

IDEPI: rapid prediction of HIV-1 antibody epitopes and other phenotypic features from labeled genetic sequence data using a flexible machine learning platform.

N Lance Hepler¹, Konrad Scheffler², Steven Weaver², Ben Murrell², Douglas D Richman^{2,3}, Dennis R Burton^{4,5}, Pascal Poignard⁴, Mojgan Haddad⁶, Davey M Smith^{2,3}, Sergei L Kosakovsky Pond^{2,*}

1 Interdisciplinary Bioinformatics and Systems Biology Program, University of California San Diego, La Jolla, California, USA

2 Department of Medicine, University of California San Diego, La Jolla, California, USA

3 San Diego Veterans Affairs Healthcare System, San Diego, California, USA

4 The Scripps Research Institute, La Jolla, California, USA

5 Ragon Institute of MGH, MIT, and Harvard, Boston, Massachusetts, USA

6 Monogram Biosciences, South San Francisco, California, USA.

* **E-mail: Corresponding spond@ucsd.edu**

Abstract

Since its identification in 1983, HIV-1 has been the focus of a research effort unprecedented in scope and difficulty, whose ultimate goals — a cure and a vaccine – remain elusive. One of the fundamental challenges in accomplishing these goals is the tremendous genetic variability of the virus, with some genes differing at as many as 40% of nucleotide positions among circulating strains. Because of this, the genetic bases of many viral phenotypes, most notably the susceptibility to neutralization by a particular antibody, are difficult to identify computationally. Drawing upon the open-source general-purpose machine learning algorithms and libraries, we have developed a software package IDEPI (IDentify EPItopes) for learning predictive models of viral phenotypes from sequences with known phenotypes. IDEPI can apply learned models to classify sequences of unknown phenotypes, and also identify specific sequence features which contribute to a particular phenotype. The program is written in Python, and designed to be flexible, so that new machine learning methods and sequence features can be readily added to accommodate various applications not necessarily limited to HIV-1. We demonstrate that IDEPI achieves performance similar to or better than that of previously published approaches on four well-studied problems: finding the epitopes of broadly neutralizing antibodies (bNab), determining coreceptor tropism of the virus,

identifying compartment-specific genetic signatures of the virus, and deducing drug-resistance associated mutations.

Introduction

The challenge of predicting a viral phenotype from sequence data has many motivating examples in HIV-1 research. In this work we restrict our attention to predicting binary phenotypes, e.g. resistant *vs* susceptible. Perhaps the most established application is that of determining whether or not the viral population in a particular host harbors drug resistance associated mutations (DRAMs) [1]. Algorithms for inferring this from viral genotype alone (e.g. [2]) are well established and used both in research [3] and in clinical practice [4]. These algorithms have been developed based on large training sets using phenotypic assays, for example those measuring half maximal inhibitory concentration (IC_{50}) of an antiretroviral drug (ARV) [5] to label sequences resistant or susceptible. For many ARVs, the genetic basis of resistance is simple, and consists of specific point mutations [1], making it possible to distinguish resistant viruses from their susceptible counterparts by the presence or absence of a specific residue or a set of residues, leading to reliable prediction [6, 7]. For others ARVs including some protease, integrase, nucleoside reverse transcriptase inhibitors, and co-receptor antagonists, the resistance phenotype is defined by the interaction of many sites [8–12], or structurally [13, 14], presenting an opportunity to continue methodological development (e.g. [15–17]).

Another popular prediction problem is that of determining which of the two cellular co-receptors needed for HIV-1 fusion with the target cell and infection can be used by a particular viral strain. The ability of a virus to bind CCR5 (R5-tropic), CXCR4 (X4-tropic), or either (dual-tropic) determines the efficiency with which it can infect different types of target cells [18], predicts whether or not certain ARVs will be effective [19], and impacts the course of disease progression [20]. The primary determinant of V3 tropism is thought to be the third variable loop (V3) of the envelope glycoprotein (*env*) [21], which spans approximately 35 amino-acid residues. Specialized assays can be used to determine the tropism of a virus with a particular *env* protein [22], providing both the training sets and the gold standard against which computational prediction methods can be compared [23, 24]. Starting with the work by Fouchier and colleagues in 1992 [25], which used the computed total charge of V3 to derive and experimentally validate the simple 11/25 rule (if residues at sites 11 and 25 are positively charged, then

the virus is classified as X4 tropic), numerous authors have applied decision trees [26], random forests [27], position-specific scoring matrices [28], support vector machines (SVM) [26], neural networks [29], Bayesian networks [30], and hybrid models [31] to the problem. Various feature engineering approaches including using structural information [32], electrostatic hulls [27], sequence motifs [28], positional and segment residue frequencies [31] have also been attempted. At present the best methods achieve accuracy on the order of 85% on comprehensive training datasets, thereby justifying ongoing research to improve this value [33].

A different class of applications of machine learning techniques arises when researchers seek to infer genetic “signatures” of HIV-1 isolates from different anatomical compartments (e.g. blood vs cerebrospinal fluid [34]), individuals with different clinical attributes (e.g. those with and without neurocognitive impairment [35]), and different disease stages (e.g. acute vs chronic infection [36]). Once again, the interest is both in prediction for unlabeled sequences, for example to modify treatment before impairment occurs [35], and in finding predictive features, for instance to target vaccine research towards HIV-1 strains that are more likely to establish new infections [36].

One of the most promising avenues of HIV-1 vaccine research provides our final example of genotype to phenotype association problems, and the one that IDEPI was specifically developed to address. Rational HIV-1 vaccine design has been greatly advanced by the isolation and identification of broadly neutralizing antibodies (bNab), typically from chronically infected individuals [37]. By definition, a bNab is able to neutralize (in experimental assays) a large proportion of reference viruses (e.g. [38–40]), and understanding which epitopes are being targeted can reveal “conserved” elements shared by many circulating viruses, and help design a vaccine which elicits responses to the same epitopes [41]. While powerful and illuminating, current biochemical and structural techniques for mapping bNab epitopes (e.g. [39, 40, 42]), are expensive, time consuming, and do not necessarily lead to good predictive models (e.g. [43]). The appeal of computational epitope prediction lies in generating hypotheses for experimental validation, and in high-throughput screening of sequences with unknown resistance phenotypes. Fortunately, as a byproduct of screening for bNab, large panels of phenotypic (IC_{50}) and matched envelope sequences have been generated, and several recent efforts [44–46] have been directed at applying machine learning techniques to these data in order to predict the resistance phenotypes of HIV-1 strains, and to infer antibody epitopes.

