Title

HIV-1 subtypes maintain distinctive physicochemical signatures in Nef domains associated with immunoregulation.

Authors

Lamers, Susanna
Fogel, Gary
Liu, Enoch
et al.

Publication Date

2023-11-01

DOI

10.1016/j.meegid.2023.105514

Peer reviewed

HHS Public Access

Author manuscript

Infect Genet Evol. Author manuscript; available in PMC 2024 February 05.

Published in final edited form as:

Infect Genet Evol. 2023 November ; 115: 105514. doi:10.1016/j.meegid.2023.105514.

HIV-1 subtypes maintain distinctive physicochemical signatures in Nef domains associated with immunoregulation

Susanna L. Lamers[a,*], **Gary B. Fogel**[b], **Enoch S. Liu**[b], **David J. Nolan**[a], **Rebecca Rose**[a], **Michael S. McGrath**[c]

[a]BioInfoExperts , Thibodaux, Louisiana, USA

[b]Natural Selection, San Diego, California, USA

[c]University of California, San Francisco, California, USA

Abstract

Background: HIV subtype is associated with varied rates of disease progression. The HIV accessory protein, Nef, continues to be present during antiretroviral therapy (ART) where it has numerous immunoregulatory effects. In this study, we analyzed Nef sequences from HIV subtypes A1, B, C, and D using a machine learning approach that integrates functional amino acid information to identify if unique physicochemical features are associated with Nef functional/structural domains in a subtype-specific manner.

Methods: 2253 sequences representing subtypes A1, B, C, and D were aligned and domains with known functional properties were scored based on amino acid physicochemical properties. Following feature generation, we used statistical pruning and evolved neural networks (ENNs) to determine if we could successfully classify subtypes. Next, we used ENNs to identify the top five key Nef physicochemical features applied to specific immunoregulatory domains that differentiated subtypes. A signature pattern analysis was performed to the assess amino acid diversity in sub-domains that differentiated each subtype.

Results: In validation studies, ENNs successfully differentiated each subtype at A1 (87.2%), subtype B (89.5%), subtype C (91.7%), and subtype D (85.1%). Our feature-based domain scoring, followed by *t*-tests, and a similar ENN identified subtype-specific domain-associated features. Subtype A1 was associated with alterations in Nef CD4 binding domain; subtype B was associated with alterations with the AP-2 Binding domain; subtype C was associated with alterations in a structural Alpha Helix domain; and, subtype D was associated with alterations in a Beta-Sheet domain.

[*]Corresponding author at: CEO, BioInfoExperts LLC, 718 Bayou Lane, Thibodaux, LA 70301, United States of America. susanna@bioinfox.com (S.L. Lamers).

Declaration of Competing Interest

Susanna Lamers, David Nolan, and Rebecca Rose are employed by BioInfoExperts LLC. Gary Fogel and Enoch Liu are employed by Natural Selection Inc. No other competing interests exist.

CRediT authorship contribution statement

Susanna L. Lamers: Project administration, Conceptualization, Methodology. **Gary B. Fogel:** Conceptualization, Methodology. **Enoch S. Liu:** Conceptualization, Methodology, Writing – review & editing. **David J. Nolan:** Writing – review & editing. **Rebecca Rose:** Writing – review & editing. **Michael S. McGrath:** Project administration, Writing – review & editing, Funding acquisition.

Conclusions: Recent studies have focused on HIV Nef as a driver of immunoregulatory disease in those HIV infected and on ART. Nef acts through a complex mixture of interactions that are directly linked to the key features of the subtype-specific domains we identified with the ENN. The study supports the hypothesis that varied Nef subtypes contribute to subtype-specific disease progression.

Keywords

HIV; Nef protein; Machine-learning; Bioinformatics; HIV subtypes; Sequence analysis

1. Introduction

Despite more than three decades of research and improved antiretroviral therapies (ART), low-level immune activation during HIV infection continues to promote immunoregulatory-associated disease processes (Lamers et al., 2012; Sereti et al., 2017). These diseases include a variety of cancers, metabolic issues, and HIV-associated neurological disorders. While AIDS-defining diseases have seen a dramatic reduction in the modern-treatment era, other non-AIDS-defining diseases are on the rise in HIV-infected individuals (Shiels and Engels, 2017).

Current antiretroviral agents used to treat HIV-1 infection target several different pathways and have varied mechanisms of action (Gandhi et al., 2023). When used in combination, these drugs have been extraordinarily successful in the long-term management of HIV infection and limit transmission; however, they do not clear latent reservoirs (Pereira and daSilva, 2016; Chun and Fauci, 2012). HIV integrates into the genomes of tissue-based macrophages and T-cells (Hendricks et al., 2021; Shacklett et al., 2019), where a lower cell turnover rate combined with lower drug penetration enables HIV to persist in reservoirs (Cohn et al., 2020; Sung and Margolis, 2018). Furthermore, even with ART-reduced viral load, copies of HIV mRNA transcripts that may be translated to protein persist (e.g., Nef and Tat), which provides a mechanism for chronic inflammation, immune activation, and ongoing immunoregulatory interference (Nolan et al., 2022; Fischer et al., 2002).

Nef is an HIV accessory protein with several intracellular functions that prime host cells for HIV replication and ensures the infectivity of progeny virions (Kirchhoff et al., 1995; Deacon et al., 1995). Nef downregulates CD4, MHC-1, Tetherin and SERINC at the trans Golgi apparatus and at the interior cell surface membrane (Kwon et al., 2020). These functions are regulated through the interactions of Nef with >30 other partner proteins in both membrane-bound and cytoplasmic Nef (Fackler and Baur, 2002; Arold and Baur, 2001). Recent studies have shown that Nef is found in circulating exosomes (Mukhamedova et al., 2019; McNamara et al., 2018; Khan et al., 2016; Puzar Dominkus et al., 2017) indicating that Nef also acts extracellularly within the vesicular trafficking machinery. Extracellular Nef has the potential to alter host gene expression and dysregulate the host immune response in bystander cells (Felli et al., 2017). Intracellularly and extracellularly, the molecular actions of Nef align with the pathogenic effects of a lengthy infection in an individual on ART (Olivetta et al., 2016; Zhu et al., 2014).

Given the many immunoregulatory functions of Nef, there is increased interest in Nef-targeted therapies (Painter et al., 2020; Bouchet et al., 2011; Kamalzare et al., 2019; Lv et al., 2020). For instance, a successful drug candidate would limit the ability of Nef to downregulate cellular factors, promoting the clearance of infected cells while simultaneously reducing extracellular Nef trafficking. However, implementing any therapy targeting Nef requires an improved understanding of its sequence diversity across HIV subtypes and the role of this diversity in protein-protein interactions. To date, most Nef-associated molecular modeling studies have typically suffered from a limited focus on a single molecular clone of HIV subtype B Nef (NL.43) (Kwon et al., 2020; Shen et al., 2015). While this is an essential strategy in understanding molecular docking to Nef, such a restricted view avoids sequence/structural Nef diversity altogether, providing information about docking for only one subtype, and within that subtype, only one specific sequence.

Globally, HIV is divided into twelve major subtypes that continue to evolve and recombine, resulting in well-defined hybrid or “recombinant” forms. HIV subtypes are defined by their genotype and can be visualized through independent branching in a phylogenetic tree. Temporal studies in Africa have demonstrated that HIV disease progression also varies among the major infecting subtypes, which coincides with changes in the rates of infection with different subtypes (Conroy et al., 2010; Lamers et al., 2020; Blanquart et al., 2016). During the 1990s in Uganda, HIV subtype D represented the majority of infections; however, individuals infected with subtype D progressed faster to AIDS than another circulating infecting subtype, subtype A (Lamers et al., 2020; Collinson-Streng et al., 2009). Over an 18-year time frame, the proportion of subtype D infections in Uganda decreased from an average of 71% to 40%, whereas subtype A increased from an average of 21% to 31% in the six communities (Lamers et al., 2020). Another study demonstrated that rapid progression with subtype D infection is associated with rapid development of high viral loads (Amornkul et al., 2013). A similar study in Brazil, where subtypes B, D, and F1 cocirculate, found that subtype D and a recombinant form of B + F1 resulted in more rapid disease progression (Leite et al., 2017).

One of the problems in understanding the impacts of HIV diversity is that it is difficult to distill vast amounts of genetic data into single experiments. Using a machine learning approach called the Zoetic Amino Acid Protein Profiler (ZAPP), we previously linked specific physicochemical amino acid signatures in Nef to diseased and non-diseased tissues to provide an improved understanding and prediction of HIV progression to neurological disease (Lamers et al., 2018; Liu et al., 2020). ZAPP combines amino acid sequence data, functional domain constraints, and amino acid physicochemical scales derived from biochemistry studies regarding the behavior of amino acids under certain conditions (e.g., shape, polarity, hydrophobicity scales). Features that survive statistical pruning were provided as input to neural networks optimized on training data using evolutionary computation. This same process of neural network optimization also includes a further process of feature reduction that results in small sets of features and their associated neural networks that can be used to help predict outcomes well and yet still be explainable. The best feature sets and their associated domains can then be examined independently and understood more carefully in terms of defined outgroups.

In this study, we used ZAPP to examine Nef functional domains and their relationship to predominant Ugandan HIV subtypes (A1, C, and D) and subtype B, the predominant HIV subtype in Western countries. Our goal was an improved understanding of the structural/functional variation among Nef subtypes from a biochemical and structural perspective as this variation may be associated with varied disease progression and/or assist in the development of Nef-targeted therapies.

2. Materials and methods

2.1. Sequences

As described in Fig. 1, 2343 Nef gene nucleotide sequences from four subtypes were obtained from the Los Alamos HIV database (subtype A1 (n = 579); subtype B (n = 631); subtype C (n = 704); subtype D (n = 429)) (Apetrei et al., 2021). Search criteria included subtype and country designation. Subtype A1 and D sequences originated from Uganda, subtype C sequences originated from Uganda and nearby East African countries, and subtype B sequences were derived from the United States. Subtypes associated with these sequences were confirmed using CONtext-based Modeling for Expeditious Typing, or COMET (Struck et al., 2014), which is an alignment-free typing algorithm for HIV-1 and other viruses. Any potential recombinant sequences, identical sequences, or those that were discordant with the Los Alamos Database designation were removed, leaving a total of 2253 sequences. All nucleotide sequences were translated to Nef protein amino acid sequences and aligned using Geneious Prime software (version 2021.2.2) with manual modifications adhering to a previously described alignment protocol (Lamers et al., 2018). Sixteen well-characterized domains in Nef from the literature were identified. These were Alpha Helix A, Alpha Helix B, Alpha Helix CD, AP-2 Binding, Positions Associated with Molecular Signaling, Beta-Sheet A, Beta-Sheet B, CD4 Downregulation, Cytokine Binding, MHC-1 Downmodulation I (MHC-1 DM1), MHC1 Downmodulation II (MHC-1 DM2), MHC-1 Association with Signaling Molecules (MHC1-signaling), Myristoylation, Nef Loop, Dileucine binding motif, and Sh-3 Binding. The positions and variations within the Nef aligned sequences are represented graphically in Fig. 2.

