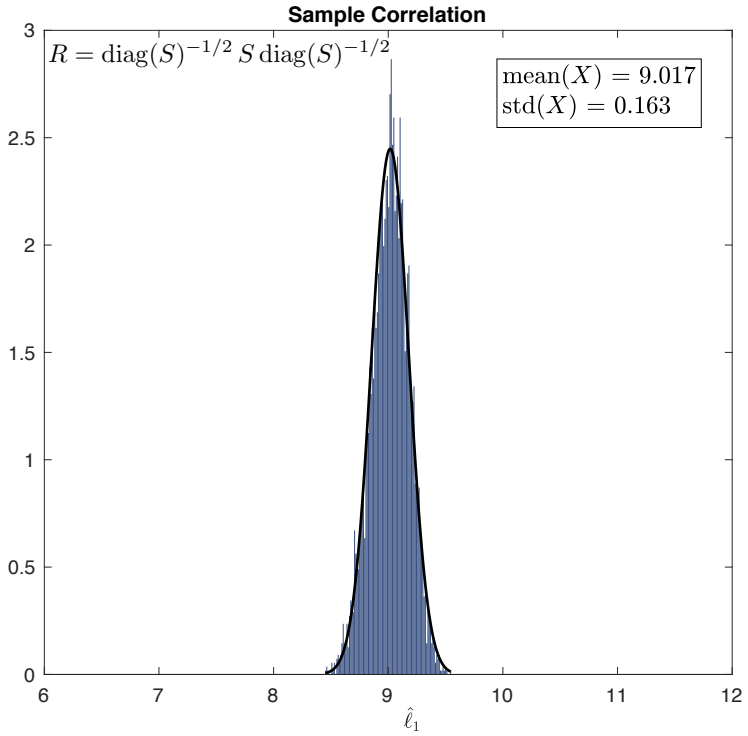
Spectral Analysis of Sample Correlation Matrices for Spiked Models

- Sample correlation matrices are widely used in practice
- Well studied under classical 'large n, fixed p' regimes (Girshick 1939, Konishi 1979, Fang and Krishnaiah 1982, …)
- For high-dimensional 'large n, large p' settings, recent progress for *null* models
  - Empirical spectrum (Jiang 2004)
  - Almost sure limits of largest and smallest eigenvalues (Jiang 2004, Xiao and Zhou 2010, Heiny et. al. 2018)
  - Tracy-Widom fluctuations (Bao et. al. 2012, Pillai and Yin 2012)
  - CLTs for linear statistics (Gao 2017)
- Unlike for covariance matrices, that have been very well studied, high-dimensional results for sample correlation matrices beyond null models are scarce (e.g., El Karoui 2009, Mestre and Vallet 2017).
- *Focus of this work:* Motivated by the biological application, we seek to study the spectral properties (eigenvalues and eigenvectors) of spiked correlation matrices.

D. Morales-Jimenez, I. M. Johnstone, M. R. M., and J. Yang, "Asymptotics of eigenstructure of sample correlation matrices for high-dimensional spiked models", to appear in *Statistica Sinica*.

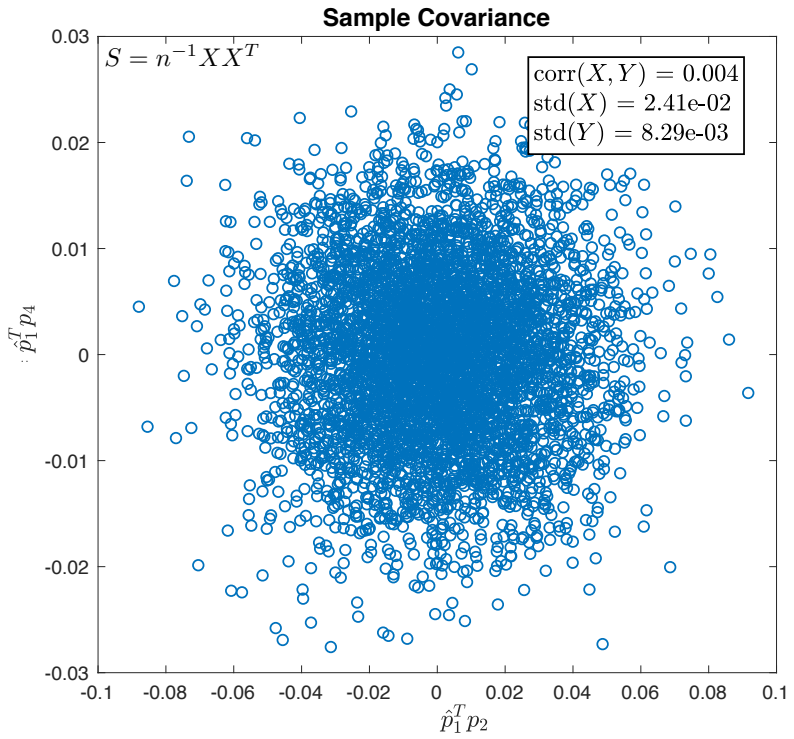
Numerical example: Largest eigenvalue distribution