Given the clear utility, relevance, and popularity of tools for predicting viral phenotypes, under-

standing sequence features that determine them, the continued development and refinement of machine learning techniques, and expanding training data sets, we designed IDEPI – a domain-specific and extensible software library for supervised learning of models that relate genotype to phenotype for HIV-1 and other organisms. IDEPI makes use of open source libraries for machine learning (scikit-learn, scikit-learn.org/), sequence alignment (HMMER, hmmerr.janelia.org/), sequence manipulation (BioPython, biopython.org), and parallelization (joblib, pythonhosted.org/joblib), and provides a programming interface to allow the users to engineer sequence features and select machine learning algorithms appropriate for their application. Our original motivation for developing IDEPI was to predict the epitopes of broadly neutralizing antibodies (hence the name), but we believe that the program can find many diverse application in the field of HIV-1 research, and more generally for predicting discrete phenotypes from genetic sequence data.

IDEPI is *powerful and accurate*: when we compare its performance with that of specialized tools on four classes of problems outlined above, we find that even without feature and machine learning method tuning, IDEPI closely hews to or even outperforms existing methods on the same data. IDEPI infers *biologically meaningful features*: for each studied problem IDEPI identified many or most of the genetic sequence features that have been previously shown to affect phenotype. IDEPI is *convenient*: by standardizing data manipulation, e.g. aligning sequences to standard reference coordinates, extracting features to be modeled, reading and handling phenotype annotation, and providing means to save learned models and easily reuse them later, IDEPI can empower researchers interested in tackling new problems to focus on innovation, instead of “plumbing”; IDEPI makes tasks like retraining a classifier on different data sets trivial – something that is difficult to impossible to do with many published algorithms. IDEPI is *fast*: automatic parallelization of independent tasks (e.g. cross-validation) on multi-core architectures greatly accelerates model learning and performance evaluation; for the default linear support vector machine (LSVM) classifier, classification of new sequences given a model can be done at a rate of $10^4 - 10^5$ sequences per minute, making the program suitable for the analysis of next generation sequencing data. IDEPI is *customizable*: different machine learning algorithms implemented in scikit-learn can be used; new sequence features can be defined using a well-specified application programming interface (API); various feature selection approaches (e.g. forward or backward selection) can be used; performance can be optimized with respect to many metrics (e.g. sensitivity).

Design and Implementation

IDEPI architecture and dependencies

IDEPI is implemented in the Python 3 programming language, and leverages open-source and community-developed libraries to implement reusable functionality: BioPython for biological sequence data structures and for parsers of FASTA- and Stockholm-format files; NumPy (numpy.org) and SciPy (scipy.org) for vector, matrix, and other common numerical recipes; and scikit-learn (scikit-learn.org) for various machine-learning functionality. When extending the facilities provided by these libraries, IDEPI provides compatible application programming interfaces so that its components are reusable and similarly extensible.

IDEPI accepts two forms of input data – a specially-crafted SQLite database (sqlite.org) or a combination of FASTA-formatted sequences with supplemental phenotypic data in comma-separated value (CSV) format. These input data are transformed by IDEPI into a multiple sequence alignment (MSA) using HMMER (version 3.1b1). Because the authors of HMMER recommend providing amino-acid sequences to the program, IDEPI will by default translate the input sequences if they are determined to have a DNA alphabet. A user-provided reference MSA is modeled by HMMER to guide the construction of a MSA from the input data. Additionally, IDEPI includes a user-provided reference sequence in the alignment to label the columns of the MSA in a conventional manner (e.g. N332 for an asparagine at site 332).

Feature extraction techniques included with IDEPI

For feature extraction, IDEPI provides four classes (all scikit-learn compatible) for the vectorization of labeled MSAs: one for sequence identity at a given site (e.g. N301N), optionally reduced (e.g. Stanfel encoding, N301[DENQ] [47]); one the for sequence identity at two sites within a specified radius (e.g. N301N+S334S), also optionally reduced; one for the identification of sequence motifs describable by regular expression (e.g. potential N-linked glycosylation sites (PNGS), using the regular expression “N[[^]P][TS][[^]P]”); and another for motifs at two sites. For label extraction, IDEPI provides a class which converts phenotype data to a form usable by scikit-learn.

Feature selection and learning algorithms used by IDEPI

To enable rapid learning and prevent overfitting, IDEPI performs feature selection using the minimum redundancy maximum relevance (mRMR) algorithm [48]. mRMR greedily chooses new features that are maximally informative of our label while also minimally informative of already-chosen features. This is implemented in the separate `sklrmr` package, also scikit-learn compatible, and uses Cython (`cython.org`) for high performance.

Default model learning is implemented using a soft-margin, linear support vector machine. The soft-margin parameter, C , is chosen by (inner) grid search. Both of these functions are implemented within scikit-learn, and parallelized when possible.

Tools included with IDEPI

IDEPI provides three scripts for end users not wishing to directly program their own pipelines. “`idepi discrete`” accepts labeled sequence data and will generate a MSA from these data, extract features and labels, perform N-fold cross-validation on models built from a pipeline of mRMR and soft-margin linear SVMs, and finally report the models’ performance along with the labels of the most frequently selected features and their relationship to the models (e.g. is the presence or absence indicative of the feature indicative of an outcome). “`idepi learn`” will similarly accept labeled sequence data, learn a model, and save it to disk for later use. Lastly, “`idepi predict`” accepts a saved model and some unlabeled data and will predict its label. All the results presented in the manuscript have been generated using these three scripts.

Extensible API for feature engineering

IDEPI defines a `LabeledMSA` class as a wrapper around BioPython’s `MultipleSeqAlignment` for the column-wise labeling of a MSA. Together with classes facilitating alphabet encoding, IDEPI provides simple facilities enabling rapid feature engineering for biological sequence data. Examples of how these facilities can be used can be found within IDEPI’s source code – the `SiteVectorizer` and `MotifVectorizer` classes for feature extraction. Additionally, motif features are trivially supported by the `MotifVectorizer` class, which accepts a regular expression argument describing the motif. IDEPI uses this functionality to extract features for putative N-linked glycosylation sites, using the regular expression described above.

Results

Because IDEPI is designed to be an extensible general framework for predicting binary phenotypes, we first tested it on simulated data, and well-studied problems of drug-resistance and tropism prediction and detection of associated genetic features. The large number of published methods make a comprehensive comparison infeasible, hence we selected methods based on their popularity, recency, performance, and the availability of training data. IDEPI was evaluated for (i) its performance in phenotype prediction using standard cross-validation metrics, and on previously published independent datasets; and (ii) the veracity of inferred genetic features inferred to be informative of a particular phenotype. All the datasets and instructions needed to run them with IDEPI are provided with the package distribution.