2.2. Feature generation

For each amino acid in the 16 Nef domains we calculated 70 physicochemical characteristics. These 70 characteristics were collected from sources in the literature (Lamers et al., 2016) and are representative of six broad classes, including amino acid size, shape or structure (n = 25), polarity (n = 6), composition (n = 3), hydrophobicity (n = 27), and other features such as HPLC and pKa (n = 9) (Table 1).

In order to generate domain-level scores, for each domain and for each characteristic we summed the per-amino acid scores across each domain as a single value. For instance, one characteristic such as “Molecular Weight” applied to the domain “Alpha Helix A" would be the summed score of the value of Molecular Weight for each amino acid within the Alpha Helix A domain. This process was repeated for all 16 Nef domains over all 70 characteristics generating 1120 domain-characteristics (termed “features”) gathered from Nef sequences representing subtypes A1, C, D and B. Initial box-and-whisker plots were

used to qualitatively observe differences across the four HIV-1 subtypes over these 1120 features.

2.3. Evolved neural networks

Initially, we used evolved neural networks to determine if there was sufficient information represented in the 1120 features for reasonable classification across all four subtypes simultaneously (A1 vs. B vs. C vs. D) without feature reduction. We divided the 2253 Nef sequences from subtypes A1, B, C, and D into three separate training, testing, and validation sets. A subset of 10% (n = 225) of the total data was chosen at random across all four subtypes as a held-out validation set. The remaining 2028 sequences were assigned at random into training (n = 1352) and testing (n = 676) sets. Each of these three divisions were then used for the development of feed-forward neural networks using an optimization process of evolutionary computation rather than traditional backpropagation. Using evolutionary optimization, it is possible to adjust both the weights associated with each connection and the features that are used as input to the classifier simultaneously. In this manner, we could pre-specify the number of features to use that were sub-selected from a larger set of possible features, thus allowing evolutionary optimization to determine which features in their combination provide the most information. Such sub-selection offers the opportunity to evaluate the predictive performance of neural networks designed with an equal and small number of features as input, providing a rough means of comparison of predictive utility. For the purpose of this section of the effort, we chose to sub-select 10 features as input to the evolved neural networks, with 5 hidden nodes, and 1 output node, where the output node was in the range [0,3] with thresholds assigned on the training set to infer categorization on the testing and validation sets (Table 2A).

A population of feedforward fully-connected multilayer neural networks with sigmoid activation functions were generated and subjected to initial random weight assignments to all connections. Each neural network was scored for its ability to correctly predict the HIV subtype represented by each input sequence. Those neural networks with minimum squared error (MSE) (e.g., a Brier score) in each generation of the evolutionary process were used as “parent” neural networks for the generation of “offspring” neural networks, subject to another round of random variation in the weights associated with the connections. Each neural network population comprised 50 parents and 50 offspring with a tournament selection of 4. As evolutionary optimization proceeded on the training data, the best neural network performance in the population was assayed every 50 generations on the testing data to monitor possible overfitting.

The number of generations of optimization in training that led to a minimum MSE on testing without overfitting was then used to re-run the experiment again for each same fold of the data using the same random seed. At the end of each of these processes, the best neural network was assayed for performance on the training, testing, and, most importantly, the validation set. Decision thresholds were used on the output node of the training data to associate the output with four bins representing each of the four subtypes with those thresholds used to determine performance on the testing and validation sets.

We then sought to understand subtype-specific differences by examining four comparisons: (1) subtype A1 vs. [C, D and B]; (2) subtype B vs. [A1, C and D]; (3) subtype C vs. [A1, B, and D]; and (4) subtype D vs. [A1, B, and C]. For each of these four comparisons we used a two-tailed Student's *t*-test with equal variance over all of the 1120 features to determine if the samples in each comparison were derived from populations with the same mean. Those features with low *p*-values thus had a statistically significant difference in one subtype relative to the others. For each of the four comparisons, we ranked the features according to their p-value and retained the features with the 25 lowest p-values (Table 3). We noticed that many of these features were associated with one or two Nef domains. For instance in the comparison of Subtype C vs. Subtype A1, B, or D, all of the top 25 features were associated with the Alpha Helix A domain of Nef. For Subtype D vs. Subtype A1, B, or C, 16 of the top 25 features were associated with Beta Sheet A. This suggested that characteristics associated with specific Nef domains could be directly relevant to HIV subtype.

To verify this finding, we then used evolved neural networks to produce a binary classification about each subtype relative to the set of all other subtypes combined (e.g., subtype A1 (0) vs. subtype B, C, or D (1)). Given the above understanding that specific Nef domains may be relevant, for each of the four comparisons we restricted the sub-selection only to the primary domain that was associated 25 lowest *p*-values from the *t*-test process. These processes combined enabled us to determine the degree to which is it possible to predict subtypes using very specific Nef domains combined with small feature sets (functional/structural characteristics) of those domains, rather than use long amino acid sequences to infer subtype but provide no meaningful biological differentiation of the samples, other than subtype alone.

For each domain, using the lowest *p*-value features as input, we allowed the evolutionary process to sub-select five features as input to the neural networks using very simple 5 input, 3 hidden node, 1 output node perceptrons. This process was repeated three times with separate divisions of the training and testing data as noted previously and then the three resulting sets of five features for each subtype were examined for similarity and performance on the validation set (Table 4).

We used a notation that includes N possible input features, F sub-selected input features, H hidden nodes, and 1 output node (e.g., N [F]-H-1) as these varied by experiment to help force the evolution into smaller and smaller sets of input features to assist with downstream biological interpretation. For instance, for subtype A1 vs. all other subtypes here we present neural networks with a 15[5]-3–1, an architecture where the output node was over the range [0,1] with specific thresholds on the output that maximized performance over the training examples per subtype decision. Using the training data, the threshold was adjusted until there was roughly equal accuracy on true positives and true negatives to avoid any bias. The code base for this approach was developed internally at Natural Selection, Inc. and has been presented previously in the literature (Lamers et al., 2018; Liu et al., 2020; Lamers et al., 2016; Liu et al., 2021; Lamers et al., 2017; Fogel et al., 2015; Fogel et al., 2014; Lamers et al., 2008).

2.4. Signature pattern analysis and protein modeling

The initial Nef multiple sequence alignment was separated into subtypes A1, B, C, and D and then further pruned to the domains of interest (CD4-I, AP-2 Binding, Alpha Helix A, and Beta-Sheet A). VESPA (Korber and Myers, 1992) was used to identify amino acids and their alignment positions that may have caused the significant *t*-test *p*-values and the ENN subtype-specific domain assignment. VESPA calculates the frequency of each amino acid at each position in an alignment for the query and background set, selects the positions for which the most common character in the query set differs from that in the background set, and provides a frequency score. Comparisons were made between the following populations of sequences: 1) Background: subtype A1 CD4 Downregulation vs. Query: subtype B, C, or D; 2) Background subtype B AP-2 Binding domain vs. Query subtype A1, C, or D; 3) Background subtype C Alpha Helix A vs. Query subtype A1, B, or D; 4) Background subtype D Beta-Sheet A vs. Query subtype A1, B, or C). When a signature was identified, it was noted in along with its associated domain-characteristic and the value of the amino acid in the characteristic scale. Domains that were important for distinguishing subtypes were mapped onto a previously published Nef protein model (Lamers et al., 2018) using Pymol (The PyMOL Molecular Graphics System, 2021)(ver 2.5.4).

3. Results

3.1. ENNs utilizing domain specific amino acid features for the classification across all four subtypes

In this experiment, our goal was to determine if there was sufficient information represented in the 1120 features for reasonable classification across all four subtypes simultaneously. For each of the three divisions of the data (training, testing, and validation) different sets of 10 features were identified in the best-evolved neural network during training with different thresholds (Table 2A). The ENN performed reasonably well in classifying subtypes based on functional domain amino acid physicochemical properties. Validation performance across subtype A1, B, C and D sequences was 87.2%, 89.5%, 91.7%, and 85.1% respectively (Table 2B). A box and whisker plot for the first division as a representation of performance is provided (Fig. 3).

3.2. ENNs to determine if the 25 features with lowest p-value differentiated subtypes

In this experiment, we compared each subtype to all of the others (e. g., subtype A1 vs {subtypes B, C, D}; subtype B vs. {subtypes A1, D, C}; subtype C vs. {subtypes A1, B, D}; and subtype D vs {subtypes A1, B, C}. *t*-tests were used to identify the top 25 features with the lowest p-value that differentiated each subtype (Table 3). All significance values for the features shown were highly significant ($p \langle 10^{-20}$). Interestingly, when each of the four subtypes was compared to the other three, different Nef domains for each comparison were considered most informative. For instance, in the comparison of subtype A1 vs. {B, C, D}, the majority of the features with lowest *p*-values (13 out of 25), were found in the positions associated with CD4 Downregulation. For subtype B, there was more variability, with a mixture of AP-2 Binding ($n = 10$) and Alpha Helix CD ($n = 11$) domains being most discriminatory; however, AP-2 Binding was a unique feature not found in the lists for subtypes A, C, or D domain features. For subtype C, all 25 of the

top features were associated with the Alpha Helix A domain, which was not represented in the subtype A1, B, D discriminating features list. For subtype D, 15 of the top 25 features were associated with the Beta-Sheet A domain and 7 top features were associated with the complementary Beta-Sheet B domain. It is also interesting to note that many Nef domains had no representation in the top features from this test (Association with Signaling Molecules, MHC-1 Downmodulation I (MHC-1 DM1), MHC-1 Downmodulation II (MHC-1 DM2), MHC1-signaling, Myristoylation, and Sh-3 Binding), indicating that only specific Nef domains are associated with subtype differences.

3.3. ENNs to determine if domain-specific features alone perform better than the 25 lowest p-values

For subtype A1 vs. {B, C, D}, we used the 15 top features associated with CD4 Downregulation identified by *t*-tests with $p < 10^{-20}$ and provided these to a neural network in a 15[5]-3–1 architecture. Each of the three divisions of the data resulted in best neural networks with different sets of 5 input features. One feature (pI at 25 °C) was present in all three best neural networks, while four features (HPLC/TFA, Welling, Browne, and Retention at pH 2.1) were found in two out of the three best neural networks. The best classification by the ENN used two hydrophobicity characteristics (Welling and Browne) and three characteristics from the “other” category (pI at 25 °C, HPLC/TFA, and Retention at pH 2.1) (Fig. 4A). Using just these five features in combination, the ENN successfully differentiated subtype A1 from the other subtypes in training, testing, and validation with perfect accuracy (Fig. 4B and C). The AP-2 domain consists of only four amino acids, with two of the positions highly conserved among all Nef sequences. An amino acid signature was identified at position 1 between subtype A1 and D, where 89% of subtype A1 Nef sequences preserved an alanine (A) in contrast to 25% of subtype D Nef; alternatively, 58% of Subtype D sequences preserved an aspartic (D) or glutamic acid (E) at this position, both of which are strongly negatively charged amino acids. The same signature was identified between subtype A1 and subtype B and C, where subtype B and C preserved an asparagine (N) at >80% of the sequences and 65% of subtype A1 sequences preserved a threonine (T) (Fig. 4D).

