TLmutation: predicting the effects of mutations using transfer learning.

Zahra Shamsi,[†] Matthew Chan,[†] and Diwakar Shukla[*,†,‡,¶]

†*Department of Chemical and Biomolecular Engineering, University of Illinois at Urbana-Champaign, Urbana, IL, 61801*

‡*Center for Biophysics and Quantitative Biology, University of Illinois at Urbana-Champaign, Urbana, IL, 61801*

¶*National Center for Supercomputing Applications, University of Illinois, Urbana, IL, 61801*

E-mail: diwakar@illinois.edu

Abstract

A reoccurring challenge in bioinformatics is predicting the phenotypic consequence of amino acid variation in proteins. Due to recent advances in sequencing techniques, sufficient genomic data is becoming available to train models that predict the evolutionary statistical energies for each sequence, but there is still inadequate experimental data to directly predict functional effects. One approach to overcome this data scarcity is to apply transfer learning and train more models with available datasets. In this study, we propose a set of transfer learning algorithms, we call TLmutation, which implements a supervised transfer learning algorithm that transfers knowledge from survival data to a protein function of interest in the same protein followed by an unsupervised transfer learning algorithm that extends the knowledge to a homologous protein. We explore the application of our algorithms in three cases. First, we test the supervised transfer on dozens of previously published mutagenesis datasets to complete and refine

missing datapoints. We further investigate these datasets to identify which variants build better predictors of variant functions. In the second case, we apply the algorithm to predict higher-order mutations solely from single point mutagenesis data. Finally, we perform the unsupervised transfer learning algorithm to predict mutational effects of homologous proteins from experimental datasets. These algorithms are generalized to transfer knowledge between Markov random field models. We show the benefit of our transfer learning algorithms to utilize informative deep mutational data and provide new insights into protein variant functions. As these algorithms are generalized to transfer knowledge between Markov random field models, we expect these algorithms to be applicable to other disciplines.

Introduction

Proteins are intricate molecular machines that regulate all biological processes. The function of a particular protein is intrinsically linked to its structure, which governs its stability and conformational dynamics. Consequently, mutations in protein sequences, in which one amino acid is replaced by another amino acid, can affect a protein's structure, stability, and inevitably its function. While some mutations may have little to no effect on a protein's function, others have larger implications for disease and antibiotic resistance.[1,2] Recent advancements in large-scale genomic sequencing have provided tools and resources for both consumers and clinicians to identify disease-potential mutations in one's proteome at an affordable cost. However, due this large influx of genomic data, a recurring challenge is predicting the phenotypic consequence in proteins due to amino acid variations.[3,4]

Several workflows, both experimentally and computationally, have been developed to identify, predict, and model the effects of mutations.[5–8] Engineering approaches, such as deep mutational scanning, provide a unique glimpse into the sequence-function relationship of proteins by surveying all mutations in the sequence and assessing its altered function.[9,10] These methods provide large quantitative datasets of mutational effects for a particular

protein. Alternatively, statistical models have been used as standalone approaches or to compliment biophysical experiments. PolyPhen2[11] and SIFT[12] are examples of common frameworks that use multiple sequence and structural-based alignments to characterize variants. Other models, such as SNAP2,[13] CADD,[14] and Envision,[15] employ machine learning algorithms to classify and predict mutations and are popular due to their robustness with large datasets. Of these methods, Envision is a prime example of utilizing experimental deep mutational scanning datasets and implements a gradient boosting learning algorithm to predict mutations of various proteins.[15]

One successful approach employed for predicting mutational effects is EVmutation which uses evolutionary sequence conservation.[16] In addition to applying evolutionary conservation to predict the effect of mutations, EVmutation also considers genetic interactions between mutations and the sequence background. By accounting for the interactions between all residue pairs, the model predicts the effects of mutations more accurately as compared to other predictors.[16] EVmutation utilizes a graph based Markov random field known as the Potts model which is trained on natural sequences.[16] This means, for a given sequence, the algorithm searches through the UniProt database,[17] locates all natural sequences in its family, and uses these sequences as data to train the Potts model. However, these unsupervised probabilistic models do not directly predict the effects of mutations on the functionality of the protein; rather they predict if the mutant species is fit to survive, which is not always directly correlated to a specific function.[16]

As genomic sequencing and mutational libraries become readily available, it is an opportunity to utilize this data-rich regime to enhance predictors of protein mutations. However, training Potts models on specific experimental data is usually not feasible. While mutational libraries of various proteins have been completed due to advancements in deep mutational scanning and directed evolution methods,[18–21] these datasets are inadequate for training. As the datasets are limited, we must rely on alternative methods to obtain insights and predictions. One approach to overcome this data scarcity is to apply transfer learning al-

gorithms in which we apply knowledge from one task to a different, yet related task. In a traditional machine learning approach, different task will be individually learned to build a model. However, in many real-world applications, collecting training data and rebuilding models may be computationally expensive.[22] These difficulties are akin to deep mutational scanning and other biophysical approaches. Many experimental methods in characterizing protein variants are susceptible to noise or missing datapoints.[7] Moreover, it is difficult to infer informational about other proteins from a single deep mutational scan.