Simulated data

In order to establish baseline performance of IDEPI where the true “phenotype” is known we simulated the evolution of $N = 241$ HIV-1 protein envelope sequences subject to a directional selective pressure applied to sites in an epitope. For this moderate size data set (chosen to represent a typical bNab training set), IDEPI performs very well (Table 1) overall, both in terms of classification performance, and in recovering the locations/residue identity of epitopes. In the simplest case, when a single mutation in a 5-site epitope confers resistance, IDEPI delivers Matthew’s Correlation Coefficient (MCC) of 0.98, and recovers $> 50\%$ of sites within epitopes if they are sufficiently variable. Because positions in epitopes are likely quite correlated, mRMR redundant feature selection explains essentially all the signal with a median of 2 features per replicate. For a fixed training data set size, with the increased epitope length and complexity, the performance degrades predictably, but MCC remains excellent for intermediate epitope complexity (0.94) and good (0.78) for high epitope complexity. Encouragingly there seems to be no undue false positive results due to phylogenetic relatedness of the samples: IDEPI yields a median MCC of 0.04 (MCC of 0 corresponds to “no-better than random prediction” performance) for randomly assigned phenotypes.

Drug resistance

We used a large publicly available data set associating viral sequences (reverse transcriptase) with IC_{50} values for the PhenoSense assay (available from the Stanford HIV Drug Resistance Database,

hivdb.stanford.edu) to train an IDEPI classifier for resistance to a non-nucleoside reverse transcriptase inhibitor nevirapine (NVP). We chose this drug as a test case because (i) the basis for its resistance is well understood making the assessment of IDEPI predictions easy; (ii) testing for NVP resistance is biomedically relevant, for example in the context of preventing mother to child HIV-1 transmission; (iii) a recent study [49] used resistance data from the Stanford database to train specialized classifiers for NVP resistance providing a basis for comparison.

Using 80 features IDEPI achieves the same accuracy (0.92, Table 2) as a state-of-the-art custom-built prediction tool using structural information [49]. The first three selected features (Table 3), correspond to three canonical sites of strong phenotypic resistance, and the maintenance of the wildtype residue at each of the positions is strongly predictive of susceptibility –a classifier built on just these three features achieves MCC of 0.74, compared to that of 0.83 for the 80-feature model. Other genetic features implicated in the development of NVP resistance recovered by the IDEPI model include major resistance mutations K101P, K103N, V106A, Y181C, Y188L, G190A, and accessory/weak resistance mutations L100I, E138Q, H221Y, and V108V, P236P (the latter two associated with susceptibility) [1].

We compared the SVM model learned by IDEPI against perhaps the most commonly used drug resistance prediction algorithm –the Stanford HIVdb (expert curated, and evidence based) [2], using a large dataset collected in Mexico for the purposes of drug resistance surveillance [50]. Because no phenotype measurements were available for these sequences (as is common in practice), we computed the degree of concordance between IDEPI and HIVdb using Cohen’s κ [51]. Since surveillance HIV-1 pol sequences are obtained from mixed viral populations and often contain ambiguous bases, not directly handled by default IDEPI feature sets, we considered all possible amino-acid resolutions of nucleotide level ambiguities at the positions involved in the model, and called a sequence resistant if any of the resolutions were predicted as resistant. The two methods of resistance prediction were in excellent agreement overall ($\kappa = 0.85$), including all cases of “highly-resistant” sequences, on par with the numbers reported in a recent comparison of several rule-based resistance prediction algorithms [52].

Co-receptor usage/tropism

In 2010, Dybowski et al [27] presented a sophisticated multi-level classifier including structural and biochemical properties of the V3 loop, performed extensive training and validation of their approach, and compared it to previous work. Because a large training data set of V3 amino-acid sequences and

associated phenotypic measurements was provided as a part of the publication, we were able to train an IDEPI classifier on the same data to enable a direct comparison.

As has been documented before (e.g. [27]), most of the predictive power of V3 sequences is captured by only a few features – in the case of IDEPI, a model using only two features already achieves an MCC of 0.67, while the full model with 90 features improves it to 0.78. The first selected feature is a potential N-linked glycosylation site (PNGS) at position 301; several sites in this 4-residue motif have been implicated as critical to CCR5 receptor binding [53], hence a single composite feature is able to encapsulate the discriminating positions for many sequences. The second feature is one of the two residues in the 11/25 rule [25]; interestingly, the two positions are sufficiently correlated in the training sample that mRMR feature selection eliminates position 25 once 11 has been included. IDEPI appears to be surprisingly well suited to the problem of tropism prediction, and delivers nearly the same accuracy (0.94 vs 0.96, the latter number obtained in the original publication by tuning algorithmic cutoffs to maximize accuracy on the training data) as the much more complex feature engineering approach undertaken by Dybowski and colleagues. Furthermore, on an independent dataset, IDEPI attains accuracy of 0.905, whereas the best of the 5 methods compared previously [27] attained accuracy of 0.86.

HIV-1 associated dementia

A recent comprehensive study by Holman and Gabuzda [35] applied a machine learning pipeline (based on decision trees) to partial envelope sequences to identify signatures (defined as collections of residues or biochemical properties at specific genomic positions) of sequences isolated from brain tissue of subjects who developed HIV-1 associated dementia (HAD). Since the training set of sequences and corresponding diagnoses has been kindly made available by the authors through the HIV Brain Sequence Database [54], it was straightforward to apply IDEPI to the same data to learn a classifier. The Holman and Gabuzda study also included an independent validation data set of 10 individuals diagnosed with HAD, and we used it here to test the learned model.

IDEPI excels at this classification problem, with specificity and sensitivity exceeding 0.9, and accuracy of 0.95. The original authors reported accuracy of 0.75, but their model was restricted to a subset of the available sequence length, HXB2 coordinates 265-369. As with previous two applications, a single prominent feature (T297K) attains MCC of 0.57; unlike the other problems, the next four features appear to be of about the same informative content (based on the order in which they are selected in

cross-validation folds), and MCC performance increases gradually as the features are added (Figure 2). Interestingly, features reported as associated with HAD previously (see [35] for a summary), are not added to the model until later: for example site 283 is the 8th ranked feature, site 308 – the 38th, and site 304 – the 65th. Furthermore, the 90-feature IDEPI correctly classifies all 10 individuals in the validation data set, whereas the original method correctly classified 8/10 cases.

Broadly neutralizing antibodies

Because IDEPI was designed with the specific problem of finding bNab epitopes and predicting the resistance phenotype from sequence data, we compared its performance against three recently published machine learning approaches to solving same problem.