For subtype B vs. {A1, C, D}, because AP-2 binding was considered a unique feature associated with the subtype, we used the 17 top features associated with AP-2 binding with a *t*-test $p < 10^{-20}$ and provided these to a neural network in a 17[5]-3–1 architecture. Each of the three divisions of the data resulted in best neural networks with different sets of 5 input features. One feature (Charge Polarity) was present in all three best neural networks, while three features (Refractivity, Retention at pH 2.1, and Grantham) were found in two out of the three best neural networks. The best ENN identified the following five features as most discriminatory, which included two polarity characteristics (Grantham and Charge Polarity), two hydrophobicity characteristics (Cowan and Whittaker and Browne), and one characteristic based on the amino acid retention coefficient in HPLC at pH 2.1 (Fig. 5A) and on average the best neural networks over the three splits of the data retained reasonable classification accuracy; however, this subtype was the most difficult of the four to differentiate via the ENNs, with validation experiments correctly classifying subtype B sequences at 91% (Fig. 5B and C). The AP-2 Binding domain is located within the Nef loop,

a highly unstructured and multifunctional binding domain. A variety of signature positions were identified between the subtypes in the AP-2 Binding domain. The most interesting, and ubiquitous among subtypes was position 17, where a leucine (L) appears in >50% of subtype B sequences; however, subtypes A1 and D maintain a glutamine (Q) at >90%, and subtype C a glutamine appears in 49% of sequences. Cysteines (C) appear in a majority of the AP-2 domain in subtype A1 at position 16 and in subtypes C and D (position 10). Cysteines appear less frequently in subtype B AP-2 domains, and when they do, it is in position 10 (Fig. 5D). This is interesting considering the unique ability of cysteines to participate in protein structure and folding through the formation of stable intramolecular and intermolecular disulfide bonds.

For subtype C vs. {A1, B, D}, we used the 45 features associated with Alpha Helix A domain with a *t*-test $p < 10^{-20}$ and provided these to a neural network in a 45[5]-3–1 architecture. Each of the three divisions of the data resulted in best neural networks with different sets of 5 input features. A maximum of two features (Exchange, Bull and Breese) were common in two out of the three best neural networks. By using three hydrophobicity characteristics (Rose, Bull and Breese, and Cowan), the Exchange characteristic, which is a categorical scale based on the frequency of amino acid substitution rates, and one categorical characteristic associated with amino acid surface exposure (Surface Exposure) (Fig. 6A), the ENN successfully differentiated subtype C from the other subtypes with validation tests at 98.6% (Fig. 6B and C). At position 8, 98% of subtype C sequences preserved a phenylalanine (F), whereas >80% of subtypes A1, B, and D preserved a histidine (H) (Fig. 6D). This is a rare substitution (as observed in the surface exposure and exchange scales). Phenylalanine (F) is strongly hydrophobic, whereas histidine (H) is hydrophilic. Slight alterations in Nef subtype C Alpha Helix domains have been associated with reducing the stability of the protein (Johnson et al., 2016).

For subtype D vs. {A1, B, C}, we used 22 features from the Beta-Sheet A domain identified with a *t*-test $p < 10^{-20}$ and provided these to a neural network in a 22[5]-3–1 architecture. Each of the three divisions of the data resulted in best neural networks with different sets of 5 input features. One feature (Wilson) was found in common to best neural networks resulting from all three divisions, while relative mutability and charge were found in two of the three best neural networks. By using the following combination of features that included two hydrophobicity scales (Wilson and Janin), one polarity scale (Charge Conversion), one compositional characteristic (Relative Mutability), and one structural scale (Alpha and Chou Fasman) (Fig. 7A) the ENN successfully differentiated subtype C from the other subtypes with 89.4% accuracy (Fig. 7B and C). Only one signature amino acid was observed here, a glutamic acid (E) was found in 82% of subtype D sequences, whereas a lysine (K) was found >99% of subtype A1, B, and C sequence populations (Fig. 7D). This is considered a radical amino acid substitution as the amino acids are oppositely charged, differing in the properties of their side chains: glutamic acid (E) is acidic, and lysine (K) is basic. A key feature of the Nef structure is how it takes advantage of conformational changes, primarily through beta sheet folding in the protein core, to facilitate the recruitment of specific target proteins into clathrin-coated vesicles (Kwon et al., 2020). While there are a variety of forces that contribute to protein folding, hydrophobicity is thought to be one of the primary forces

(Moelbert et al., 2004). Beta-sheet changes are also associated with AP-2 Binding pocket formation.

4. Discussion

In this study we assembled a large number of HIV Nef genetic sequences comprising of four HIV subtypes and extracted structural and functional characteristics of amino acids in domains associated with Nef's ability to modulate cellular interactions. Using ENNs, we identified that Nef may modify cellular functions in a subtype-specific manner. As Nef proteins have a major role in establishing and maintaining infection and persist in patients on ART, these subtle genetic differences among subtypes may influence disease progression and reservoir maintenance.

The domains where we identified subtype-specific alterations in the Nef structure are shown (Fig. 8). Subtype A1 sequences were associated with modifications in positions associated with CD4 downregulation, which is comprised of four amino acids within the flexible AP-2 and C-terminal flexible loop. After forming a complex with AP-2, Nef uses these amino acids to bind the CD4 cytoplasmic tail on the cellular membrane for internalization (Ren et al., 2014). Interestingly, subtype B sequences were associated with alterations in additional positions in the AP-2 Binding domain, which could impact CD4 internalization efficiency via a different mechanism. While CD4 binding is a direct means of Nef to remove CD4, highly efficient AP-2 binding can result in internal folding of the cell wall, generating endosomes containing both CD4 and SERINC5, which are then shuttled to lysosomes for degradation (Buffalo et al., 2019). Nef's ability to successfully downregulate CD4 and SERINC5 is associated with AP-2 and AP-1 binding at the cell wall and trans-Golgi-network (TGN) respectively, however with differing impacts. Without Nef-associated down-regulation of CD4, viral release is impacted, and super-infection is more likely (Lindwasser et al., 2007). Furthermore, if SERINC5 persists on cell surfaces, it becomes incorporated into viral particles, making them less infective (Staudt and Smithgall, 2020; Chai et al., 2021).

One of Nef's key features is that it can deviate from its original structure to facilitate binding processes. The folding of Nef is accomplished through complex mechanisms in Nef's core, including the alpha helix and beta sheet domains (Buffalo et al., 2019). Subtype C signatures in the alpha helix domain are associated with reduced MHC-1 trafficking from the TGN (Mangasarian et al., 1999). Subtype D signatures mapped to the Beta-Sheet A, another structurally conserved domain within Nef sequences that is important for successful protein folding and downstream binding of AP-1 and AP-2 and down-regulation of MHC-1. Nef's fold-dependent interaction with AP-1 molecules is also required for the recruitment of specific target proteins into clathrin-coated vesicles (Kwon et al., 2020).

5. Conclusions

Nef is a complex disease regulator with multiple functions. Nef is the first HIV protein that is produced in abundance (Pereira and daSilva, 2016) and also the first HIV protein to display host-specific adaptations (Lamers et al., 2015). The finding that Nef protein

persists in tissues absent a measurable viral load, adds additional support that Nef may be a driver of disease-associated processes under ART therapy (Nolan et al., 2022; Rose et al., 2016; Duette et al., 2022; Ferdin et al., 2018). In fact, one study found that Nef derived from exosomes from patients with HIV-associated dementia could increase the secretion of beta-amyloid in bystander cells, thus providing even more direct evidence of the Nef-disease correlation (Khan et al., 2016).

Only a few groups have assessed how naturally occurring Nef subtype diversity may influence its ability to modulate signaling. Using a transient expression assay, one study revealed that SERINC internalization function varied significantly among subtypes in 339 HIV Nef sequences from subtypes A1, B, C, and D (Jin et al., 2020). Interestingly, in this study, subtype B demonstrated the least significant difference in SERINC internalization to any of the other three subtypes. We also found that subtype B Nef had the most variability in terms of the domains and features relative to the other subtypes. Another inter-subtype study used CD4 transfection assays of 360 Nef sequences followed by flow cytometry to assess CD4 and HLA class 1 surface levels (Mann et al., 2013). CD4 downregulation was marginally higher in subtype B than subtypes A1, C, and D; however HLA downregulation significantly varied among all subtypes, with subtype B displaying the greatest ability to downregulate, followed by subtypes D/A1, and then C. Finally, one other inter-subtype study demonstrated subtype B and D Nef clones were more successful in inhibiting T-cell receptor-mediated NFAT signaling when compared to subtypes A1 and C Nef clones (Naidoo et al., 2019). This difference alone could impact humoral immune responses, immunological tolerance, and immune metabolism regulation (Vaeth and Feske, 2018).

These studies complement ours by providing additional evidence that Nef subtypes vary in their ability to modulate the immune system, which may contribute to the variation in the pathogenicity of the subtypes. Further functional studies like the ones just mentioned using cultured cells, and studies utilizing animal models, will have to be performed to explore and validate the subtype distinctions in nef found by our ENN analysis. Continued functional characterization of different HIV-1 subtypes may improve our understanding of viral pathogenesis and spread and provide insights to improve anti-Nef therapies under consideration.

Acknowledgements

Funding: This work was supported by the National Institutes of Health grant #R01CA239263-03. SLL, GBF, ESL, DJN, RR and MSM assisted in data review and manuscript preparation and. SLL, GBF, ESL designed and implemented the study. MSM provided funding for the study.

Data availability

The study used previously published sequences data derived from the Los Alamos HIV sequence database. All alignments used for the study are available by making a request to the corresponding author.

References

Amornkul PN, Karita E, Kamali A, Rida WN, Sanders EJ, Lakhi S, et al. , 2013. Disease progression by infecting HIV-1 subtype in a seroconverter cohort in sub-Saharan Africa. AIDS. 27 (17), 2775–2786. [PubMed: 24113395]

Apetrei C, Hanh B, Rambaut A, Wolinski S, Brister J, Faser C, HIV Sequence Compendium, 2021. Los Alamos National Laboratory. Theoretical Biology and Biophysics Group, NM.

Arold ST, Baur AS, 2001. Dynamic Nef and Nef dynamics: how structure could explain the complex activities of this small HIV protein. Trends Biochem. Sci 26 (6), 356–363. [PubMed: 11406408]

Blanquart F, Grabowski MK, Herbeck J, Nalugoda F, Serwadda D, Eller MA, et al. , 2016. A transmission-virulence evolutionary trade-off explains attenuation of HIV-1 in Uganda. Elife. 5.