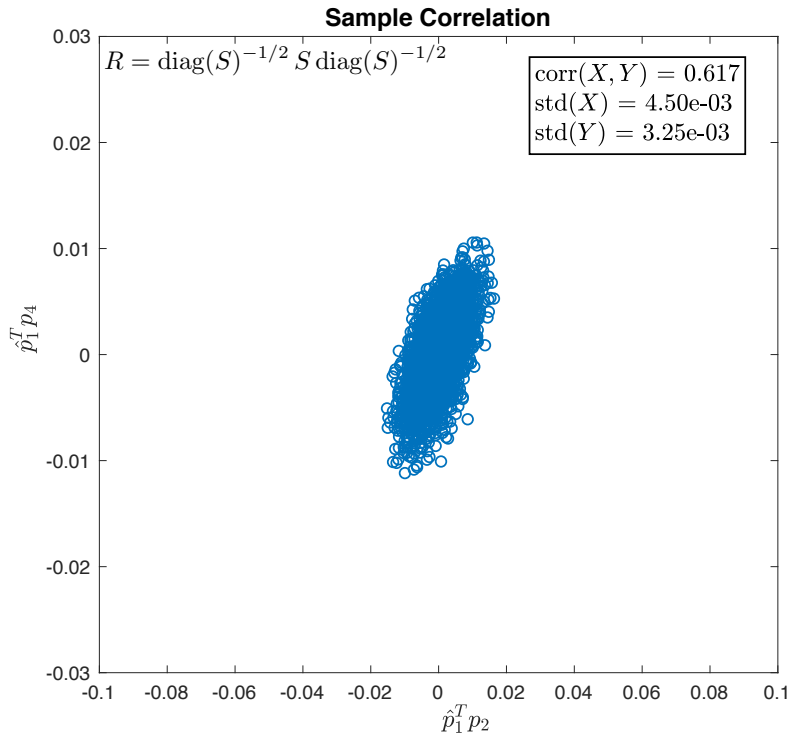




- $n = 200$ i.i.d. samples, $x_i \sim N(0, \Sigma)$. Covariance: $\Sigma = \mathrm{blkdiag}(\Sigma_s, I_{90})$, $(\Sigma_s)_{i,j=1}^{10} = \left(0.95^{|i-j|}\right)_{i,j=1}^{10}$

Numerical example: Eigenvector projections





- $n = 200$ i.i.d. samples, $x_i \sim N(0, \Sigma)$. Covariance: $\Sigma = \mathrm{blkdiag}(\Sigma_s, I_{90})$, $(\Sigma_s)_{i,j=1}^{10} = \left(0.95^{|i-j|}\right)_{i,j=1}^{10}$

Model M - Construction

- Let $x = \begin{bmatrix} \xi \\ \eta \end{bmatrix}$ have zero-mean entries with finite $(4+\delta)$th moment, for some $\delta > 0$.
- $\xi \in \mathbb{R}^m, \eta \in \mathbb{R}^p$, independent, $\quad \mathrm{Cov}\,(\xi) = \Sigma, \quad \mathrm{Cov}\,(\eta) = I$
- Correlation matrix of ξ: $\quad \Gamma \;=\; \mathrm{diag}(\Sigma)^{-1/2}\Sigma\,\mathrm{diag}(\Sigma)^{-1/2} \;=\; PLP^T$

  $$P = [p_1, \ldots, p_m], \quad L = \mathrm{diag}(\ell_1, \ldots, \ell_m), \quad \ell_1 \geq \cdots \geq \ell_m > 0$$

- Correlation matrix of x: $\quad \Gamma_x = \mathrm{blkdiag}(\Gamma, I)$

  Eigenvalues: $\quad \ell_1, \ldots, \ell_m, 1, \ldots, 1$

  Eigenvectors: $\quad \mathfrak{p}_1, \ldots, \mathfrak{p}_m, e_{m+1}, \ldots, e_{m+p}$ $\qquad$ Partition: $\quad \mathfrak{p}_i = [p_i^T \;\; 0_p^T]^T$