1. Gnanakaran et al [44] proposed and tested an ensemble framework combining pattern analysis and logistic regression to predicting the neutralization phenotype and mapping the epitopes of the b12 bNab [55] which targets the CD4 receptor binding site [56]. We used the genotypic and associated phenotypic data from this paper to train and test an IDEPI classifier for the b12 bNab.
2. West et al [45] applied a direct optimization (implemented in the Antibody Database program [ADP]) to predict the continuous IC_{50} value using sequence based features and applied in to data from 25 antibodies. We compared the predictions derived by IDEPI models for some of the same antibodies (chosen to represent one of the the remaining three types of bNab classified by their targets [56]) using publicly available neutralization assay data, distributed with IDEPI, or the training data set from [46]
3. Chuang et al [46] developed an epitope feature selection based on evaluating various measures based on mutual information between sequence sites and IC_{50} values –an idea shared and extended by the mRMR approach. We used the genotype and phenotype data for two of the antibodies (8ANC131 and 8ANC195, the latter also studied by West et al) whose epitopes were mapped and experimentally confirmed by Chuang et al.

b12 bNab prediction Unlike the previous three cases, b12 epitope prediction results in both a simpler model (only 5 features) and a considerably lower performance (Table 2), with MCC of only 0.35. Comparing IDEPI with the ensemble method developed by Gnanakaran et al [44], IDEPI achieves lower

accuracy on the training data (note that the original reference does not report a cross-validation value), but higher accuracy on validation data (Table 4), suggesting that the ensemble model may have been over-fitting the training data. Also unlike other classification problems, only a single residue (D185D, Table 3) is supported by the majority of cross-validation folds. Taken together, these results suggest that the training data set is too small (or that the feature set is suboptimal) to reliably identify the complex structurally-defined epitope for b12, although IDEPI seems to outperform a previously published method on an independent validation dataset, and its 5-feature epitope includes residue 424 which is a part of the CD4 binding site [57] targeted by the antibody.

Other broadly neutralizing antibodies

1. **PG9** is a broadly neutralizing antibody targeting the V1/V2 loop in HIV-1 env [58], whose canonical epitope is anchored by the PNGS at position 160, which is also the single most important position in the 30-feature model fitted by IDEPI. Due to the relatively small training data set (174 sequences), an MCC of 0.42 is achieved, with the model showing fairly low sensitivity (0.49, Table 2). The 30 feature model has a remarkably high accuracy on the training data (0.96), but the small number of resistant sequences in it makes it difficult to generalize the features past the N160 (Figure 2). A direct comparison with West et al is difficult to formulate, because ADP’s performance is measured by the proportion of IC_{50} variance explained by the model, which cannot be measured for IDEPI. However, in terms of relevance of epitope features, IDEPI appears to produce a more useful result. Indeed, it finds the three features found by ADP, but ranks them differently (more in agreement with the structural studies): G732G (resistant, ADP: strong support, IDEPI: mean feature rank 18/30), PNGS (N160) (susceptible, supported by structure [58], ADP: intermediate support, IDEPI: mean feature rank 1/30), and K171K (susceptible, supported by structure [58], ADP: strong support, IDEPI: mean feature rank 11/30). Further, IDEPI places another structurally confirmed residue in the inferred epitope: V169E (resistant, mean rank 8/30).
2. **PGT-121** is a broadly neutralizing antibody targeting glycans in the V3 loop [40]. IDEPI infers a single feature model (Table 3), which associates the presence of a pair of PNGS (at positions 301 and 332) as strongly predictive (MCC=0.58) of susceptibility. Interestingly, while PNGS (N332) is the key part of the canonical PGT-121 epitope, PNGS (N301) – previously thought relatively unimportant – appears to act together with N332 to effect PGT-121 binding [59]. ADP predicts the

importance of PNGS (N332), but also lists four other sites whose role in antibody-virus interaction is unclear, and does not report N301 as important.

3. **10E8** is a broadly neutralizing antibody targeting the Membrane-Proximal External Region (MPER) [60], showing unusual potency on the reference panel viruses. Because of this, the training sample (Table 2) includes only 4% of resistant sequences, making meaningful learning difficult, as evidenced by the low MCC of 0.23, and poor sensitivity. There are no top ranked features in the model (the ranking changes significantly between cross-validations, Table 2), but one of the structurally defined epitope sites (T676T) is included among the top 5, whereas ADP finds no such sites and also performs poorly. The relevance of other inferred model features associated with resistance, e.g. PNGS(T413+E824), K171E and E153Q is questionable, and larger training datasets containing more resistant samples are needed for computational prediction to improve.
4. **8ANC195** is a broadly neutralizing antibody whose epitope has not been structurally confirmed [61], but it was used a test case for computational epitope prediction and experimental confirmation by two independent groups [45, 46]. IDEPI achieves a good MCC of 0.67 on the training data from Chuang et al, and does so with only two features in the epitope: two pairs of PNGS sites (Table 3). The top feature is that the absence of either a PNGS anchored at site 234 or that anchored at 276 confers resistance. This single pair of PNGS subsumes three features (N234, N276, and T236) experimentally validated by previous work. This example highlights that feature engineering (pairs on PNGS) provides a more parsimonious and powerful model of neutralization that either single PNGS [45], or single residues [46] can. The second feature selected by IDEPI is another pair of PNGS (N160 and N230), which is predicted to confer resistance, and does so at a weak level [45].
5. **8ANC131** is a broadly neutralizing antibody whose epitope has been structurally mapped, but not yet published [46], and the same authors performed computational prediction of epitope sites and tested them experimentally. Unlike 8ANC195, where the epitope features are clean and experimentally confirmed, computational predictions have not been found nearly as useful, with the top sites conferring only marginal resistance [46]. IDEPI finds a diffuse signal for 15 features (Figure 2, Table 2), and MCC of 0.19 (the lowest of all problems). There seems to be little overlap between the features found in 3 or more cross-validation folds (susceptible: K151G, V169R, resistant: N463K, D474N, PNGS(N339+Q442), PNGS(142a+N234)), and those reported by [46] [top 10:

456,78,79,466,280,326,96,80,282,461] although many are in the same region of the three-dimensional structure.

Availability and future directions

IDEPI and sklmrmr are installable via the PyPI Python package system, through all the standard tools (`easy_install/pip`), and their source code is available on GitHub (`github.com/veg/idepi` and `github.com/nlhepler/sklmrmr`). A Virtual Machine for Oracle’s VirtualBox has also been built to provide easy access to IDEPI for users unfamiliar with the intricacies of Python package management, and is available from the main package distribution page.

IDEPI will likely be extended in the future to include a larger array of built-in feature extraction mechanisms. In the future, we intend to release an update that includes a feature extractor that maps sequence data to a provided structure to perform “patch analysis”. Downstream users that build novel feature extractors are recommended to submit their creations to IDEPI, via GitHub’s pull request mechanism, for inclusion in a future release. Additionally, in providing APIs compatible with BioPython and scikit-learn, IDEPI will prove ever more useful as advances are made in those fast-moving software packages.