Bouchet J, Basmaciogullari SE, Chrobak P, Stolp B, Bouchard N, Fackler OT, et al. , 2011. Inhibition of the Nef regulatory protein of HIV-1 by a single-domain antibody. Blood. 117 (13), 3559–3568. [PubMed: 21292773]

Buffalo CZ, Iwamoto Y, Hurley JH, Ren X, 2019. How HIV Nef proteins hijack membrane traffic to promote infection. J. Virol 93 (24).

Chai Q, Li S, Collins MK, Li R, Ahmad I, Johnson SF, et al. , 2021. HIV-1 Nef interacts with the cyclin K/CDK13 complex to antagonize SERINC5 for optimal viral infectivity. Cell Rep. 36 (6), 109514. [PubMed: 34380030]

Chun TW, Fauci AS, 2012. HIV reservoirs: pathogenesis and obstacles to viral eradication and cure. AIDS. 26 (10), 1261–1268. [PubMed: 22472858]

Cohn LB, Chomont N, Deeks SG, 2020. The biology of the HIV-1 latent reservoir and implications for cure strategies. Cell Host Microbe 27 (4), 519–530. [PubMed: 32272077]

Collinson-Streng AN, Redd AD, Sewankambo NK, Serwadda D, Rezapour M, Lamers SL, et al. , 2009. Geographic HIV type 1 subtype distribution in Rakai district, Uganda. AIDS Res. Hum. Retrovir 25 (10), 1045–1048. [PubMed: 19803713]

Conroy SA, Laeyendecker O, Redd AD, Collinson-Streng A, Kong X, Makumbi F, et al. , 2010. Changes in the distribution of HIV type 1 subtypes D and a in Rakai District, Uganda between 1994 and 2002. AIDS Res. Hum. Retrovir 26 (10), 1087–1091. [PubMed: 20925575]

Deacon NJ, Tsykin A, Solomon A, Smith K, Ludford-Menting M, Hooker DJ, et al. , 1995. Genomic structure of an attenuated quasi species of HIV-1 from a blood transfusion donor and recipients. Science. 270 (5238), 988–991. [PubMed: 7481804]

Duette G, Hiener B, Morgan H, Mazur FG, Mathivanan V, Horsburgh BA, et al. , 2022. The HIV-1 proviral landscape reveals that Nef contributes to HIV-1 persistence in effector memory CD4+ T cells. J. Clin. Invest 132 (7).

Fackler OT, Baur AS, 2002. Live and let die: Nef functions beyond HIV replication. Immunity. 16 (4), 493–497. [PubMed: 11970873]

Felli C, Vincentini O, Silano M, Masotti A, 2017. HIV-1 Nef signaling in intestinal mucosa epithelium suggests the existence of an active inter-kingdom crosstalk mediated by exosomes. Front. Microbiol 8, 1022. [PubMed: 28642743]

Ferdin J, Goricar K, Dolzan V, Plemenitas A, Martin JN, Peterlin BM, et al. , 2018. Viral protein Nef is detected in plasma of half of HIV-infected adults with undetectable plasma HIV RNA. PLoS One 13 (1), e0191613. [PubMed: 29364927]

Fischer M, Wong JK, Russenberger D, Joos B, Opravil M, Hirschel B, et al. , 2002. Residual cell-associated unspliced HIV-1 RNA in peripheral blood of patients on potent antiretroviral therapy represents intracellular transcripts. Antivir. Ther 7 (2), 91–103. [PubMed: 12212929]

Fogel G, Liu ES, Salemi M, L SL, McGrath MS, 2014. Evolved neural networks for HIV-1 co-receptor identification. In: IEEE Congress on Evolutionary Computation (CEC), pp. 2778–2784.

Fogel GB, Lamers SL, Liu ES, Salemi M, McGrath MS, 2015. Identification of dual-tropic HIV-1 using evolved neural networks. Biosystems. 137, 12–19. [PubMed: 26419858]

Gandhi RT, Bedimo R, Hoy JF, Landovitz RJ, Smith DM, Eaton EF, et al. , 2023. Antiretroviral drugs for treatment and prevention of HIV infection in adults: 2022 recommendations of the international antiviral society-USA panel. JAMA. 329 (1), 63–84. [PubMed: 36454551]

Hendricks CM, Cordeiro T, Gomes AP, Stevenson M, 2021. The interplay of HIV-1 and macrophages in viral persistence. Front. Microbiol 12, 646447. [PubMed: 33897659]

Jin SW, Mwimanzi FM, Mann JK, Bwana MB, Lee GQ, Brumme CJ, et al. , 2020. Variation in HIV-1 Nef function within and among viral subtypes reveals genetically separable antagonism of SERINC3 and SERINC5. PLoS Pathog. 16 (9), e1008813. [PubMed: 32925973]

Johnson AL, Dirk BS, Coutu M, Haeryfar SM, Arts EJ, Finzi A, et al. , 2016. A Highly Conserved Residue in HIV-1 Nef Alpha Helix 2 Modulates Protein Expression. mSphere 1 (6).

Kamalzare S, Noormohammadi Z, Rahimi P, Atyabi F, Irani S, Tekie FSM, et al. , 2019. Carboxymethyl dextran-trimethyl chitosan coated superparamagnetic iron oxide nanoparticles: an effective siRNA delivery system for HIV-1 Nef. J. Cell. Physiol 234 (11), 20554–20565. [PubMed: 31144311]

Khan MB, Lang MJ, Huang MB, Raymond A, Bond VC, Shiramizu B, et al. , 2016. Nef exosomes isolated from the plasma of individuals with HIV-associated dementia (HAD) can induce Abeta(1-42) secretion in SH-SY5Y neural cells. J. Neuro-Oncol 22 (2), 179–190.

Kirchhoff F, Greenough TC, Brettler DB, Sullivan JL, Desrosiers RC, 1995. Brief report: absence of intact nef sequences in a long-term survivor with nonprogressive HIV-1 infection. N. Engl. J. Med 332 (4), 228–232. [PubMed: 7808489]

Korber B, Myers G, 1992. Signature pattern analysis: a method for assessing viral sequence relatedness. AIDS Res. Hum. Retrovir 8 (9), 1549–1560. [PubMed: 1457200]

Kwon Y, Kaake RM, Echeverria I, Suarez M, Karimian Shamsabadi M, Stoneham C, et al. , 2020. Structural basis of CD4 downregulation by HIV-1 Nef. Nat. Struct. Mol. Biol 27 (9), 822–828. [PubMed: 32719457]

Lamers SL, Salemi M, McGrath MS, Fogel GB, 2008. Prediction of R5, X4, and R5X4 HIV-1 Coreceptor usage with evolved neural networks. IEEE/ACM Trans. Comput. Biol. Bioinform 5, 291–300.

Lamers SL, Fogel GB, Singer EJ, Salemi M, Nolan DJ, Huysentruyt LC, et al. , 2012. HIV-1 Nef in macrophage-mediated disease pathogenesis. Int. Rev. Immunol 31 (6), 432–450. [PubMed: 23215766]

Lamers SL, Nolan DJ, Rife BD, Fogel GB, McGrath MS, Burdo TH, et al. , 2015. Tracking the emergence of host-specific simian immunodeficiency virus env and nef populations reveals nef early adaptation and convergent evolution in brain of naturally progressing Rhesus macaques. J. Virol 89 (16), 8484–8496. [PubMed: 26041280]

Lamers SL, Fogel GB, Liu ES, Salemi M, McGrath MS, 2016. On the physicochemical and structural modifications associated with HIV-1 subtype B tropism transition. AIDS Res. Hum. Retrovir 32 (8), 829–840. [PubMed: 27071630]

Lamers SL, Fogel GB, Liu ES, Nolan DJ, Salemi M, Barbier AE, et al. , 2017. Predicted coreceptor usage at end-stage HIV disease in tissues derived from subjects on antiretroviral therapy with an undetectable plasma viral load. Infect. Genet. Evol 51, 194–197. [PubMed: 28392467]

Lamers SL, Fogel GB, Liu ES, Barbier AE, Rodriguez CW, Singer EJ, et al. , 2018. Brain-specific HIV Nef identified in multiple patients with neurological disease. J. Neuro-Oncol 24 (1), 1–15.

Lamers SL, Rose R, Cross S, Rodriguez CW, Redd AD, Quinn TC, et al. , 2020. HIV-1 subtype distribution and diversity over 18 years in Rakai, Uganda. AIDS Res. Hum. Retrovir 36 (6), 522–526. [PubMed: 32281387]

Leite TC, Campos DP, Coelho AB, Teixeira SL, Veloso V, Morgado MG, et al. , 2017. Impact of HIV-1 subtypes on AIDS progression in a Brazilian cohort. AIDS Res. Hum. Retrovir 33 (1), 41–48. [PubMed: 27418261]

Lindwasser OW, Chaudhuri R, Bonifacino JS, 2007. Mechanisms of CD4 downregulation by the Nef and Vpu proteins of primate immunodeficiency viruses. Curr. Mol. Med 7 (2), 171–184. [PubMed: 17346169]

Liu ES, Fogel GB, D.J. N, L SL, M MS, 2020. Using Neural Networks to Identify Features Associated with HIV Nef Protein and Cancer. In: IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology.

Liu ES, Fogel GB, Nolan D, Lamers SL, McGrath MS, 2021. Using Evolved Neural Networks to Elucidate Nef Features Associated with HIV-1 Subtype Differentiation. In: IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB), pp. 1–8.

Lv Y, Meng Y, Zhong J, Wan J, Zou W, 2020. Research Progress of HIV-1 Nef inhibitors. AIDS Rev. 22 (4), 221–226. [PubMed: 33105470]

Mangasarian A, Piguet V, Wang JK, Chen YL, Trono D, 1999. Nef-induced CD4 and major histocompatibility complex class I (MHC-I) down-regulation are governed by distinct determinants: N-terminal alpha helix and proline repeat of Nef selectively regulate MHC-I trafficking. J. Virol 73 (3), 1964–1973. [PubMed: 9971776]

Mann JK, Byakwaga H, Kuang XT, Le AQ, Brumme CJ, Mwimanzi P, et al. , 2013. Ability of HIV-1 Nef to downregulate CD4 and HLA class I differs among viral subtypes. Retrovirology. 10, 100. [PubMed: 24041011]

McNamara RP, Costantini LM, Myers TA, Schouest B, Maness NJ, Griffith JD, et al. , 2018. Nef secretion into extracellular vesicles or exosomes is conserved across human and simian immunodeficiency viruses. mBio. 9 (1).

Moelbert S, Emberly E, Tang C, 2004. Correlation between sequence hydrophobicity and surface-exposure pattern of database proteins. Protein Sci. 13 (3), 752–762. [PubMed: 14767075]

Mukhamedova N, Hoang A, Dragoljevic D, Dubrovsky L, Pushkarsky T, Low H, et al. , 2019. Exosomes containing HIV protein Nef reorganize lipid rafts potentiating inflammatory response in bystander cells. PLoS Pathog. 15 (7), e1007907. [PubMed: 31344124]

Naidoo L, Mzobe Z, Jin SW, Rajkoomar E, Reddy T, Brockman MA, et al. , 2019. Nef-mediated inhibition of NFAT following TCR stimulation differs between HIV-1 subtypes. Virology. 531, 192–202. [PubMed: 30927712]

Nolan DJ, Rose R, Zhang R, Leong A, Fogel GB, Scholte LLS, et al. , 2022. The persistence of HIV diversity, transcription, and Nef protein in Kaposi's sarcoma tumors during antiretroviral therapy. Viruses. 14 (12).