Model M - Estimation

- Data matrix: $X = [x_1, \ldots, x_n] \in \mathbb{R}^{(m+p)\times n}$, where x_i are i.i.d. copies of x
- Sample covariance: $S = n^{-1} X X^T$
- Sample correlation: $R = \mathrm{diag}(S)^{-1/2}\, S\, \mathrm{diag}(S)^{-1/2}$

Eigenvalues: $\hat{\ell}_1 \geq \cdots \geq \hat{\ell}_m \geq \hat{\ell}_{m+1} \geq \cdots \geq \hat{\ell}_{m+p} \geq 0$

Eigenvectors: $\hat{\mathfrak{p}}_1, \cdots, \hat{\mathfrak{p}}_m, \hat{\mathfrak{p}}_{m+1}, \cdots, \hat{\mathfrak{p}}_{m+p}$ Partition: $\hat{\mathfrak{p}}_\nu = [\hat{p}_\nu^T, \hat{v}_\nu^T]^T$

- Asymptotics: *m* fixed, while *p* and *n* grow large with $\gamma_n = p/n \to \gamma > 0$ as $p, n \to \infty$.
- For an index ν such that $\ell_\nu > 1 + \sqrt{\gamma}$ is a simple eigenvalue, define

$$\rho_\nu = \rho(\ell_\nu, \gamma)\,, \qquad \rho_{\nu n} = \rho(\ell_\nu, \gamma_n)\,, \qquad \dot{\rho}_\nu = \dot{\rho}(\ell_\nu, \gamma)\,, \qquad \dot{\rho}_{\nu n} = \dot{\rho}(\ell_\nu, \gamma_n)$$

where $\rho(\ell, \gamma) = \ell + \gamma \dfrac{\ell}{\ell - 1}\,, \qquad \dot{\rho}(\ell, \gamma) = \dfrac{\partial \rho(\ell, \gamma)}{\partial \ell} = 1 - \dfrac{\gamma}{(\ell - 1)^2}$

Asymptotic properties of spiked eigenvalues

- **Theorem 1:** Assume Model M, and that $\ell_\nu > 1 + \sqrt{\gamma}$ is a simple eigenvalue. As $p/n \to \gamma > 0$,

  i) $\hat{\ell}_\nu \xrightarrow{\text{a.s.}} \rho_\nu$

  ii) $\sqrt{n}(\hat{\ell}_\nu - \rho_{\nu n}) \xrightarrow{\mathcal{D}} N(0, \tilde{\sigma}_\nu^2)$

  where $\tilde{\sigma}_\nu^2 = 2\dot{\rho}_\nu \ell_\nu^2 + \dot{\rho}_\nu^2[\mathcal{P}^\nu, \kappa] + \dot{\rho}_\nu^2[\mathcal{P}^\nu, \check{\kappa}]$

  Gaussian, covariance | Non-Gaussian correction | Correlation correction (variance norm)

$$[\mathcal{P}^\nu, A] = \sum_{i,j,i',j'} P^\nu_{iji'j'} A_{iji'j'}$$

$$P^\nu_{iji'j'} = p_{\nu,i} p_{\nu,j} p_{\nu,i'} p_{\nu,j'}$$

$$\kappa_{iji'j'} = \mathbb{E}[\bar{\xi}_i \bar{\xi}_j \bar{\xi}_{i'} \bar{\xi}_{j'}] - \kappa_{ij}\kappa_{i'j'} - \kappa_{ij'}\kappa_{ji'} - \kappa_{ii'}\kappa_{jj'}$$

$$\kappa_{ij} = \mathbb{E}\bar{\xi}_i \bar{\xi}_j \qquad \bar{\xi}_i = \xi_i / \sigma_i$$

- **Corollary 1:** For Gaussian data, the asymptotic variance simplifies to

$$\tilde{\sigma}_\nu^2 = 2\ell_\nu^2 \dot{\rho}_\nu \left[1 - \dot{\rho}_\nu \left(2\ell_\nu \text{tr} P_{D,\nu}^4 - \text{tr}(P_{D,\nu} \Gamma P_{D,\nu})^2\right)\right]$$