Here, we propose an algorithm, TLmutation, which is an adaptation of the successful variant effect predictor EVmutation. We implement the algorithm in two fashions. First, TLmutation transfers knowledge from a model, trained on natural sequences, to a new protein function for the same protein. We call this algorithm supervised TLmutation. This is followed by an unsupervised transfer learning algorithm that expands the knowledge to a related protein, and referred to as unsupervised TLmutation. It is worth mentioning that EVmutation is shown to be able to capture the function related conformational dynamics of proteins.[23] We conducted multiple experiments using the proposed transfer learning algorithms to evaluate the practical efficiency in predicting the effects of mutations in different proteins, with different types of training and test datasets. In the first case, we explore the application of the proposed algorithm on 11 previously published mutagenesis datasets to complete missing datapoints. We further investigate different sampling approaches to delineate which variants provide more accurate predictors of variant functions. In the second case, we apply our algorithm to predict higher order mutations (i.e. double mutations, triple mutations) solely from single point mutatgensis data. Finally, we implement the unsupervised transfer learning algorithm to predict mutational effects of homologous proteins from experimental datasets.

Methods

Potts model for protein sequences

Potts models are a specific type of Markov random field (MRF) models. MFR models have widely been used to tackle large datasets in different disciplines such as genomic biology,[24] physics,[25] natural language processing,[26] and computer vision.[27] The proposed supervised and unsupervised transfer algorithms can be generalized and applied to other MRF models. A MRF is an undirected, probabilistic graphical model that represents statistical dependencies among a set of random variables, $\sigma = (\sigma_1, ..., \sigma_N)$, where $\forall\ \sigma_i \in \{1, 2, .., l\}$. In this study, let $\sigma = (\sigma_1, \sigma_2, ..., \sigma_N)$ represent the amino acid sequence of a protein with length N. Each σ_i takes on values in $\{1, 2, ..., 21\}$ (one state for each of the 20 naturally occurring amino acids and one additional state to represent a gap). The probability of $\sigma_1, ..., \sigma_N$ is then given by:

$$P(\sigma_1, ..., \sigma_N) = \frac{1}{Z} \exp(\sum_{i=1}^{N} h_i(\sigma_i) + \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} J_{ij}(\sigma_i, \sigma_j)) \tag{1}$$

where h_i is the potentials for site i (fields), J_{ij} is potentials between residue pair constraints (couplings) of i and j,[16] and Z is the partition function.[28]

Assume we have two similar domains, source and target domain, in which we have a base and a new task (Figure 1). The new task has to be a subset of the base task. This means the base task has to the result of multiple smaller tasks, where the new task is one of them. We are given two MRF models for the source domain and the target domain, both trained on the base task (M_s^{base} and M_t^{base} represent the MRF models, where their domain is shown as a subscript and task's type as a superscript). Our aim is to find MRF models on the new task for both of these domains (M_s^{new} and M_t^{new}). A set of n labeled training data points for the new task ($X_s^{new,tr}, Y_s^{new,tr}$) is available only for the source domain. Each $X_{s,i}^{new,tr} \in \Re^{d_x}$ and

$Y_{s,i}^{new,tr} \in \Re^{d_x}$. We have sets of test data for both the source and the target domain ($X_s^{new,te}$ and $X_t^{new,te}$, with corresponding labels as $Y_s^{new,te}$ and $Y_t^{new,te}$). The first goal is to utilize the given training data and the MRF model to find a predictive model for the new task in the source domain (M_s^{new}). Then, we extend the knowledge from this supervised transfer learning step to learn a model in the target domain (M_t^{new}) using an unsupervised transfer learning algorithm. The predictions ($\hat{Y}_i$) were ranked and compared to the actual labels (Y_i) for calculation of the Spearman correlation coefficient (ρ) to assess the performance of the algorithm.[29]