Olivetta E, Arenaccio C, Manfredi F, Anticoli S, Federico M, 2016. The contribution of extracellular Nef to HIV-induced pathogenesis. Curr. Drug Targets 17 (1), 46–53. [PubMed: 26424397]

Painter MM, Zimmerman GE, Merlino MS, Robertson AW, Terry VH, Ren X, et al. , 2020. Concanamycin a counteracts HIV-1 Nef to enhance immune clearance of infected primary cells by cytotoxic T lymphocytes. Proc. Natl. Acad. Sci. U. S. A 117 (38), 23835–23846. [PubMed: 32900948]

Pereira EA, daSilva LL, 2016. HIV-1 Nef: taking control of protein trafficking. Traffic. 17 (9), 976–996. [PubMed: 27161574]

Puzar Dominkus P, Ferdin J, Plemenitas A, Peterlin BM, Lenassi M, 2017. Nef is secreted in exosomes from Nef.GFP-expressing and HIV-1-infected human astrocytes. J. Neuro-Oncol 23 (5), 713–724.

Ren X, Park SY, Bonifacino JS, Hurley JH, 2014. How HIV-1 Nef hijacks the AP-2 clathrin adaptor to downregulate CD4. Elife. 3, e01754. [PubMed: 24473078]

Rose R, Lamers SL, Nolan DJ, Maidji E, Faria NR, Pybus OG, et al. , 2016. HIV maintains an evolving and dispersed population in multiple tissues during suppressive combined antiretroviral therapy in individuals with Cancer. J. Virol 90 (20), 8984–8993. [PubMed: 27466425]

Sereti I, Krebs SJ, Phanuphak N, Fletcher JL, Slike B, Pinyakorn S, et al. , 2017. Persistent, albeit reduced, chronic inflammation in persons starting antiretroviral therapy in acute HIV infection. Clin. Infect. Dis 64 (2), 124–131. [PubMed: 27737952]

Shacklett BL, Ferre AL, Kiniry BE, 2019. Tissue issues: mucosal T-cell responses in HIV-1 infection. Curr. Opin. HIV AIDS 14 (2), 100–107. [PubMed: 30601239]

Shen QT, Ren X, Zhang R, Lee IH, Hurley JH, 2015. HIV-1 Nef hijacks clathrin coats by stabilizing AP-1:Arf1 polygons. Science. 350 (6259), aac5137. [PubMed: 26494761]

Shiels MS, Engels EA, 2017. Evolving epidemiology of HIV-associated malignancies. Curr. Opin. HIV AIDS 12 (1), 6–11. [PubMed: 27749369]

Staudt RP, Smithgall TE, 2020. Nef homodimers down-regulate SERINC5 by AP-2-mediated endocytosis to promote HIV-1 infectivity. J. Biol. Chem 295 (46), 15540–15552. [PubMed: 32873704]

Struck D, Lawyer G, Ternes AM, Schmit JC, Bercoff DP, 2014. COMET: adaptive context-based modeling for ultrafast HIV-1 subtype identification. Nucleic Acids Res. 42 (18), e144. [PubMed: 25120265]

Sung JM, Margolis DM, 2018. HIV persistence on antiretroviral therapy and barriers to a cure. Adv. Exp. Med. Biol 1075, 165–185. [PubMed: 30030793]

The PyMOL Molecular Graphics System, 2021. Version 2.5 Schrödinger. LLC.

Vaeth M, Feske S, 2018. NFAT control of immune function: new Frontiers for an abiding trooper. F1000Res. 7, 260. [PubMed: 29568499]

Zhu X, Guo Y, Yao S, Yan Q, Xue M, Hao T, et al. , 2014. Synergy between Kaposi's sarcoma-associated herpesvirus (KSHV) vIL-6 and HIV-1 Nef protein in promotion of angiogenesis and oncogenesis: role of the AKT signaling pathway. Oncogene. 33 (15), 1986–1996. [PubMed: 23604117]

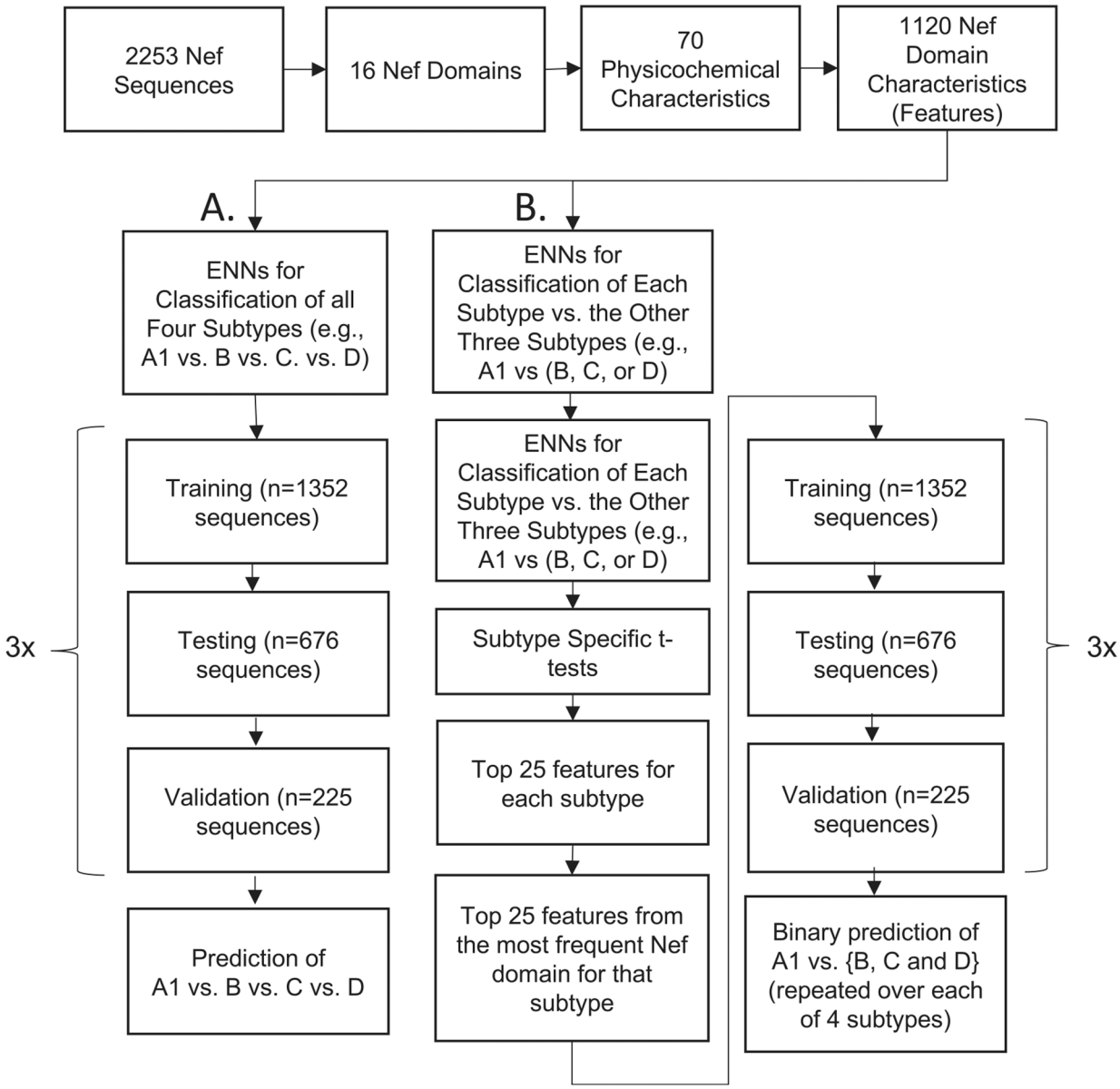


Fig. 1.

Data flow for two data classifications systems (A and B) using evolved neural networks. Each classification system had essentially four stages, 1) sequence identification, domain assignment, calculation of features for all domains, 2) design of evolved neural networks (ENNs) through training, testing and validation sets, in triplicate with random sampling, 3) classification of individual subtype (A) or classification of one subtype relative to all other subtypes (B). In classification system A, we determined that sufficient information was present in domain-specific regions to adequately predict each subtype using an ENN approach. The 10 features that identify each subtype and their performance can be found in Tables 2A&B. In classification system B, we determined via t-tests that each subtype had signal in a specific domain that differentiated it from the other subtypes (Table 3).

We validated this assumption using ENNs and the top 25 features associated with each subtype-specific domain with validation scores >86% (Table 4).

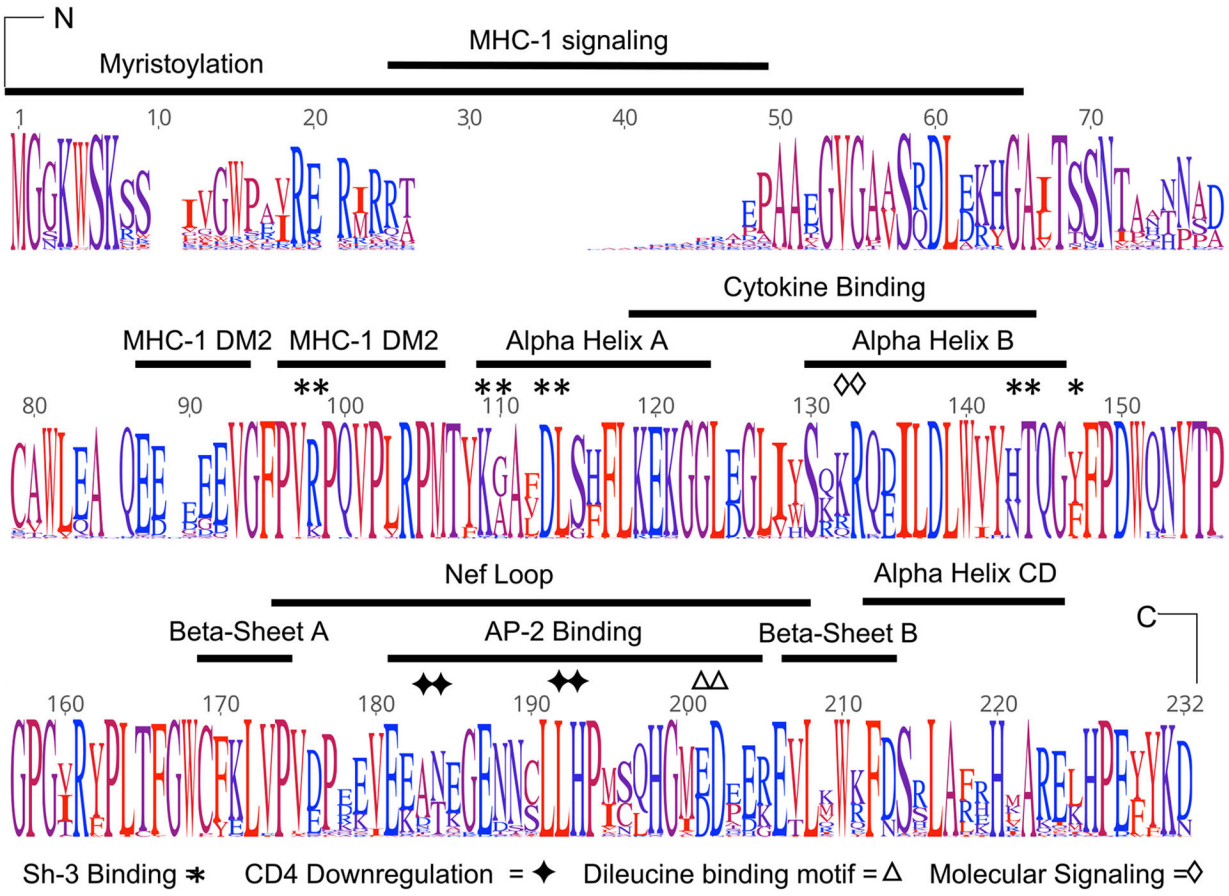


Fig. 2.