Often positive

where $P_{D,\nu} = \text{diag}(p_{\nu,1}, \ldots, p_{\nu,m})$

Asymptotic properties of spiked eigenvalues

- **Theorem 1:** Assume Model M, and that $\ell_\nu > 1 + \sqrt{\gamma}$ is a simple eigenvalue. As $p/n \to \gamma > 0$,

  i) $\hat{\ell}_\nu \xrightarrow{\text{a.s.}} \rho_\nu$

  ii) $\sqrt{n}(\hat{\ell}_\nu - \rho_{\nu n}) \xrightarrow{\mathcal{D}} N(0, \tilde{\sigma}_\nu^2)$

  where $\tilde{\sigma}_\nu^2 = 2\dot{\rho}_\nu \ell_\nu^2 + \dot{\rho}_\nu^2[\mathcal{P}^\nu, \kappa] + \dot{\rho}_\nu^2[\mathcal{P}^\nu, \check{\kappa}]$

  Gaussian, covariance | Non-Gaussian correction | Correlation correction (variance norm)

$$[\mathcal{P}^\nu, A] = \sum_{i,j,i',j'} P^\nu_{iji'j'} A_{iji'j'}$$

$$P^\nu_{iji'j'} = p_{\nu,i} p_{\nu,j} p_{\nu,i'} p_{\nu,j'}$$

$$\check{\kappa}_{iji'j'} = \mathrm{Cov}(\psi_{ij}, \psi_{i'j'}) - \mathrm{Cov}(\psi_{ij}, \chi_{i'j'}) - \mathrm{Cov}(\chi_{ij}, \psi_{i'j'})$$

$$\chi_{ij} = \bar{\xi}_i \bar{\xi}_j \quad \psi_{ij} = \kappa_{ij}(\bar{\xi}_i^2 + \bar{\xi}_j^2)/2$$

- **Corollary 1:** For Gaussian data, the asymptotic variance simplifies to

$$\tilde{\sigma}_\nu^2 = 2\ell_\nu^2 \dot{\rho}_\nu \left[1 - \dot{\rho}_\nu \left(2\ell_\nu \mathrm{tr} P_{D,\nu}^4 - \mathrm{tr}(P_{D,\nu} \Gamma P_{D,\nu})^2\right)\right]$$

Often positive

where $P_{D,\nu} = \mathrm{diag}(p_{\nu,1}, \ldots, p_{\nu,m})$

Asymptotic properties of spiked eigenvectors

- **Theorem 2:** Assume Model M, and that $\ell_\nu > 1 + \sqrt{\gamma}$ is a simple eigenvalue. As $p/n \to \gamma > 0$,

  i) $\langle \hat{\mathfrak{p}}_\nu, \mathfrak{p}_\nu \rangle^2 \xrightarrow{\text{a.s.}} \dot{\rho}_\nu \ell_\nu / \rho_\nu$

  ii) $\sqrt{n}(P^T \hat{p}_\nu / \|\hat{p}_\nu\| - e_\nu) \xrightarrow{\mathcal{D}} N(0, \Sigma_\nu)$

  where $\Sigma_\nu = \mathcal{D}_\nu \tilde{\Sigma}_\nu \mathcal{D}_\nu$ with $\mathcal{D}_\nu = \sum_{k \neq \nu}^{m} (\ell_\nu - \ell_k)^{-1} e_k e_k^T$

  $$\tilde{\Sigma}_{\nu,kl} = \underbrace{\dot{\rho}_\nu^{-1} \ell_k \ell_\nu \, \delta_{k,l}}_{\text{Gaussian, covariance}} + \underbrace{[\mathcal{P}^{k\nu l\nu}, \kappa]}_{\substack{\text{Non-Gaussian} \\ \text{correction}}} + \underbrace{[\mathcal{P}^{k\nu l\nu}, \check{\kappa}]}_{\substack{\text{Correlation} \\ \text{correction}}}$$

  $$P^{k\nu l\nu}_{iji'j'} = p_{k,i} p_{\nu,j} p_{l,i'} p_{\nu,j'}$$

- **Corollary 2:** For Gaussian data, the asymptotic covariance simplifies to $\Sigma_\nu = \mathcal{D}_\nu \tilde{\Sigma}_\nu \mathcal{D}_\nu$,

  $$\tilde{\Sigma}_\nu = \frac{\ell_\nu}{\dot{\rho}_\nu} L + (\ell_\nu I + L) \left(\tfrac{1}{2} \mathcal{Z} - \ell_\nu \mathcal{Y} \right) (\ell_\nu I + L) + \ell_\nu (\ell_\nu^2 \mathcal{Y} - L \mathcal{Y} L)$$