Supplementary Methods

Simulated data

We simulated the evolution of $N = 241$ HIV-1 protein envelope sequences subject to a directional selective pressure applied to sites in an epitope using the HyPhy package [62]: the reference HXB2 sequence was evolved along a phylogenetic tree representing the diversity of circulating HIV-1 group M strains (inferred from biological isolates), subject to an HIV-1 specific substitution model [63], with site-to-site substitution rate heterogeneity modeled by a 3-bin general discrete distribution [64]. The development of resistance to a particular simulated epitope in a subset of sequences (defined as a set of positions in the genome and “escape” residue), was modeled by accelerating the rate of amino-acid substitution towards the escape residue along the terminal tree branch leading to a resistant sequence. For each replicate (100 replicates per set), an epitope of desired complexity was generated (Table 1), and each simulated

sequence was assigned a phenotype. We also performed a simulation where phenotypes were assigned to sequences randomly, in order to establish the degree to which phylogenetic relatedness can drive spurious associations due to the non-independence of samples [44].

Drug resistance

We labeled a sequence resistant to NVP if the measured fold change in IC_{50} was 5 or greater. A feature was reported if it appeared were selected in 3 or more out of 5 cross-validation replicates. We investigated the complexity of the genotypic basis of resistance by a simple grid search (the number of features was one of the following values: 1,2,3,4,5,10,15,20,25,30,35,40,50,60,70,80,90,100 see Figure 2)

Co-receptor usage/tropism

The number of features maximizing 5-fold cross-validation MCC was determined by a simple grid search. In addition to cross-validation performance metrics, we compared the performance of the IDEPI model to the methods considered by Dybowski et al [27] on an independent validation dataset with 74 sequences.

Broadly neutralizing antibodies

IDEPI labeled sequences with IC_{50} of $\geq 20\mu g/ml$ for a given bNab as resistant, except for the 10E8 bNab(which shows unusually low titers for the reference panel), where the threshold was lowered to $5\mu g/ml$. The number of features maximizing 5-fold cross-validation MCC was determined by a simple grid search (as before, the number of features was one of: 1,2,3,4,5,10,15,20,25,30,35,40,50,60,70,80,90,100).

Computational Resources and Software Versions

All experiments were performed with IDEPI v0.17, sklrmr v0.2.0, scikit-learn v0.14.1, scipy v0.12.0, numpy v1.7.1, BioPython v1.62, and Python v3.3.3 on an Apple MacBook Pro (Quad-core 2.6GHz Intel Core i7) running Mac OS X 10.9.1.

Acknowledgments

This research was supported in part by the National Institutes of Health (AI47745 and AI57167, AI74621, GM093939), the International AIDS Vaccine Initiative (AI090970), the UC Laboratory Fees Research

Program grant 12-LR-236617 ...

References

1. Tang MW, Shafer RW (2012) HIV-1 antiretroviral resistance: scientific principles and clinical applications. *Drugs* 72: e1-25.
2. Liu TF, Shafer RW (2006) Web resources for HIV type 1 genotypic-resistance test interpretation. *Clin Infect Dis* 42: 1608-18.
3. Poon AFY, Aldous JL, Mathews WC, Kitahata M, Kahn JS, et al. (2011) Transmitted drug resistance in the CFAR network of integrated clinical systems cohort: prevalence and effects on pre-therapy CD4 and viral load. *PLoS One* 6: e21189.
4. Hirsch MS, Günthard HF, Schapiro JM, Brun-Vézinet F, Clotet B, et al. (2008) Antiretroviral drug resistance testing in adult HIV-1 infection: 2008 recommendations of an International AIDS Society-USA panel. *Clin Infect Dis* 47: 266-85.
5. Zhang J, Rhee SY, Taylor J, Shafer RW (2005) Comparison of the precision and sensitivity of the Antivirogram and PhenoSense HIV drug susceptibility assays. *J Acquir Immune Defic Syndr* 38: 439-44.
6. Frentz D, Boucher CAB, Assel M, De Luca A, Fabbiani M, et al. (2010) Comparison of HIV-1 genotypic resistance test interpretation systems in predicting virological outcomes over time. *PLoS One* 5: e11505.
7. Jiamsakul A, Kantor R, Li PCK, Sirivichayakul S, Sirisanthana T, et al. (2012) Comparison of predicted susceptibility between genotype and virtual phenotype HIV drug resistance interpretation systems among treatment-naive HIV-infected patients in Asia: TASER-M cohort analysis. *BMC Res Notes* 5: 582.
8. Poon AFY, Kosakovsky Pond SL, Richman DD, Frost SDW (2007) Mapping protease inhibitor resistance to human immunodeficiency virus type 1 sequence polymorphisms within patients. *J Virol* 81: 13598-607.

9. Rhee SY, Taylor J, Fessel WJ, Kaufman D, Towner W, et al. (2010) HIV-1 protease mutations and protease inhibitor cross-resistance. *Antimicrob Agents Chemother* 54: 4253-61.
10. Melikian GL, Rhee SY, Taylor J, Fessel WJ, Kaufman D, et al. (2012) Standardized comparison of the relative impacts of HIV-1 reverse transcriptase (RT) mutations on nucleoside RT inhibitor susceptibility. *Antimicrob Agents Chemother* 56: 2305-13.
11. Garrido C, Villacian J, Zahonero N, Pattery T, Garcia F, et al. (2012) Broad phenotypic cross-resistance to elvitegravir in HIV-infected patients failing on raltegravir-containing regimens. *Antimicrob Agents Chemother* 56: 2873-8.
12. Van der Borgh K, Verheyen A, Feyaerts M, Van Wesenbeeck L, Verlinden Y, et al. (2013) Quantitative prediction of integrase inhibitor resistance from genotype through consensus linear regression modeling. *Virol J* 10: 8.
13. Zhang J, Hou T, Wang W, Liu JS (2010) Detecting and understanding combinatorial mutation patterns responsible for HIV drug resistance. *Proc Natl Acad Sci U S A* 107: 1321-6.
14. Yuan Y, Yokoyama M, Maeda Y, Terasawa H, Harada S, et al. (2013) Structure and dynamics of the gp120 V3 loop that confers noncompetitive resistance in R5 HIV-1(JR-FL) to maraviroc. *PLoS One* 8: e65115.
15. Prosperi MCF, Altmann A, Rosen-Zvi M, Aharoni E, Borgulya G, et al. (2009) Investigation of expert rule bases, logistic regression, and non-linear machine learning techniques for predicting response to antiretroviral treatment. *Antivir Ther* 14: 433-42.
16. Heider D, Senge R, Cheng W, Hüllermeier E (2013) Multilabel classification for exploiting cross-resistance information in HIV-1 drug resistance prediction. *Bioinformatics* 29: 1946-52.
17. Jessen LE, Hoof I, Lund O, Nielsen M (2013) SigniSite: Identification of residue-level genotype-phenotype correlations in protein multiple sequence alignments. *Nucleic Acids Res* 41: W286-91.
18. Moore JP, Trkola A, Dragic T (1997) Co-receptors for HIV-1 entry. *Curr Opin Immunol* 9: 551-62.
19. Vandekerkhove L, Verhofstede C, Vogelaers D (2009) Maraviroc: perspectives for use in antiretroviral-naive HIV-1-infected patients. *J Antimicrob Chemother* 63: 1087-96.