SeqLogo consensus sequence showing Nef variation among subtypes and functional domains studied. The image represents the consensus alignment of 1628 sequences from HIV subtypes A, B, C, and D. The height of each 1-letter amino acid code correlates with its presence in the aligned set of sequences. Amino acids are colored by hydrophobicity value, where red is the most hydrophobic and blue is the most hydrophilic.

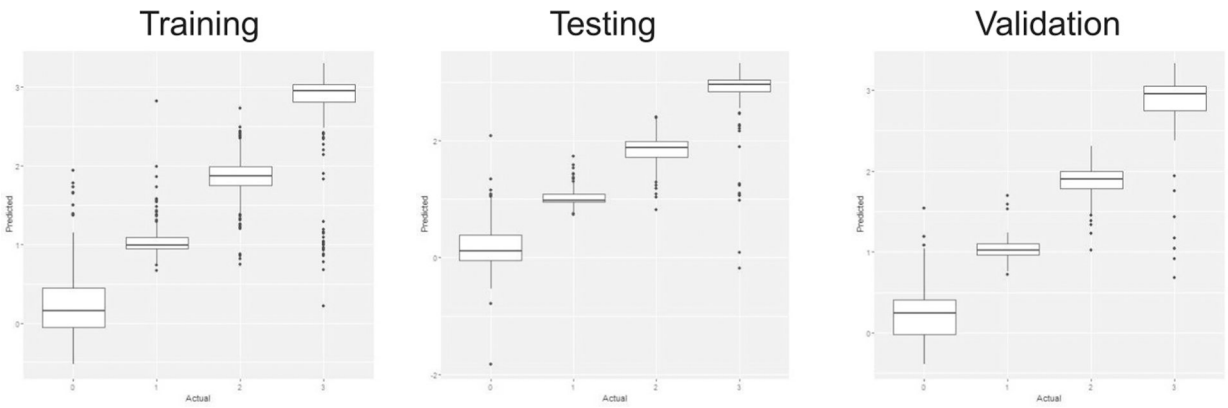
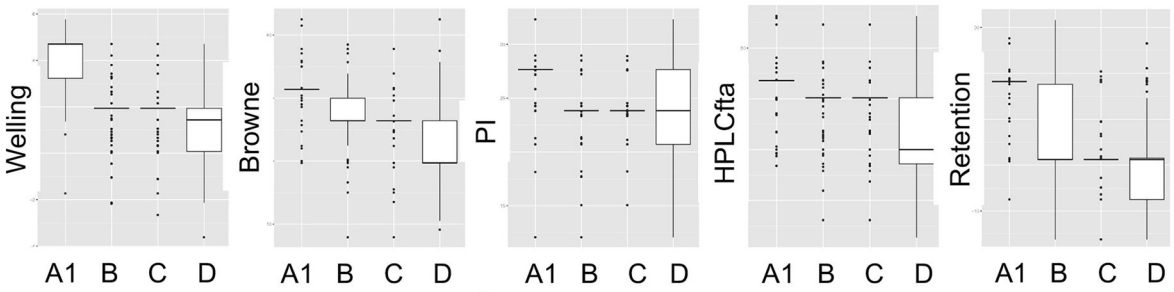


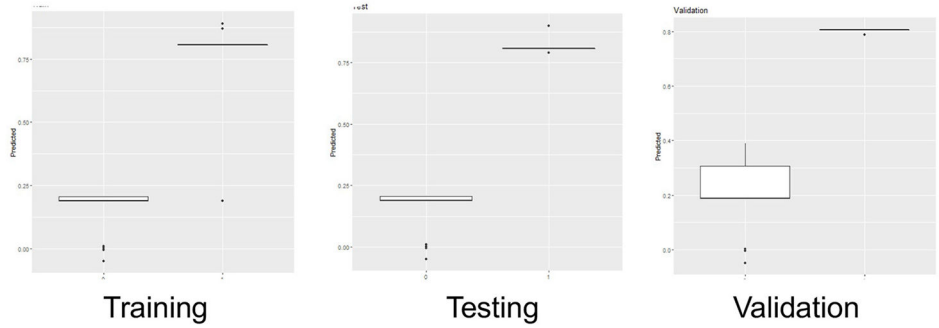
Fig. 3.

Performance of evolved neural networks on dataset 1 for subtype classification over all subtypes. Performance is shown graphically for training (A), testing (B), and validation (C) with the four subtypes A1 through D on the x-axis and the neural network output prediction on the y-axis from 0 to 3 representing the four classes.

A.



B.



C.

		Actual	
		Other	**Subtype A**
Predicted	**Other**	188/188=100.0%	0/39=0.0%
	Subtype A	0/188=0.0%	39/39=100.0%

D.

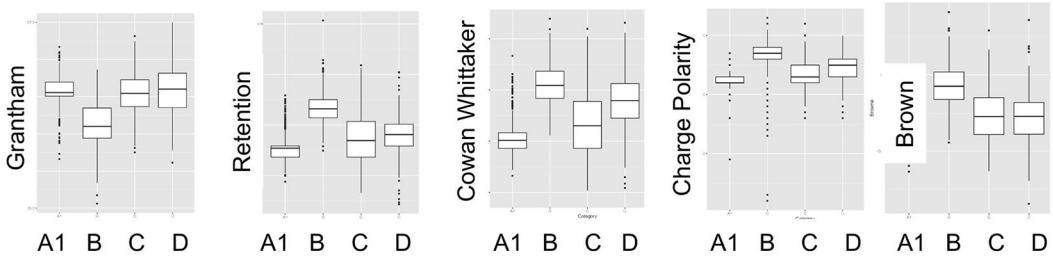
Domain	Position	Amino Acids				Welling				Browne			
		A	B	C	D	A	B	C	D	A	B	C	D
CD4-D	1	A			D	1.15			0.65	3.9			-2.8
	2	T	N	N		-0.45	-0.77	-0.77		1.1	-2.8	-2.8	
Domain	Position	PI				HPLCfta				Retention			
		A	B	C	D	A	B	C	D	A	B	C	D
CD4-D	1	6			5.41	7.3			-2.9	0.5			-8.2
	2	5.6	2.77	2.77		0.8	-5.7	-5.7		2.7	0.8	0.8	

Fig. 4.

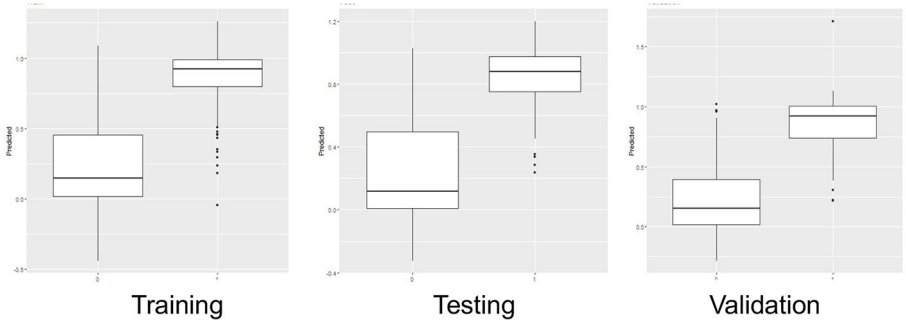
Subtype A and CD4 downregulation. A) Box plots showing the ranges of the 5 features used by ENNs for classification. B) Box plot showing training, testing, and validation ENNs using the top 25 features for CD4 downregulation. C) Cross tables of predicted vs. actual ability of ENN to identify subtypes using features applied to functional domains. D) Comparative signature pattern analysis (VESPA) among subtypes over the subtype-distinguishing domain. Only the domain positions where a > 50% difference appeared between the background (subtype A1) and query (subtype B, C, D) occurred are shown. Colored headings indicate the top five feature groups used by the ENN within the CD4 downregulation domain with yellow = hydrophobicity scale; green = HPLC or “other” scale. The absolute score for each amino acid for each feature scale is shown. Red amino acids indicate the background data set (subtype A1) and black values indicate the query data for subtypes B, C, and D. Gray boxes indicate the query and the background sequences agree at

the specified position. Values in bold indicate that the amino acid occurred >80% of the time in the specified position.

A.



B.



C.

		Actual	
		Other	Type B
Predicted	Other	143/158=90.5%	6/67=9.0%
	Type B	15/158=9.5%	61/67=91.0%

D.