  Non-diagonal (Asymptotic dependencies)

  where $\mathcal{Z} = P^T P_{D,\nu} (\Gamma \circ \Gamma) P_{D,\nu} P$ and $\mathcal{Y} = P^T P_{D,\nu}^2 P$

Subcritical case

- **Theorem 3:** Assume Model M, and that $1 < \ell_\nu \leq 1 + \sqrt{\gamma}$ is a simple eigenvalue. As $p/n \to \gamma > 0$,

  i) $\hat{\ell}_\nu \xrightarrow{\text{a.s.}} (1 + \sqrt{\gamma})^2$

  ii) $\langle \hat{\mathfrak{p}}_\nu, \mathfrak{p}_\nu \rangle^2 \xrightarrow{\text{a.s.}} 0$

- Derivations leverage technical approaches from (Paul, 2007) and (Bai and Yao, 2008)

- Results are "correlation" versions of a set of theorems in (Paul, 2007), which considered covariance matrices and Gaussian data
- Parallel results (with consistent notation) given for covariance matrices and non-Gaussian data in the companion review notes:

  I. M. Johnstone and J. Yang, "Notes on asymptotics of sample eigenstructure for spiked covariance models with non-Gaussian data". ArXiv: 1810.10427, 2019.

Setup (Standard steps)

- Similar to the analysis of spiked covariance models (e.g., Bai and Yao, 2008), the derivations involve computing deterministic (almost sure) limits and CLTs for matrix-valued quadratic forms.

- Define the partition: $R = \mathrm{diag}(S)^{-1/2}\, S\, \mathrm{diag}(S)^{-1/2} = n^{-1} \begin{bmatrix} \bar{X}_1 \bar{X}_1^T & \bar{X}_1 \bar{X}_2^T \\ \bar{X}_2 \bar{X}_1^T & \bar{X}_2 \bar{X}_2^T \end{bmatrix}$

- Partitioning of eigenvector equation $R\,\hat{\mathfrak{p}}_\nu = \hat{\ell}_\nu \hat{\mathfrak{p}}_\nu$ along with $\hat{\mathfrak{p}}_\nu = [\hat{p}_\nu^T, \hat{v}_\nu^T]^T$ gives:

$$\bar{X}_1 \bar{X}_1^T \hat{p}_\nu + \bar{X}_1 \bar{X}_2^T \hat{v}_\nu = \hat{\ell}_\nu \hat{p}_\nu$$
$$X_2 \bar{X}_1^T \hat{p}_\nu + \bar{X}_2 \bar{X}_2^T \hat{v}_\nu = \hat{\ell}_\nu \hat{v}_\nu$$

- From the second equation $\hat{v}_\nu = (\hat{\ell}_\nu I_p - \bar{X}_2 \bar{X}_2^T)^{-1} \bar{X}_2 \bar{X}_1^T \hat{p}_\nu$, which gives for the first equation:

$$K(\hat{\ell}_\nu)\hat{p}_\nu = \hat{\ell}_\nu \hat{p}_\nu \qquad K(t) = \bar{X}_1 \bar{X}_1^T + \bar{X}_1 \bar{X}_2^T (t I_p - \bar{X}_2 \bar{X}_2^T)^{-1} \bar{X}_2 \bar{X}_1^T$$

  - Hence, $\hat{\ell}_\nu$ is an eigenvalue of $K(\hat{\ell}_\nu)$ with associate eigenvector $\hat{p}_\nu$

Technical challenge – (Normalized) matrix quadradic forms

- The derivations involve computing deterministic (almost sure) limits and CLTs for the matrix-valued quadratic form:

$$K(t) = n^{-1}\bar{X}_1 B_n(t)\bar{X}_1^T \qquad B_n(t) = I_n + n^{-1}\bar{X}_2^T(tI_p - n^{-1}\bar{X}_2^T\bar{X}_2)^{-1}\bar{X}_2$$
$$= t(tI_n - n^{-1}\bar{X}_2^T\bar{X}_2)^{-1}$$

- The main difference to the covariance case is that the data matrix is *variance-normalized*, creating new dependencies