20. Philpott SM (2003) HIV-1 coreceptor usage, transmission, and disease progression. *Curr HIV Res* 1: 217-27.
21. de Jong JJ, Goudsmit J, Keulen W, Klaver B, Krone W, et al. (1992) Human immunodeficiency virus type 1 clones chimeric for the envelope V3 domain differ in syncytium formation and replication capacity. *J Virol* 66: 757-65.
22. Wilkin TJ, Goetz MB, Leduc R, Skowron G, Su Z, et al. (2011) Reanalysis of coreceptor tropism in HIV-1-infected adults using a phenotypic assay with enhanced sensitivity. *Clin Infect Dis* 52: 925-8.
23. Sánchez V, Masiá M, Robledano C, Padilla S, Ramos JM, et al. (2010) Performance of genotypic algorithms for predicting HIV-1 tropism measured against the enhanced-sensitivity Trofile coreceptor tropism assay. *J Clin Microbiol* 48: 4135-9.
24. Portsmouth S, Valluri SR, Däumer M, Thiele B, Valdez H, et al. (2013) Correlation between genotypic (V3 population sequencing) and phenotypic (Trofile ES) methods of characterizing coreceptor usage of HIV-1 from 200 treatment-naïve HIV patients screened for Study A4001078. *Antiviral Res* 97: 60-5.
25. Fouchier RA, Groenink M, Kootstra NA, Tersmette M, Huisman HG, et al. (1992) Phenotype-associated sequence variation in the third variable domain of the human immunodeficiency virus type 1 gp120 molecule. *J Virol* 66: 3183-7.
26. Pillai S, Good B, Richman D, Corbeil J (2003) A new perspective on V3 phenotype prediction. *AIDS Res Hum Retroviruses* 19: 145-9.
27. Dybowski JN, Heider D, Hoffmann D (2010) Prediction of co-receptor usage of HIV-1 from genotype. *PLoS Comput Biol* 6: e1000743.
28. Jensen MA, Li FS, van 't Wout AB, Nickle DC, Shriner D, et al. (2003) Improved coreceptor usage prediction and genotypic monitoring of R5-to-X4 transition by motif analysis of human immunodeficiency virus type 1 env V3 loop sequences. *J Virol* 77: 13376-88.

29. Resch W, Hoffman N, Swanstrom R (2001) Improved success of phenotype prediction of the human immunodeficiency virus type 1 from envelope variable loop 3 sequence using neural networks. *Virology* 288: 51-62.
30. Díez-Fuertes F, Delgado E, Vega Y, Fernández-García A, Cuevas MT, et al. (2013) Improvement of HIV-1 coreceptor tropism prediction by employing selected nucleotide positions of the env gene in a Bayesian network classifier. *J Antimicrob Chemother* 68: 1471-85.
31. Kumar R, Raghava GPS (2013) Hybrid approach for predicting coreceptor used by HIV-1 from its V3 loop amino acid sequence. *PLoS One* 8: e61437.
32. Sander O, Sing T, Sommer I, Low AJ, Cheung PK, et al. (2007) Structural descriptors of gp120 V3 loop for the prediction of HIV-1 coreceptor usage. *PLoS Comput Biol* 3: e58.
33. Low AJ, Dong W, Chan D, Sing T, Swanstrom R, et al. (2007) Current V3 genotyping algorithms are inadequate for predicting X4 co-receptor usage in clinical isolates. *AIDS* 21: F17-24.
34. Pillai SK, Pond SLK, Liu Y, Good BM, Strain MC, et al. (2006) Genetic attributes of cerebrospinal fluid-derived HIV-1 env. *Brain* 129: 1872-83.
35. Holman AG, Gabuzda D (2012) A machine learning approach for identifying amino acid signatures in the HIV env gene predictive of dementia. *PLoS One* 7: e49538.
36. Gnanakaran S, Bhattacharya T, Daniels M, Keele BF, Hraber PT, et al. (2011) Recurrent signature patterns in HIV-1 B clade envelope glycoproteins associated with either early or chronic infections. *PLoS Pathog* 7: e1002209.
37. Burton DR, Weiss RA (2010) A boost for HIV vaccine design. *Science* 329: 770-3.
38. Walker LM, Phogat SK, Chan-Hui PY, Wagner D, Phung P, et al. (2009) Broad and potent neutralizing antibodies from an African donor reveal a new HIV-1 vaccine target. *Science* 326: 285-9.
39. Zhou T, Georgiev I, Wu X, Yang ZY, Dai K, et al. (2010) Structural basis for broad and potent neutralization of HIV-1 by antibody VRC01. *Science* 329: 811-7.

40. Walker LM, Huber M, Doores KJ, Falkowska E, Pejchal R, et al. (2011) Broad neutralization coverage of HIV by multiple highly potent antibodies. *Nature* 477: 466-70.
41. Burton DR, Ahmed R, Barouch DH, Butera ST, Crotty S, et al. (2012) A Blueprint for HIV Vaccine Discovery. *Cell Host Microbe* 12: 396-407.
42. Pancera M, Shahzad-Ul-Hussan S, Doria-Rose NA, McLellan JS, Bailer RT, et al. (2013) Structural basis for diverse N-glycan recognition by HIV-1-neutralizing V1-V2-directed antibody PG16. *Nat Struct Mol Biol* 20: 804-13.
43. O'Rourke SM, Schweighardt B, Phung P, Fonseca DPAJ, Terry K, et al. (2010) Mutation at a single position in the V2 domain of the HIV-1 envelope protein confers neutralization sensitivity to a highly neutralization-resistant virus. *J Virol* 84: 11200-9.
44. Gnanakaran S, Daniels MG, Bhattacharya T, Lapedes AS, Sethi A, et al. (2010) Genetic signatures in the envelope glycoproteins of HIV-1 that associate with broadly neutralizing antibodies. *PLoS Comput Biol* 6: e1000955.
45. West AP Jr, Scharf L, Horwitz J, Klein F, Nussenzweig MC, et al. (2013) Computational analysis of anti-HIV-1 antibody neutralization panel data to identify potential functional epitope residues. *Proc Natl Acad Sci U S A* 110: 10598-603.
46. Chuang GY, Acharya P, Schmidt SD, Yang Y, Louder MK, et al. (2013) Residue-level prediction of HIV-1 antibody epitopes based on neutralization of diverse viral strains. *J Virol* 87: 10047-58.
47. Stanfel LE (1996) A New Approach to Clustering the Amino Acid. *Journal of Theoretical Biology* 183: 195 - 205.
48. Peng H, Long F, Ding C (2005) Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 27: 1226-1238.
49. Ravich VL, Masso M, Vaisman II (2011) A combined sequence-structure approach for predicting resistance to the non-nucleoside HIV-1 reverse transcriptase inhibitor Nevirapine. *Biophys Chem* 153: 168-72.