Domain	Position	amino acids				Grantham				Retention			
		A	B	C	D	A	B	C	D	A	B	C	D
AP2 Binding	2	**K**	E			11.3	10.5			K	-4.8		
	3		**A**		D		**8.1**		13		**0.5**		-8.2
	4	T	**N**		**T**	8.6	**11.6**		**8.6**	2.7	**0.8**		**2.7**
	5		E	K			10.5	11.3			-4.8	K	
	10		S	**C**	**C**		9.2	**5.5**	**5.5**		1.2	**-6.8**	**-6.8**
	15	I	M			5.2	5.7			13.9	4.8		
	16	**C**	**S**		C	**5.5**	**9.2**		5.5	**-6.8**	**1.2**		-6.8
	17	**Q**	L	Q	**Q**	**12.3**	4.9	12.3	**12.3**	**-16.9**	8.8	-16.9	**-16.9**
	21		D	**E**	**E**		13	**10.5**	**10.5**		-8.2	**-4.8**	**-4.8**
	23	**E**	**P**	E		**10.5**	**8**	10.5		**-4.8**	**6.1**	-4.8	

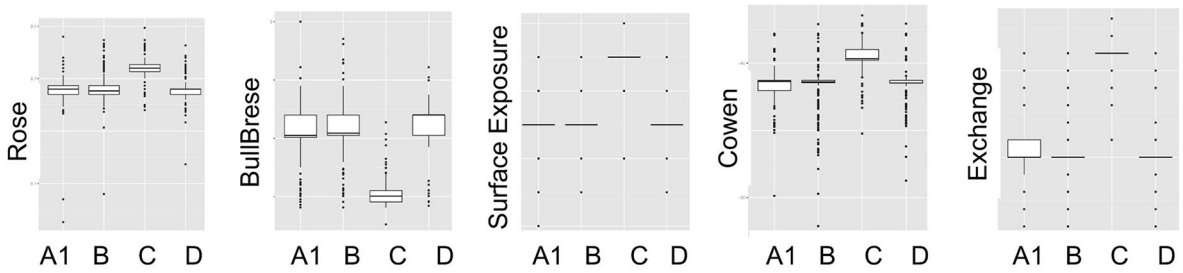
Domain	Position	Cowen Whitaker				Charge Polarity				Brown			
		A	B	C	D	A	B	C	D	A	B	C	D
AP2 Binding	2	-1.54	-0.93			1	2			-2.5	1.8		
	3		**0.35**		-2.15		4		2		**3.9**		-2.8
	4	-0.27	**-0.99**		**-0.27**	3	3		3	1.1	**-2.8**		**1.1**
	5		-0.93	-1.54			2	1			1.8	-2.5	
	10		-0.63	**0.76**	**0.76**		S	**3**	**3**		-3.5	**-14.3**	**-14.3**
	15	I	1.1			4	4			11	4.1		
	16	**0.76**	**-0.63**		0.76	**3**	**3**		3	**-14.3**	**-3.5**		-14.3
	17	**-1.95**	**1.8**	-1.95	**-1.95**	**3**	**4**	3	**3**	**-7.5**	**15**	-7.5	**-7.5**
	21		-2.015	**-0.93**	**-0.93**		2	**2**	**2**		-2.8	**1.8**	**1.8**
	23	**-0.093**	**1.69**	-0.93		**2**	**4**	2		**1.8**	**5.6**	1.8	

Fig. 5.

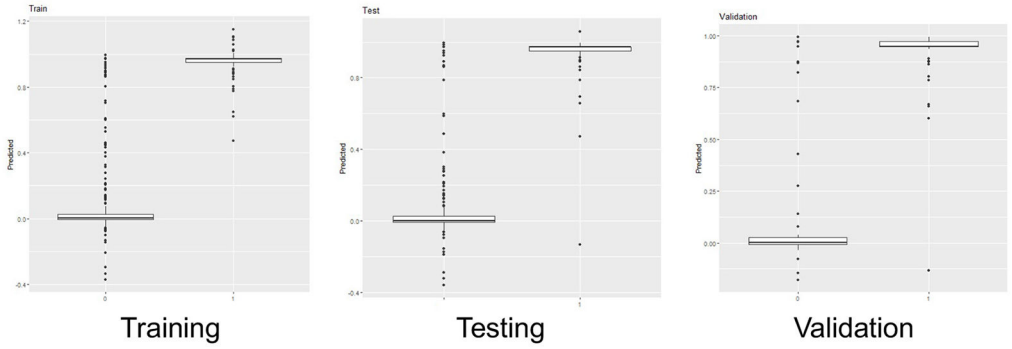
Subtype B and AP2 binding. A) Box plots showing the ranges of the 5 features used by ENNs for classification. B) Box plot showing training, testing, and validation ENNs using the top 25 features for AP2 binding. C) Cross tables of predicted vs. actual ability of ENN to identify subtypes using features applied to functional domains. D) Comparative signature pattern analysis (VESPA) among subtypes over the subtype-distinguishing domain. Only the domain positions where a > 50% difference appeared between the background (subtype A1) and query (subtype B, C, D) occurred are shown. Colored headings indicate the top five feature groups used by the ENN within the CD4 downregulation domain with yellow = hydrophobicity scale; green = HPLC or “other” scale; orange = polarity scale. The absolute score for each amino acid for each feature scale is shown. Red amino acids indicate the

background data set (subtype A1) and black values indicate the query data for subtypes B, C, and D. Gray boxes indicate the query and the background sequences agree at the specified position. Values in bold indicate that the amino acid occurred >80% of the time in the specified position.

A.



B.



C.

		Actual	
		Other	Type C
Predicted	Other	137/153=89.5%	1/72=1.4%
	Type C	16/153=10.5%	71/72=98.6%

D.

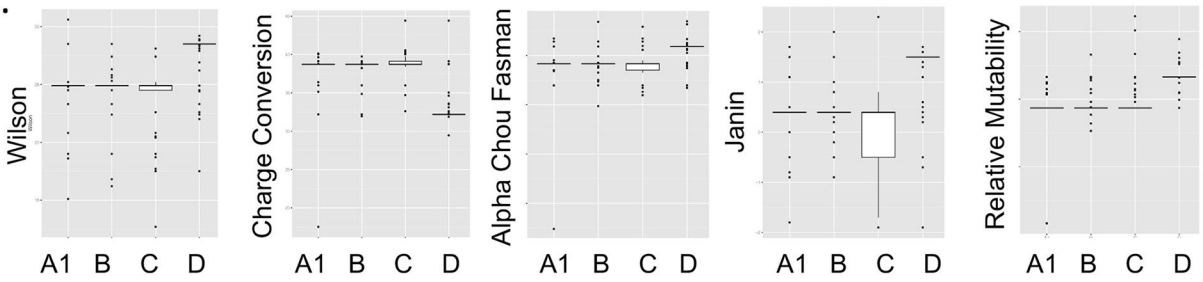
Domain	Position	Amino Acid				Rose				BullBrese			
		A	B	C	D	A	B	C	D	A	B	C	D
Alpha Helix A	2	G	G	A		0.72	0.72	0.74		0.81	0.81	0.61	
	4		L	F	H		0.85	0.88	0.78		-1.65	-1.52	0.69
	8	**H**	**H**	**F**	**H**	**0.78**	**0.78**	**0.88**	**0.78**	**0.69**	**0.69**	**-1.52**	**0.69**
Domain	Position	Surface Exposure				Cowen				Exchange			
		A	B	C	D	A	B	C	D	A	B	C	D
Alpha Helix A	2	2	2	2		0	0	0.35		4	4	4	
	4		3	3	1		1.8	1.69	-0.65		5	6	1
	8	**1**	**1**	**3**	**1**	**-0.65**	**-0.65**	**1.69**	**-0.65**	**1**	**1**	**6**	**1**

Fig. 6.

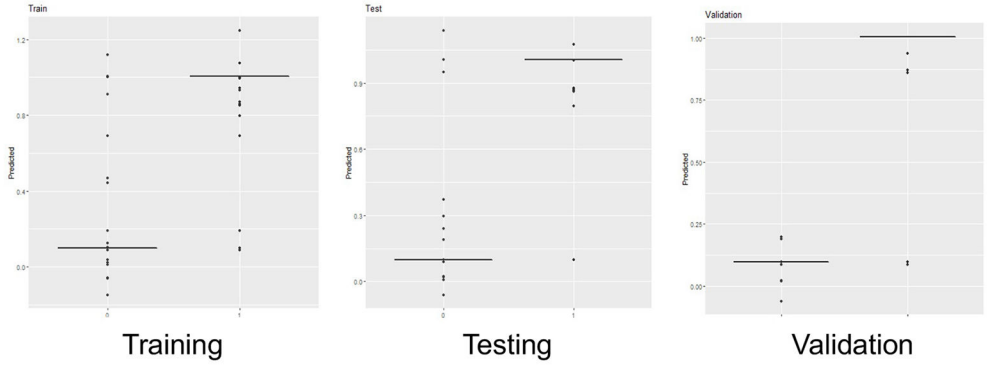
Subtype C and alpha-helix A. A) Box plots showing the ranges of the 5 features used by ENNs for classification. B) Box plot showing training, testing, and validation ENNs using the top 25 features for alpha-helix A. C) Cross tables of predicted vs. actual ability of ENN to identify subtypes using features applied to functional domains. D) Comparative signature pattern analysis (VESPA) among subtypes over the subtype-distinguishing domain. Only the domain positions where a > 50% difference appeared between the background (subtype A1) and query (subtype B, C, D) occurred are shown. Colored headings indicate the top five feature groups used by the ENN within the CD4 downregulation domain with yellow = hydrophobicity scale; green = HPLC or “other” scale; blue = structural scale. The absolute score for each amino acid for each feature scale is shown. Red amino acids indicate the background data set (subtype A1) and black values indicate the query data for subtypes B, C, and D. Gray boxes indicate the query and the background sequences agree at the

specified position. Values in bold indicate that the amino acid occurred >80% of the time in the specified position.

A.



B.



C.

		Actual	
		Other	Subtype D
Predicted	Other	178/178=100.0%	5/47=10.6%
	Subtype D	0/178=0.0%	42/47=89.4%

D.

Domain	Position	Amino Acid				Wilson				Charge Conversion			
		A	B	C	D	A	B	C	D	A	B	C	D
Beta Sheet A	3	**K**	**K**	**K**	**E**	**-3.6**	**-3.6**	**-3.6**	**-0.2**	**9.74**	**9.74**	**9.74**	**3.22**
Domain	Position	Alpha Chou fasman				Janin				Relative Mutability			
		A	B	C	D	A	B	C	D	A	B	C	D
Beta Sheet A	3	**1.16**	**1.16**	**1.16**	**1.11**	**-1.8**	**-1.8**	**-1.8**	**-0.7**	**56**	**56**	**56**	**102**

Fig. 7.

Subtype D and beta sheet A. A) Box plots showing the ranges of the 5 features used by ENNs for classification. B) Box plot showing training, testing, and validation ENNs using the top 25 features for beta sheet A. C) Cross tables of predicted vs. actual ability of ENN to identify subtypes using features applied to functional domains. D) Comparative signature pattern analysis (VESPA) among subtypes over the subtype-distinguishing domain. Only the domain positions where a > 50% difference appeared between the background (subtype A1) and query (subtype B, C, D) occurred are shown. Colored headings indicate the top five feature groups used by the ENN within the CD4 downregulation domain yellow = hydrophobicity scale; green = HPLC or "other" scale; purple = composition scale. The absolute score for each amino acid for each feature scale is shown. Red amino acids indicate the background data set (subtype A1) and black values indicate the query data for subtypes B, C, and D. Gray boxes indicate the query and the background sequences agree at the specified position. Values in bold indicate that the amino acid occurred >80% of the time in the specified position.