- CLT result: As $p/n \to \gamma > 0$,

$$\sqrt{n}\left[K(\rho_{\nu n}) - n^{-1}\mathrm{tr}B_n(\rho_{\nu n})\Gamma\right] \xrightarrow{\mathcal{D}} W^\nu$$

where W^ν is a symmetric Gaussian matrix with entries W_{ij}^ν having zero-mean and covariance

$$\mathrm{Cov}[W_{ij}^\nu, W_{i'j'}^\nu] = \frac{\rho_\nu^2}{\ell_\nu^2\dot{\rho}_\nu}\left(\kappa_{ij'}\kappa_{ji'} + \kappa_{ii'}\kappa_{jj'}\right) + \frac{\rho_\nu^2}{\ell_\nu^2}\left(\kappa_{iji'j'} + \check{\kappa}_{iji'j'}\right)$$

  - Derivation involves extension of a martingale CLT approach of Baik and Silverstein (in Appendix of Capitaine, Donati-Martin, Feral 2009)

D. Morales-Jimenez, I. M. Johnstone, M. R. M., and J. Yang, "Asymptotics of eigenstructure of sample correlation matrices for high-dimensional spiked models", to appear in *Statistica Sinica*.

Other application involving high-dimensional covariance matrices

Maximum entropy model inference using MPF-BML method

Available E2 data

...ASHEGSFTNPARC...
...ASDEGSFTNPARC...
...ASDEGSFTNPARC...
...ASDEGSFTNPARC...
...ASDEGSFTNPARC...
...ASDEGSFTNPARC...
...ASDEGSFTNPARC...
...ASDEGSFTNPARC...
...ASDEGSFTNPARC...
...ASDEGSFTNVARC...

Number of sequences (samples):
~10^3 – 10^4
Number of protein residues (variables):
~10^2 – 10^3

Statistical model:
Maximum entropy (prevalence) model

Fitness Prevalence

$$f(\mathbf{x}) \sim p(\mathbf{x}) = \frac{\exp[-E(\mathbf{x})]}{Z}$$

$$E(\mathbf{x}) = \sum_{i=1}^{L} h_i(x_i) + \sum_{i=1}^{L} \sum_{j=i+1}^{L} J_{ij}(x_i, x_j), \quad Z = \sum_{\mathbf{x}'} \exp[-E(\mathbf{x}')]$$

Maximum entropy formulation

$$\max S = -\sum_{\mathbf{x}} p(\mathbf{x}) \log p(\mathbf{x})$$

$$\text{s.t.} \sum_{\mathbf{x}} p(\mathbf{x}) = 1$$

$$\sum_{\mathbf{x}} p(\mathbf{x}) \delta(x_i, a) = p_i^{obs}(a)$$

$$\sum_{\mathbf{x}} p(\mathbf{x}) \delta(x_i, a) \delta(x_i, b) = p_{ij}^{obs}(a, b)$$

...ASDEGSFT...
...ASDEGSFT...
...ASDEGSFT...
...ASDEGSFT...
...ASDEGSFT...

...ASDEGSFT...
...ASDEGSFT...
...ASDEGSFT...
...ASDEGSFT...
...ASDEGSFT...

Solution:

$$(\mathbf{h}^*, \mathbf{J}^*) = \arg\min_{\mathbf{h},\mathbf{J}} \mathrm{KL}\left(p_0 \,||\, p_{\mathbf{h},\mathbf{J}}\right)$$

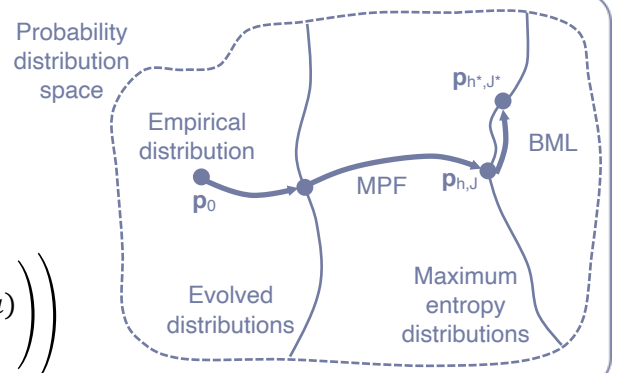
...ASNEGSFTNPARC...
...ASDEGSFTNPARC...
...ASDEGSFTNPARC...
...ASDEGSFTNPARC...
...ASDEGSFTNPARC...
...ASDEGSFTNPARC...
...ASDEGSFTNPARC...
...ASDEGSFTNPARC...
...ASDEGSFTNPARC...
...ASDEGSFTNVARC...