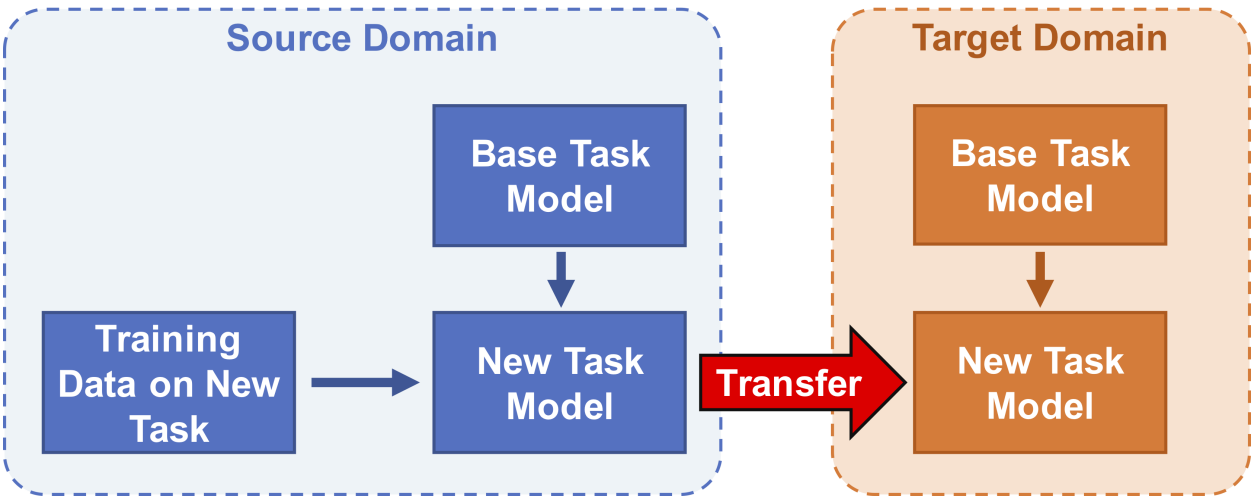

Figure 1: Proposed transfer learning algorithms for MFR models. Knowledge is transferred from the source domain where training data is available to the target domain.

Supervised transfer from the evolutionary statistical energy to a functional assay.

We want to modify an existing predictive model that is trained on the base task to be able to predict a new task. In the supervised transfer section, all the transfer is within a same domain (source domain), therefore the subscripts that determine the domains are dropped. First, we train a MRF model on the base task (M^{base}), and calculate the values for the potentials of θ_c^{base} using available methods in the literature.[16] Then, we introduce a new

MRF model that can predict the new task as following:

$$P(\sigma) = \frac{1}{Z} \exp(\sum_{c \in C} \theta_c^{new}) \tag{2}$$

where

$$\theta_c^{new} = w_c . \theta_c^{base} \tag{3}$$

w is a binary weight matrix ($w_c \in \{0, 1\}$), which is calculated by maximizing the correlation between the predicted values of labels and the actual labels ($\hat{Y}^{new,tr}$ and $Y^{new,tr}$) as

$$\max_{w} \mid \rho(rg_{\hat{Y}^{new,tr}}, rg_{Y^{new,tr}}) \mid \tag{4}$$

with $rg_{\hat{Y}^{new,tr}}$ and $rg_{Y^{new,tr}}$ are ranks of $\hat{Y}^{new,tr}$ and $Y^{new,tr}$, respectively. ρ is the Spearman's rank correlation coefficient.

In this particular study, we want to learn a Potts model that can predict the effect of mutations on a particular protein function. We are given a Potts model trained on natural sequences (EVmutation model) and a limited experimental data of the effects of point mutations on the function. In Potts models, the physical meaning of each J_{ij} parameter is the chemical/physical interaction between the corresponding residues. Since the model is trained on survival data, the J_{ij}'s with large values are coding for critical interactions for survival. A critical interaction for survival may have roles in expression level, folding, thermal stability, or other vital functions. Assuming our function of interest is one of these functions, we want to decouple J_{ij}'s and keep the ones in the function of our interest. This forms the basis of the proposed supervised transfer algorithm (supervised TLmutation). We decouple these effects by nullifying the J_{ij}'s that do not contribute to the function.

In the supervised transfer learning section, all the transfer occurs within the same protein (referred to as source protein). First, we train a Potts model on survival (e.g. EVmutation), and calculate the values for the potentials of J_{ij} using available methods in the literature.[16] Then, we introduce a new Potts model that can predict the function using the following

modified potentials:

$$J_{ij}^{func.} = w_{ij}.J_{ij} \tag{5}$$

$$h_i^{func.} = w'_i.h_i \tag{6}$$

where w and w' are binary weight matrices ($w_c \in \{0, 1\}$), which are calculated by maximizing the correlation between the predicted values of the experiments and the actual experimental values in the training set ($\hat{Y}^{tr}$ and Y^{tr}) as

$$\max_{w,w'} \mid \rho(rg_{\hat{Y}^{tr}}, rg_{Y^{tr}}) \mid \tag{7}$$

and $rg_{\hat{Y}^{tr}}$ and $rg_{Y^{tr}}$ are ranks of $\hat{Y}^{tr}$ and Y^{tr}, respectively. Again, ρ is the Spearman's rank correlation coefficient. Here, we eliminate potentials which do not contribute to the enhancement in predicting the $Y^{new,tr}$. This type of transfer is suitable when the base task is an outcome of a series of tasks and the new task is one of these task. In this way, we eliminate the effects of other tasks and focus on the new task. We call this model a TLmutation model. The maximization algorithm is explained in the Supporting Information.

Unsupervised transfer between proteins.