50. Avila-Ríos S, García-Morales C, Garrido-Rodríguez D, Ormsby CE, Hernández-Juan R, et al. (2011) National prevalence and trends of HIV transmitted drug resistance in Mexico. *PLoS One* 6: e27812.
51. Cohen J (1960) A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* 20: 37–46.
52. Liu L, May S, Richman DD, Hecht FM, Markowitz M, et al. (2008) Comparison of algorithms that interpret genotypic HIV-1 drug resistance to determine the prevalence of transmitted drug resistance. *AIDS* 22: 835-9.
53. Cormier EG, Tran DN, Yukhayeva L, Olson WC, Dragic T (2001) Mapping the determinants of the CCR5 amino-terminal sulfopeptide interaction with soluble human immunodeficiency virus type 1 gp120-CD4 complexes. *J Virol* 75: 5541-9.
54. Holman AG, Mefford ME, O'Connor N, Gabuzda D (2010) HIVBrainSeqDB: a database of annotated HIV envelope sequences from brain and other anatomical sites. *AIDS Res Ther* 7: 43.
55. Burton DR, Pyati J, Koduri R, Sharp SJ, Thornton GB, et al. (1994) Efficient neutralization of primary isolates of HIV-1 by a recombinant human monoclonal antibody. *Science* 266: 1024-7.
56. Kwong PD, Mascola JR, Nabel GJ (2013) Broadly neutralizing antibodies and the search for an HIV-1 vaccine: the end of the beginning. *Nat Rev Immunol* 13: 693-701.
57. Huang Cc, Tang M, Zhang MY, Majeed S, Montabana E, et al. (2005) Structure of a V3-containing HIV-1 gp120 core. *Science* 310: 1025-8.
58. McLellan JS, Pancera M, Carrico C, Gorman J, Julien JP, et al. (2011) Structure of HIV-1 gp120 V1/V2 domain with broadly neutralizing antibody PG9. *Nature* 480: 336-43.
59. Julien JP, Sok D, Khayat R, Lee JH, Doores KJ, et al. (2013) Broadly neutralizing antibody PGT121 allosterically modulates CD4 binding via recognition of the HIV-1 gp120 V3 base and multiple surrounding glycans. *PLoS Pathog* 9: e1003342.
60. Huang J, Ofek G, Laub L, Louder MK, Doria-Rose NA, et al. (2012) Broad and potent neutralization of HIV-1 by a gp41-specific human antibody. *Nature* 491: 406-12.

61. Scheid JF, Mouquet H, Ueberheide B, Diskin R, Klein F, et al. (2011) Sequence and structural convergence of broad and potent HIV antibodies that mimic CD4 binding. *Science* 333: 1633-7.
62. Pond SLK, Frost SDW, Muse SV (2005) HyPhy: hypothesis testing using phylogenies. *Bioinformatics* 21: 676-9.
63. Nickle DC, Heath L, Jensen MA, Gilbert PB, Mullins JI, et al. (2007) HIV-specific probabilistic models of protein evolution. *PLoS One* 2: e503.
64. Pond SK, Muse SV (2005) Site-to-site variation of synonymous substitution rates. *Mol Biol Evol* 22: 2375-85.
65. Ogert RA, Lee MK, Ross W, Buckler-White A, Martin MA, et al. (2001) N-linked glycosylation sites adjacent to and within the V1/V2 and the V3 loops of dualtropic human immunodeficiency virus type 1 isolate DH12 gp120 affect coreceptor usage and cellular tropism. *J Virol* 75: 5998-6006.

Figure Legends

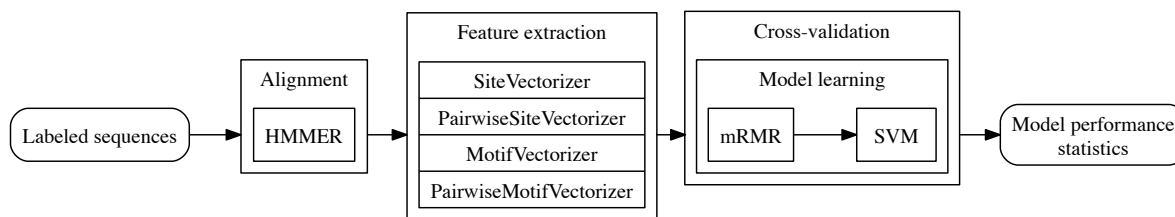


Figure 1. IDEPI workflow

Tables

Simulation	L	M	Median performance metrics				Mean epitope recovery, by class			
			Sensitivity	Specificity	MCC	Features	Slow, %	Intermediate, %	Fast, %	FP
Simple	5	≥ 1	0.98	1.0	0.98	2	11.1	56.6	80.0	0.09
Intermediate	8	≥ 2	0.95	1.0	0.94	3	10.4	42.6	71.6	0.16
Complex	10	≥ 3	0.85	0.98	0.78	3	6.0	39.4	58.3	0.16
Random	N/A	N/A	0.57	0.47	0.04	1	N/A	N/A	N/A	1

Table 1. IDEPI performance on simulated data, using forward feature selection (to optimize MCC), and 10-fold nested cross-validation. **L**: the number of sites in an epitope; **M**: how many escape mutations are needed to confer resistance; epitope recover classes are based on simulated evolutionary rates; **FP**: mean number of selected features not in a simulated epitope; a feature was counted as recovered if it were selected in 50% or more of cross-validation replicates.