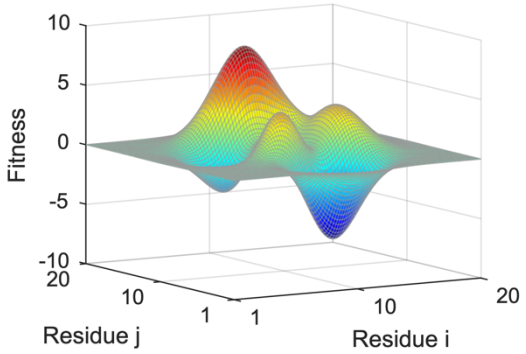
Sequence data → Maximum entropy model → Predicted fitness

Framework for model inference (**MPF-BML**)

$$(\mathbf{h}_{\mathrm{MPF}}, \mathbf{J}_{\mathrm{MPF}}) = \underset{\mathrm{h,J}}{\operatorname{argmin}} \mathrm{KL}(p_0 || p_{\mathrm{h,J;t}}) + \lambda_1 |\mathbf{J}|_1 + \lambda_2 |\mathbf{J}|_2$$

$$\mathrm{KL}(p_0 || p_{\mathrm{h,J;t}}) \approx \sum_{b,i,a} \exp\left(\frac{1}{2}\left((2y_{b,(i-1)L+a} - 1)\sum_{j,b} y_{b,(j-1)L+b} J_{ij}(a,b) - h_i(a)\right)\right)$$