Now, we want to go one step further and expand the knowledge gained from the supervised transfer to the target domain, where the training data is not available for the new task. Therefore, we have a MRF trained on the base task in the target domain (M_t^{base}) and a MRF trained on the new task in the source domain . In our generalization example, using the learned model parameter, w_c, from Equation 3, we define a new MRF model for the new task in the target domain. The new model's potential $\theta_{c,t}^{new}$ is calculated using Equation 8.

$$\theta_{c,t}^{new} = w_c.\theta_{c,t}^{base} \tag{8}$$

Since the source and target domains are similar, we assume the same corresponding potentials have effects on predicting the new task. Therefore, we use the same learned w to switch off the potentials in M_t^{base} and get the M_t^{new}. However, we use the same value of $\theta_{c,t}^{base}$ for potentials that remain active in the MRF. Here, we extended the supervised and unsupervised TLmutation algorithms to a more generic MRF model. The algorithm transfers knowledge from a well trained MRF model of one task, to other similar tasks when either limited or no training data is available.

For most proteins, we do not have sufficient training data to use a supervised transfer learning algorithm as obtaining mutation data is experimentally challenging and expensive.[30,31] We want to use the available experimental data of a protein for predicting the mutations effects in other homologous proteins. We expand the knowledge gained from the supervised TLmutation to the target protein, where no training data is available.

Assume EVmutation model for target protein and TLmutation model for source protein are available. Using the parameters of the TLmutation, w and w', from Equation 5, we define a new Potts model for predicting the effect of mutation in the target protein on its function. The new model's potentials are calculated using Equation 9 and 10.

$$J_{ij}^{func.,target} = w_{ij}.J_{ij}^{target} \tag{9}$$

$$h_i^{func.,target} = w'_i.h_i^{target} \tag{10}$$

where w_{ij} and w'_i are the binary masks from TLmutation model of the source protein. J_{ij}^{target} and h_i^{target} are the potentials from EVmodel of the target protein, and $J_{ij}^{func.,target}$ and $h_i^{func.,target}$ are the modified potentials for TLmodel of the target protein. Here, the source and target proteins should be homologs. Between homologous proteins, molecular mechanisms leading to a particular function are expected to be conserved. The residue-residue interactions participating in the function of the source protein also are anticipated to

be conserved in the target protein. Therefore in the unsupervised transfer, we eliminate the same J_{ij}'s in the target protein that did not contribute to the function in the source protein. However, we use the target's EVmodel as the basis and apply the binary masks from source (w and w') on its potentials.