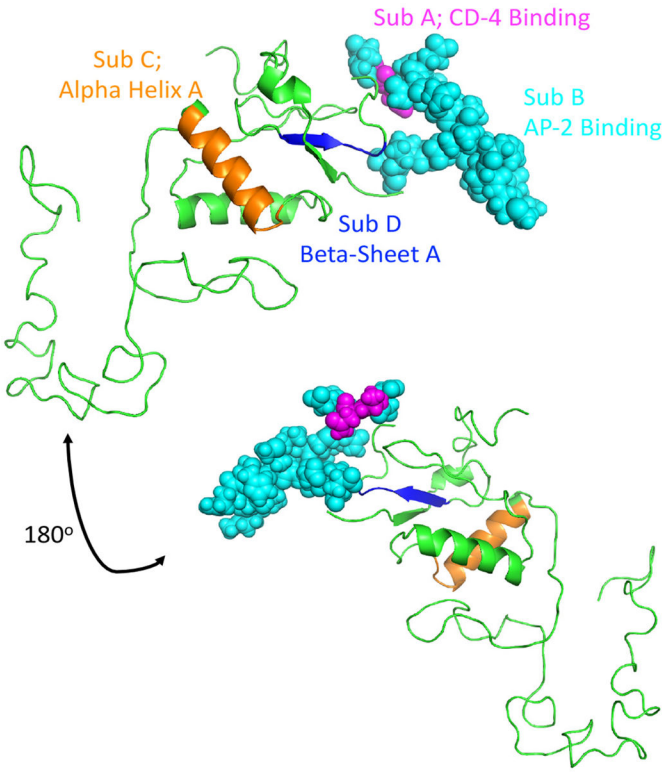


Fig. 8.

3D Nef protein structure highlighting subtype-specific domains identified via ENNs. Subtypes A and B Nef sequence populations have unique features in domains directly associated in interactions with CD4 and AP2. Subtypes C and D have unique domain features associated with Nef's structure, which could result in protein conformation and changes in the inter- or extracellular Nef interactions.

Table 1

Physicochemical and Categorical Scales (Features) Grouped by Class.

Class	Feature	
Size, Shape, Structure (***n*** **= 25)**	Molecular Weight	2D Propensity
	Beta Turn	Mass Membership Class
	Coil	Alpha Helix Levitt
	Average Flexibility	Antiparallel Beta Strand
	% Accessible Residues	Alpha Chou and Fasman
	Avg. Area Buried	Parallel Beta
	Recognition Factors	Surface Exposure
	Beta Reverse Turn	Beta Chou and Fasman
	Bulkiness	Transmembrane
	Beta-sheet Levitt	Alpha Helix
	Beta Strand	Chou and Fasman
	Volume	
	Mol. Frac. Of Buried Res.	
	Beta Sheet	
Polarity (*n* = 6)	Polarity	Grantham
	Polarity 2	Charge Polarity
	Charge (+1 for K,R; −1 for D,E)	Charge Conversion Amino Acid Composition
Composition (*n* = 3)	Amino Acid Composition	Swiss Prot
	Relative Mutability	
Hydrophobicity (*n* = 27)	Abraham and Leo	Welling
	Roseman	Parker
	Eisenberg	Aboderin
	Fauchere	Miyazawa and Jernigan
	Tanford	Sweet and Eisenberg
	Cowan and Whittaker	Guy
	Cowan	Wilson
	Meek	Manavalan and Ponnuswamy
	Black and Mould	Rao and Argos
	Kyte and Doolittle	Chothia
	Bull and Breese	Browne
	Wolfenden	Rose
		Hydrophobicity Membership
	Hopp and Woods	Class
	Janin	
HPLC and Other (*n* = 9)	Retention at pH 2.1	pKa Alpha
	HP Scale	Carboxylate
	pKa Amine	pI at 25 °C
	HPLC/TFA	Refractivity
	Exchange	

Table 2A

Features and Thresholds Used for the Three Separate Subtype Classification Experiments (all in the same ENN).

Dataset	Thresholds on Output Node	Sub-Selected Input Features
1	0 (A1) = x < 0.8; 1 (B) = 0.8 ≤ x < 1.3; 2 (C) = 1.3 ≤ x < 2.2; 3 (D) = x ≥ 2.2	α-helix B (Tanford); AP2 Binding (Polarity 2); α-helix A (Cowan); AP2 Binding (Retention at pH 2.1); CD4-II (HPLC/TFA); CD4-I (Charge); β-sheet B (Wilson); β-sheet A (Charge); β-sheet A (Wilson); β-sheet A (Charge Polarity)
2	0 (A1) = x < 0.9; 1 (B) = 0.9 ≤ x < 1.25; 2 (C) = 1.25 ≤ x < 2.2; 3 (D) = x ≥ 2.2	CD4-II (Coil); AP2 Binding (Retention at pH 2.1); α-helix A (Cowan); β-sheet A (Charge); α-helix A (pKA Alpha); β-sheet B (Wilson); β-sheet A (Exchange); AP2 Binding (Aboderin); AP2 Binding (Polarity 2); CD4-I (Retention at pH 2.1)
3	0 (A1) = x < 0.8; 1 (B) = 0.8 ≤ x < 1.3; 2 (C) = 1.3 ≤ x < 2.2; 3 (D) = x ≥ 2.2	CD4-I (Charge); α-helix A (Roseman); α-helix A (Wilson); α-helix A (Cowan); β-sheet A (Charge); AP2 Binding (Polarity 2); α-helix A (Wolfenden); CD4-II (Retention at pH 2.1); β-sheet B (Coil); CD4-II (Wilson)

Table 2B

Example Performance of Evolved Neural Networks performance over all subtypes. Performance for A) training, B) testing, and C) validation sets are shown.

Training	**Actual**				
		Subtype A	Subtype B	Subtype C	Subtype D
	Subtype A	90.7%	1.8%	0.2%	1.2%
Predicted	Subtype B	6.7%	94.0%	1.9%	8.0%
	Subtype C	2.5%	3.9%	90.0%	1.6%
	Subtype D	0.0%	0.2%	7.8%	89.1%
Testing	**Actual**				
		Subtype A	Subtype B	Subtype C	Subtype D
	Subtype A	89.6%	2.2%	0.0%	1.6%
Predicted	Subtype B	9.1%	93.8%	2.7%	7.1%
	Subtype C	0.9%	3.9%	90.9%	1.6%
	Subtype D	0.0%	0.0%	6.4%	89.7%
Validation	**Actual**				
		Subtype A	Subtype B	Subtype C	Subtype D
	Subtype A	87.2%	6.0%	0.0%	2.1%
Predicted	Subtype B	10.2%	89.5%	2.8%	6.4%
	Subtype C	2.5%	4.5%	91.7%	6.4%
	Subtype D	0.0%	0.0%	0.5%	85.1%

Table 3

Top 25 Features by Subtype that Separated Classes. All significance values were $p < 10^{-20}$.

Subtype A1		Subtype B	
Domain	**Feature**	**Domain**	**Feature**
Beta Sheet B	Coil	Nef Loop	Grantham
CD4 Downregulation	Charge	AP2 Binding	Retention
CD4 Downregulation	Charge Conversion	Alpha CD	Guy
CD4 Downregulation	Welling	AP2 Binding	Grantham
CD4 Downregulation	Retention	AP2 Binding	Hplctfa
CD4 Downregulation	Abraham-Leo	AP2 Binding	Meek
CD4 Downregulation	Wilson	AP2 Binding	Browne
CD4 Downregulation	Cowan	Alpha CD	Rose
Beta Sheet B	Wilson	AP2 Binding	Abordin
Alpha Helix B	Tanford	Alpha CD	Average Flex
Alpha Helix B	Number of Codons	AP2 Binding	Roseman
Nef Loop	Parker	AP2 Binding	Refractivity
CD4 Downregulation	Accessible Residues	Nef Loop	Hplctfa
Beta Sheet B	Parallel Beta	Alpha CD	Ave. Area Buried
Alpha Helix B	Eisenberg	Alpha CD	Accesible Residues
Cytokines Binding	Eisenberg	Alpha CD	Charge
Beta Sheet B	Manavalan	AP2 Binding	Cowan Whittaker
Nef Loop	Wilson	AP2 Binding	Wolfenden
Nef Loop	Cowan Whittaker	Alpha CD	Polarity
CD4 Downregulation	Janin	Alpha Helix B	PI
CD4 Downregulation	Welling	Alpha CD	Beta Sheet Levitt
CD4 Downregulation	Browne	Alpha CD	Abrahan-Leo
CD4 Downregulation	2D Propensity	Alpha CD	Janin
CD4 Downregulation	ChouFausman	Alpha CD	Black-Mould

Subtype C		Subtype D	
Domain	**Feature**	**Domain**	**Feature**
Alpha Helix A	Guy	Beta Sheet A	Accessible residues
Alpha Helix A	PKA Alpha	Beta Sheet A	Welling
Alpha Helix A	Exchange	Beta Sheet A	Janin
Alpha Helix A	Hydrophobicity	Beta Sheet A	Charge Conversion
Alpha Helix A	Black Mould	Beta Sheet B	Recognition Factors
Alpha Helix A	Surface Exposure	Beta Sheet A	Charge Conversion
Alpha Helix A	Chothia	Beta Sheet A	PI
Alpha Helix A	Cowan	Beta Sheet A	Cowan
Alpha Helix A	Eisenberg	Beta Sheet A	Wilson
Alpha Helix A	BullBreese	Beta Sheet B	PI
Alpha Helix A	Wilson	Beta Sheet A	Abraham-Leo

Subtype C		Subtype D	
Domain	**Feature**	**Domain**	**Feature**
Alpha Helix A	Aboderin	Beta Sheet A	Renetion
Alpha Helix A	Abraham-Leo	Beta Sheet B	PKA Alpha
Alpha Helix A	Meek	Beta Sheet A	Relative Muatability
Alpha Helix A	Tanford	Beta Sheet B	Charge
Alpha Helix A	Roseman	Beta Sheet B	Charge Conversion
Alpha Helix A	Miyazawa	Beta Sheet A	Volume
Alpha Helix A	Fauchere	Beta Sheet A	Charge Polarity
Alpha Helix A	Charge Polarity	Alpha CD	PKA Alpha
Alpha Helix A	Welling	Beta Sheet B	Alpha Chou Fasman
Alpha Helix A	Retention	Beta Sheet A	Alpha Chou Fasman
Alpha Helix A	Parker	Beta Sheet A	Alpha Helix
Alpha Helix A	Cowan Whittaker	Alpha CD	PI
Alpha Helix A	HP Scale	Beta Sheet B	Volume

Table 4

(A-D). Average Subtype Specific Classification Accuracy across all Subtypes. A) Subtype A1 vs. Subtype (B,C,D), B) Subtype B vs. Subtype (A1,C,D). C) Subtype C vs. Subtype (A1,B,D), D) Subtype D vs. Subtype (A1,B,C).

	Subtype A1	**Subtype (B,C,D)**
Train	92%	79%
Test	91%	80%
Validation	86%	80%
	Subtype B	**Subtype (A1,C,D)**
Train	92%	83%
Test	89%	84%
Validation	90%	89%
	Subtype C	**Subtype (A1,B,D)**
Train	100%	90%
Test	98%	92%
Validation	98%	89%
	Subtype D	**Subtype (A1,B,C)**
Train	91%	100%
Test	92%	99%
Validation	89%	100%