Problem	N	B	F	IDEPI performance				Benchmark (IDEPI : ref)
				5-fold cross-validation metrics				
				Sens.	Spec.	Accu.	MCC	
NVP resistance	1461	62.3%	80	0.88	0.97	0.92	0.83	CV Accu. 0.92 : 0.92 ¹
V3 tropism	1356	15.1%	90	0.89	0.94	0.94	0.78	Training Accu. 0.95:0.96 ²
Dementia	861	70.3%	90	0.96	0.93	0.95	0.89	CV Accu. 0.95:0.75 ³
b12 bNab	247	64.4%	5	0.74	0.62	0.70	0.36	Training Accu. 0.75:0.86 ⁴
10E8 bNab	178	4.0%	5	0.30	0.96	0.93	0.23	Training Accu. 0.96 vs proportion of residuals explained 0.21 ⁵
PG9 bNab	174	24.1%	30	0.49	0.89	0.79	0.42	Training Accu. 0.96 vs proportion of residuals explained 0.31 ⁵
PGT-121 bNab	118	37.2%	1	0.80	0.79	0.80	0.58	Training Accu. 0.80 vs proportion of residuals explained 0.52 ⁵
8ANC131 bNab	178	30.9%	15	0.51	0.69	0.63	0.19	
8ANC195 bNab	178	42.7%	2	0.94	0.75	0.83	0.67	Training Accu. 0.83 vs proportion of residuals explained 0.58 ⁵

Table 2. IDEPI performance in predicting phenotypes from genotypes based on training data analyzed previously. IDEPI metrics were obtained using 5-fold cross-validation. B (balance) is defined as the proportion of “positive” training samples. The number of features (F) was chosen by selecting a value from a pre-defined grid which maximized cross-validation MCC.

¹ random forests trained on combined sequence and structural features using resistance classifications from the Stanford Drug Resistance Database [49];

² a two-level classifier combining random forest predictions based on an electrostatic hull and hydrophobicity features of the V3 loop (680 features) trained on the same data [27];

³ a hierarchical decision tree classifier using composite amino-acid features trained on the same data [35].

⁴ an ensemble classifier using signature rules and logistic regression trained on the same data [44].

Problem	Features selected by IDEPI				
	Rank	Identity	Direction	MCC	Remarks
NVP resistance	1	K103K	Susceptible	0.46	Canonical NNRTI resistance site [1]
	2	Y181Y	Susceptible	0.65	Canonical NNRTI resistance site
	3	G190G	Susceptible	0.74	Canonical NNRTI resistance site
V3 tropism	1	PNGS(N301)	CCR5	0.55	Essential for CCR5 binding [53] and dual-tropic viruses [65]
	2	R306R	CCR5	0.67	Part of the 11/25 rule [25]
Dementia	1	T297K	Non-HAD	0.57	
	2	PNGS (N488)	HAD		
	2	R298D	Non-HAD		
	3	I320[]	non-HAD		
b12 bNab	4	PNGS(T188)	HAD	0.71	
	1	D185D	Susceptible	0.26	The strongest association found in [44]
8E10 bNab	3-4	T676T	Susceptible	N/A	A part of the structural epitope [60]
PG9 bNab	1	PNGS (N160)	Susceptible	0.37	Key residue for binding based on structure [58]
	8	V169E	Resistant		Forms a hydrogen bond with PG9 [58]
PGT-121 bNab	1	PNGS(N301+N332)	Susceptible	0.58	Both glycans are important for neutralization [59]
8ANC195 bNab	1	PNGS (N234+N276)	Susceptible	0.59	Encompasses the three mutants (sites 234, 236, and 276) any of which confers resistance [45, 46]
	2	PNGS(N160+N230)	Resistant	0.67	PNGS at site 230 confers weak resistance [45]
8ANC131 bNab	3.75	PNGS(N339+Q442)	Resistant		
	5	K151G	Susceptible		

Table 3. Key features selected by IDEPI for each of the example problems. Notation: T297K means that K is found in position 297 (HBX2 coordinates, T is the residue found in HXB2); PNGS (T188) – a potential N-linked glycosylation site with N at HXB2 coordinate 188; PNGS (N234+N276) – a pair of potential N-linked glycosylation site with N at HXB2 coordinates N234 and N276; [] – a deletion relative to HXB2.

Problem	Independent dataset			Performance
	N	Reference	Benchmark	
NVP resistance	1639	[50]	Stanford HIVdb	Cohen’s $\kappa = 0.85$.
V3 tropism	74	[27]	Best of 5 methods, including SVM, decision trees, and position-specific scoring matrices [27]	Accu. IDEPI 0.91 vs 0.86
Dementia	10	[35]	Ensemble of rule learning and decision trees from [35]	IDEPI 10/10 vs 8/10
b12 bNab	55	[44]	Ensemble of signatures and logistic regression [44]	Accu. IDEPI 0.73 vs 0.61

Table 4. IDEPI model performance on independent datasets and comparison with benchmark methods.

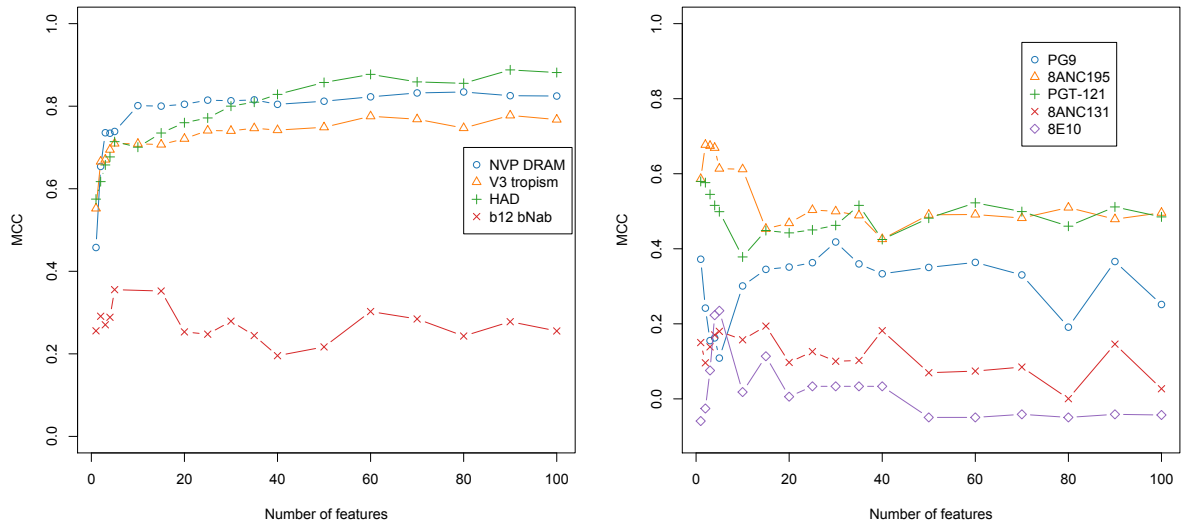


Figure 2. IDEPI performance, measured by MCC, as a function of the number of model features