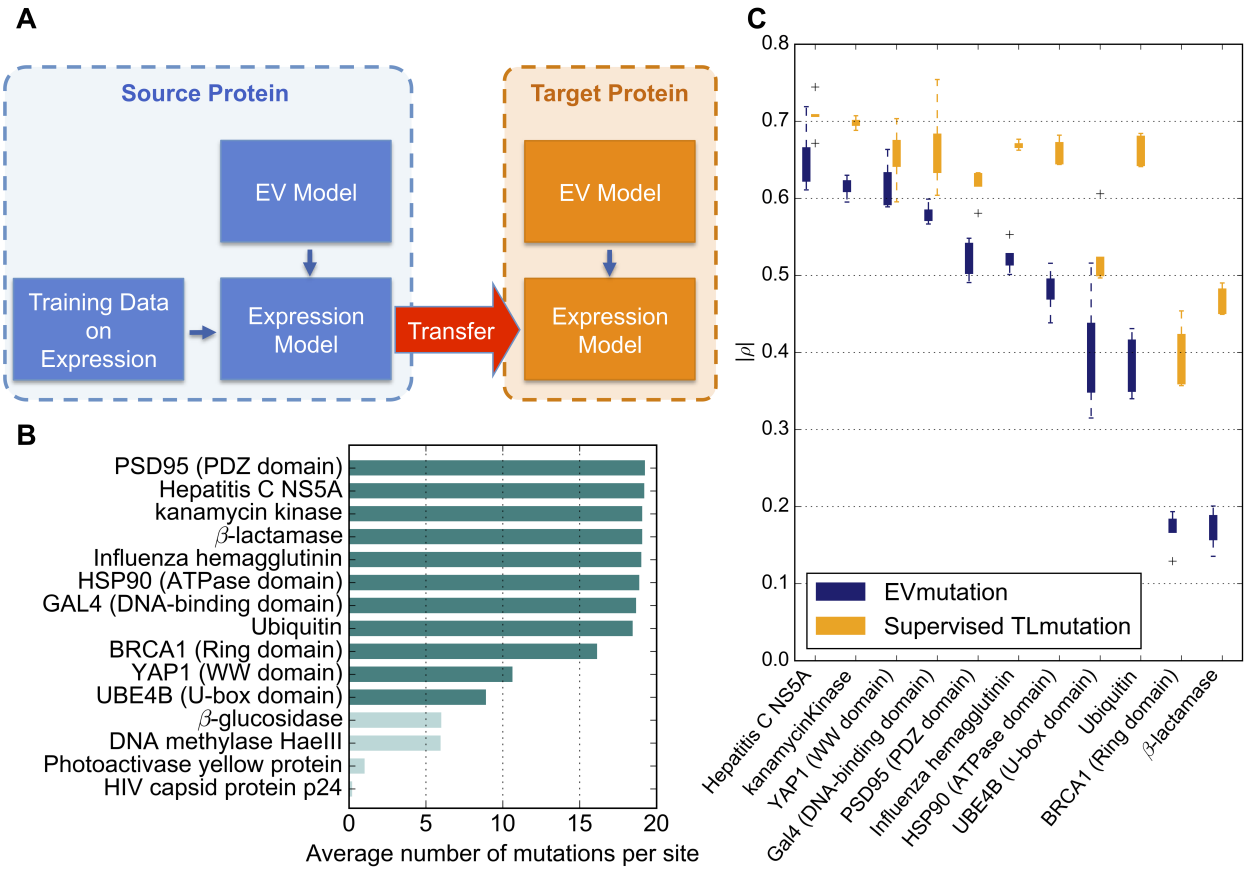

Figure 2: (**A**) schematic of the proposed transfer learning algorithms. Knowledge is transferred from the source protein where training data is available to the target protein. (**B**) Completeness of 15 studied mutagenesis datasets is shown. Datasets shown in light teal have less than 35% of all possible variants in the mutagenized region. (**C**) Computed effects of mutations agreement with experimental measurements of molecular function for 11 proteins are shown for the test set based on the EVmutation and supervised TLmutation. The agreement is measured by Spearman's rank correlation coefficient ρ. The error bars are calculated from 5-fold cross-validation.

Results

Case study 1: Completing gaps in experimental datasets

In this first case study, we want to evaluate the performance of TLmutation in refining and completing gaps in mutagenesis datasets. It is common to have incomplete datasets, where readings of certain variants were unobtainable due to experiemntal difficulties (e.g. expression or purification of protein, poor sequencing). Here, we apply the TLmutation algorithm to predict missing variants of 15 previously published, large-scale mutagensises datasets (see SI Table S1 for more details). These datasets have quantitative measurements of variant effects for various protein functions. For each dataset, the available data were randomly divided into test and training sets. 5-fold cross-validation was used to access the robustness of the supervised TLmutation models. Our initial analysis showed TLmutation does not significantly improve EVmutation models for incomplete datasets which have less than 6 mutations per site of the mutagenized region. This leads us to enforce an extra constraint on the datasets to include experimental values for at least $\sim 35\%$ of all possible variants of the mutagenized region (Figure S1). As shown in Figure 2 **B**, 11 out of 15 datasets had adequate number of experimental points. In these datasets, supervised TLmutation improved the correlation coefficient with the actual experimental values for the test set in all 11 proteins (Figure 2 **C**). Among these, the largest improvements are observed in systems with a lower correlation of EVmutation with actual experiments.

Which variants should be experimentally tested to build better predictors of variant function?

As there are more than hundreds of possible sites of a protein that can be subjected to mutagensis, it is beneficial to understand which datapoints may achieve the most generalizable information about other variants. Here, we address this question by training the TLmutation model using different sampling methods.

Simple random sampling is the most common method as it is efficient and relatively easy to implement. In this approach, samples are randomly selected with a uniform distribution. This sampling method was used in the previous section and showed a significant improvement in the performance of the model in all datasets (Figure 2 **C**, training scores are shown in Figure S2). However, mutagenesis datasets inherently are not uniform. For example, a variant with a mutation in site "A" may contain more information about other variants with mutations in the same site, as compared to variants with mutations in site "B", far from "A". More sophisticated sampling methods will be more suitable for these types of naturally ordered datasets. Here, we tested two systematic sampling approaches on the same 11 protein datasets. In the first approach, we divided the datasets based on sequence sites. The TLmutation model was trained on all available mutations for 80% of the sequence sites, and tested on the remaining data (as shown in Figure 3 **A**). As before, 5-fold cross-validation was used to assess the sampling method. This approach dramatically decreased the performance of the model. In most of the systems, no improvement was observed as compared to EVmutation (Figure 3 **B**). This observation suggests that the effects of mutations in sites far away from each other may not correlate with the mutations in other sites. This leads us to the second sampling approach, where the test/training splitting occurred for each site, meaning that 80% of available experiments for mutations on each sequence site was labeled as training, and 20% as test (as shown in Figure 3 **C**). Using this sampling, TLmutation outperformed EVmutation in all systems (Figure 3 **D**). However, the improvement is still comparable with random sampling.

Case study 2: Predicting the effects of multiple point mutations from single point mutation experiments.