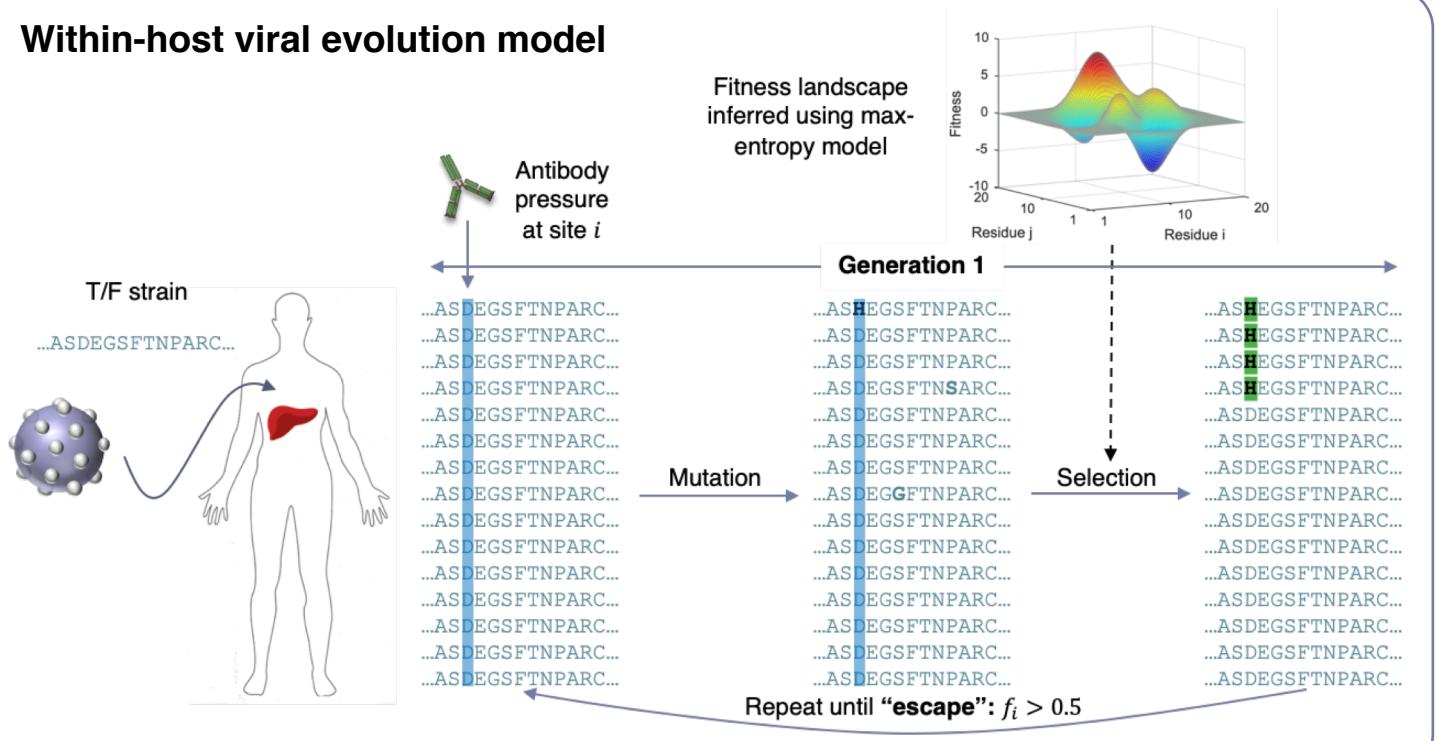




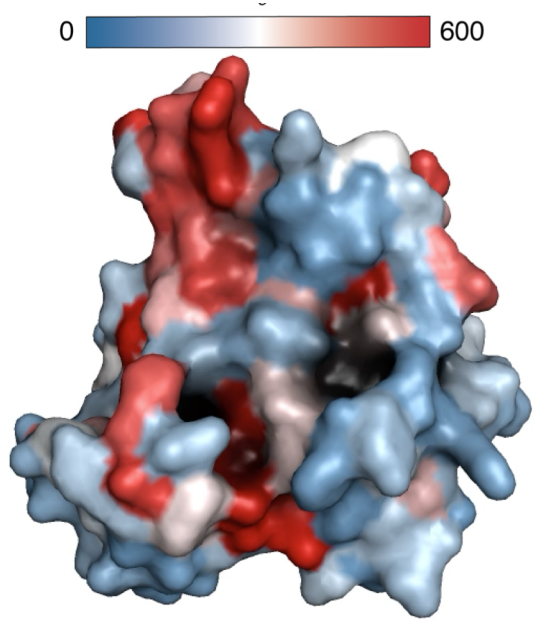
R. H. Y. Louie, K. J. Kaczorowski, J. P. Barton, A. K. Chakraborty, M. R. M., "Fitness landscape of the human immunodeficiency virus envelope protein that is targeted by antibodies," *PNAS*, 115 (**2018**)

Using maximum entropy model to **identify escape-resistant Hepatitis C antibodies**

Within-host viral evolution model

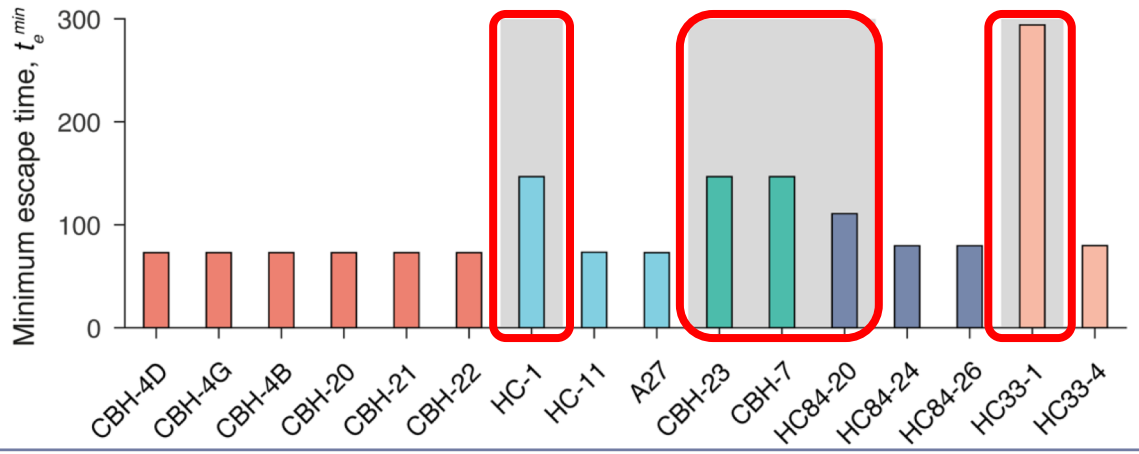


Mean escape time



Protein data bank, https://www.rcsb.org/
(PDB ID: **4MWF**)

Escape resistance of antibodies



A. A. Quadeer, R. H. Y. Louie, and M. R. M., "Identifying immunologically-vulnerable regions of the HCV E2 glycoprotein and broadly neutralizing antibodies that target them", *Nature Communications* 10 **(2019)**.

Acknowledgements

HKUST	**Ahmed Abdul Quadeer** **Syed Faraz Ahmed** Saqib Sohail
QUB	**David Morales-Jimenez**
Stanford	**Iain Johnstone, Jeha Yang**
MIT	Arup Chakraborty, Kevin Kaczorowski
UC-Riverside	John Barton
UNSW	Raymond Louie
Berkeley	Karthik Shekhar

Thanks also to:

Cambridge	Chris Illingworth
UNSW	Rowena Bull
Hong Kong U.	Leo Poon

Funding: General Research Fund of Hong Kong RGC