Another limitation of mutagenesis datasets, is the lack of higher order mutation data. While a single point mutation may not affect the protein's function, it is possible for multiple mutations to cooperatively alter its function. Datasets of single point mutations have become

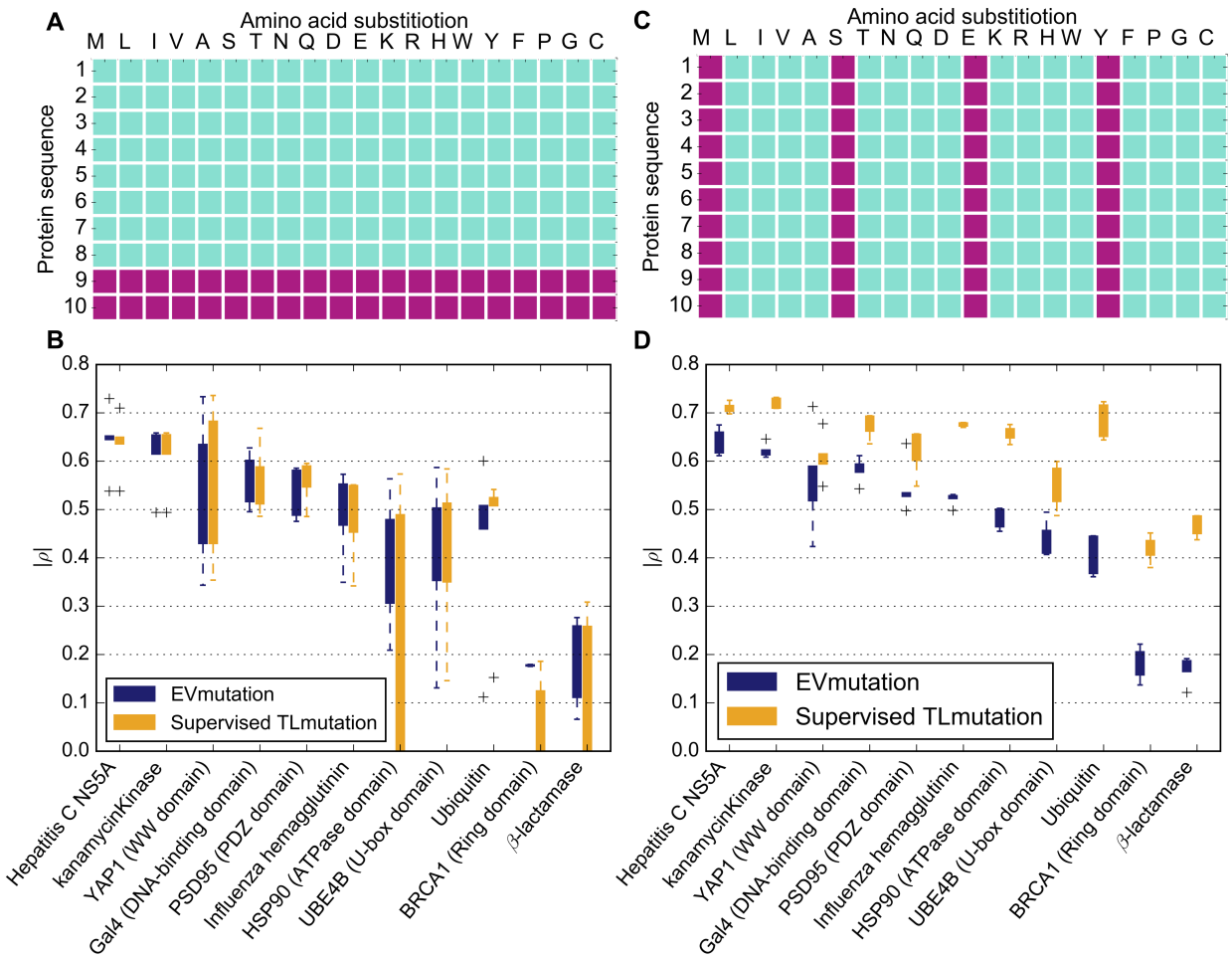

Figure 3: (**A**) A demonstration of train/test split based on sequence sites is shown. Purple points are considered in the test set and teal points in the training set. (**B**) The correlation between computed effects of mutations and experimental measurements for 11 proteins are shown for 5-fold cross validation and site based train/test split. TLmutation using site based train/test split, does not significantly improve the correlation coefficients. (**C**) A demonstration of train/test split based on substituted amino acid is shown. (**D**) The correlation between computed effects of mutations and experimental measurements for 11 proteins are shown for 5-fold cross validation and substituted amino acid based train/test split. TLmutation improves the correlation coefficients for all datasets.

increasing available over the past decade. However, our knowledge of multiple point mutations remains unclear. From an experimental perspective, the number of possible mutants increase exponentially with the increase in number of mutated residues, thus conducting a thorough mutagenesis analysis of large proteins is more difficult and expensive. Here, we use the supervised TLmutation algorithm to train on the available singe point mutations and predict the effects of multiple point mutations. We tested the performance of the algorithm on three previously published mutagenesis datasets. These datasets have been studied in the literature to evaluate EVmutation as well.[16] They contain single and double point mutations (more detail is provided in SI table S2). In these systems, the correlation coefficient was increased for both test training set as compared to EVmutation as shown in Figure 4.

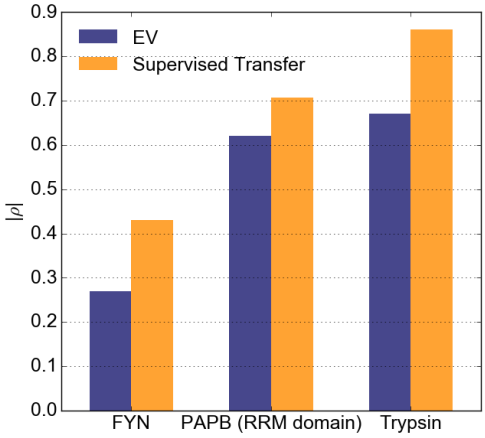

Figure 4: The agreement between predicted effects of double point mutations with experimental measurements are shown for the test sets. The supervised TLmodels were trained on limited available single point mutation data.

Case study 3: Predicting the effects of mutations using available experimental data on a homologous protein.

Mutagensis datasets provide an opportunity to utilize this data-rich regime and investigate the transferability of mutational effects among homologous proteins. The effectiveness of our transfer algorithms was tested for two chemokine receptors, CXCR4 and CCR5 (Figure

5). Chemokine receptors belong to the class of G-protein coupled receptors (GPCRs) that transmit cellular signals throughout the body.[32] These receptors bind to signaling molecules known as chemokines and regulate the movement of immune cells in the body, most notably white blood cells during inflammation.[33] Chemokine receptors play a vital role in HIV-1 infection and progression,[34] hence are considered as major drug targets for treating HIV-1, along with other autoimmune disorders and cancer.[35,36] Specifically, both CXCR4 and CCR5 have been identified as co-receptors for HIV-1 entry into immune cells. Numerous efforts have been made to understand HIV pathology and to develop new therapeutic approaches. Several molecules have been designed to block CXCR4 and CCR5 and are undergoing clinical trials.[37,38] Clinical studies have associated a lack or low expression of CCR5 to provide a natural resistance of HIV infection.[39] Mutational analysis, for example the deep mutational scanning used in this study, provides an invaluable perspective on the function of these receptors, but at a high experimental cost.[9,31] There are more than 20 chemokine receptors in the human body,[40] of which only 2 have a complete mutational dataset.[41] Evaluating TLmutation algorithms on these two datasets will allow us to uncover essential insights into other chemokine receptors and HIV pathology.

In this case study, we utilized available single point mutation datasets for two proteins, CXCR4 and CCR5, and two different experiments, expression level and bimolecular fluorescence complementation (BiFC) assay.[41] Sequence identity and similarity between CXCR4 and CCR5 are $\sim$ 30% and 50%, respectively (Figure 5). The proposed supervised and unsupervised TLmutation algorithms were implemented and shown for one case of supervised transfer and four different cases of unsupervised transfer. The EVmutation models for CCR5 and CXCR4 were built using 95619 and 94461 natural sequences using the algorithm as explained in literature.[16]

We want to learn a Potts model that can predict the effect of mutations on expression levels of CXCR4 given a Potts model trained on natural sequences related to the CXCR4 sequence and a set of 6994 expression levels of single point mutations of CXCR4. Using the

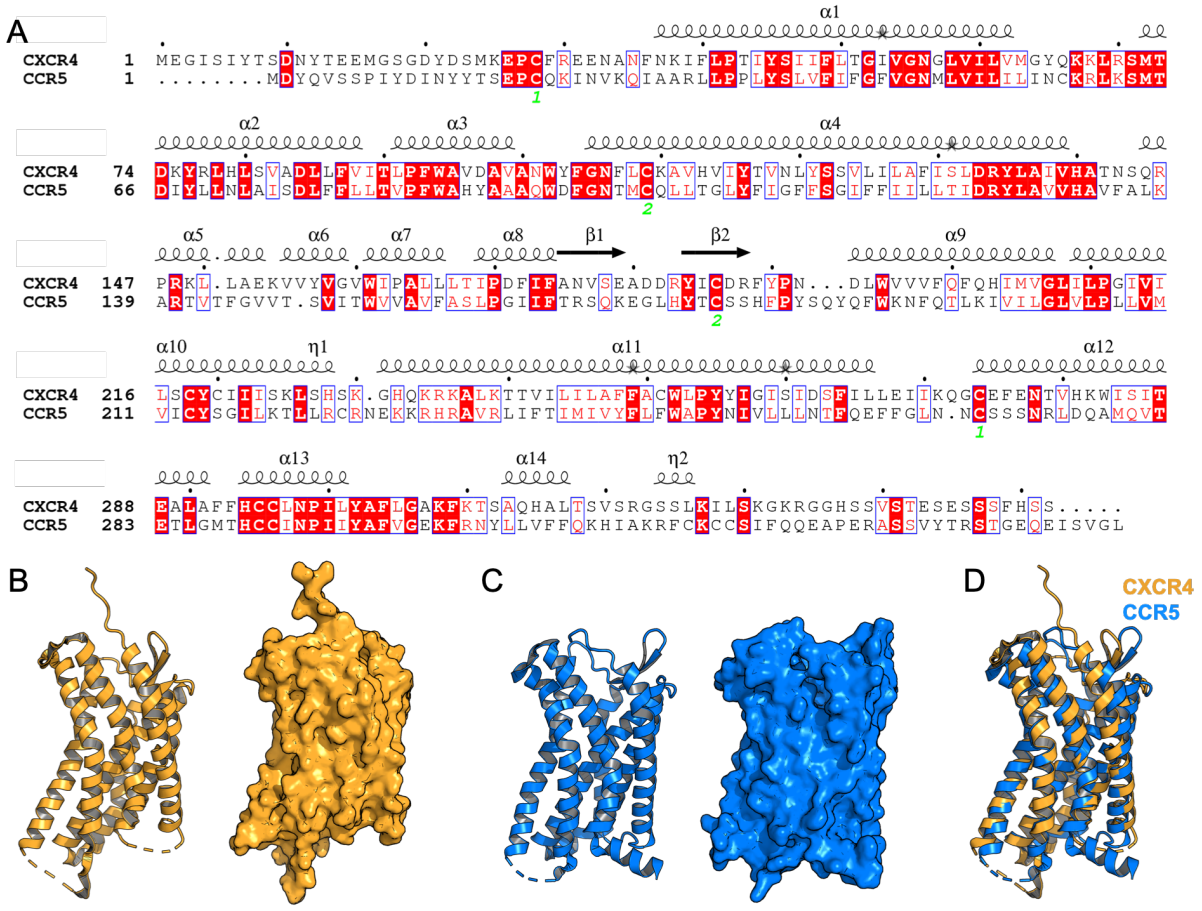

Figure 5: (**A**) Sequence alignments of chemokine receptors CXCR4 and CCR5. Identical residues colored in red boxes, where similar residues are colored in red letters. Corresponding secondary structure is placed above the aligned sequence. The amino acid sequences of the two receptors share a 30% sequence identity and 50% sequence similarity. Crystal structures, depicted in cartoon and surface representation, of CXCR4 (PDB: 4RWS) (**B**) and CCR5 (PDB: 4MBS) (**C**). Cartoon representation of CXCR4 superimposed on CCR5 shows the high sequence and structural similarity between these receptors (**D**).

Table 1: Unsupervised transfer learning between different proteins. Higher Spearman coefficient means better match with the actual data.

Source Protein	New Task (Experiment)	EV's ρ (for Source)	Supervised TL ρ (for Source)	Target Protein	EV ρ (for Target)	Unsupervised TL ρ (for Target)
CCR5	BiFC	0.135	0.348	CXCR4	0.256	0.269
CXCR4	BiFC	0.256	0.449	CCR5	0.135	0.145
CCR5	Expression	0.326	0.514	CXCR4	0.174	0.206
CXCR4	Expression	0.174	0.390	CCR5	0.326	0.357

supervised TLmutation algorithm, a 5-fold cross-validation was performed for the available experimental data on expression levels of CXCR4. The Spearman coefficients for the training data increased from 0.174 ± 0.006 to 0.403 ± 0.005 and the coefficients for the test data increased from 0.174 ± 0.023 to 0.279 ± 0.041 as it is shown in Figure 6 **C**. For all of the 5 folds, the correlation between predicted ranks and actual experiment is higher compared to EV model. For fold 3, the projection of the combined error for each residue is shown on the 3D structure of CXCR4 in Figure 6 **A** and **B**. The overall error is considerably lower for the proposed algorithm, compared to EV model[16] (as the structure is less red).

The effectiveness of the unsupervised TLmutation algorithm in predicting expression levels and BiFC was evaluated on four different cases: (1) transfer from CXCR4 to CCR5 expression, (2) CXCR4 to CCR5 BiFC, (3) CCR5 to CXCR4 expression, and (4) CCR5 to CXCR4 BiFC, as shown in Table **??** and Figure 6 **D**. Spearman coefficients in the third and fourth columns illustrate the improved performance of the supervised-transfer on the training data compared to the EV model and Envision. We note that in previous experiements in this study, we did not compare our results with Envision, as many of these proteins in the training datasets overlapped among the methods. The sixth and seventh columns indicate the Spearman coefficients of test data for EV model and using the unsupervised TL, respectively. In all the cases, we observed that the proposed approach improves the current genomic prediction method (EV model). For the first case (the first row of Table **??**), where the supervised transfer is on expression levels of CXCR4, and the unsupervised

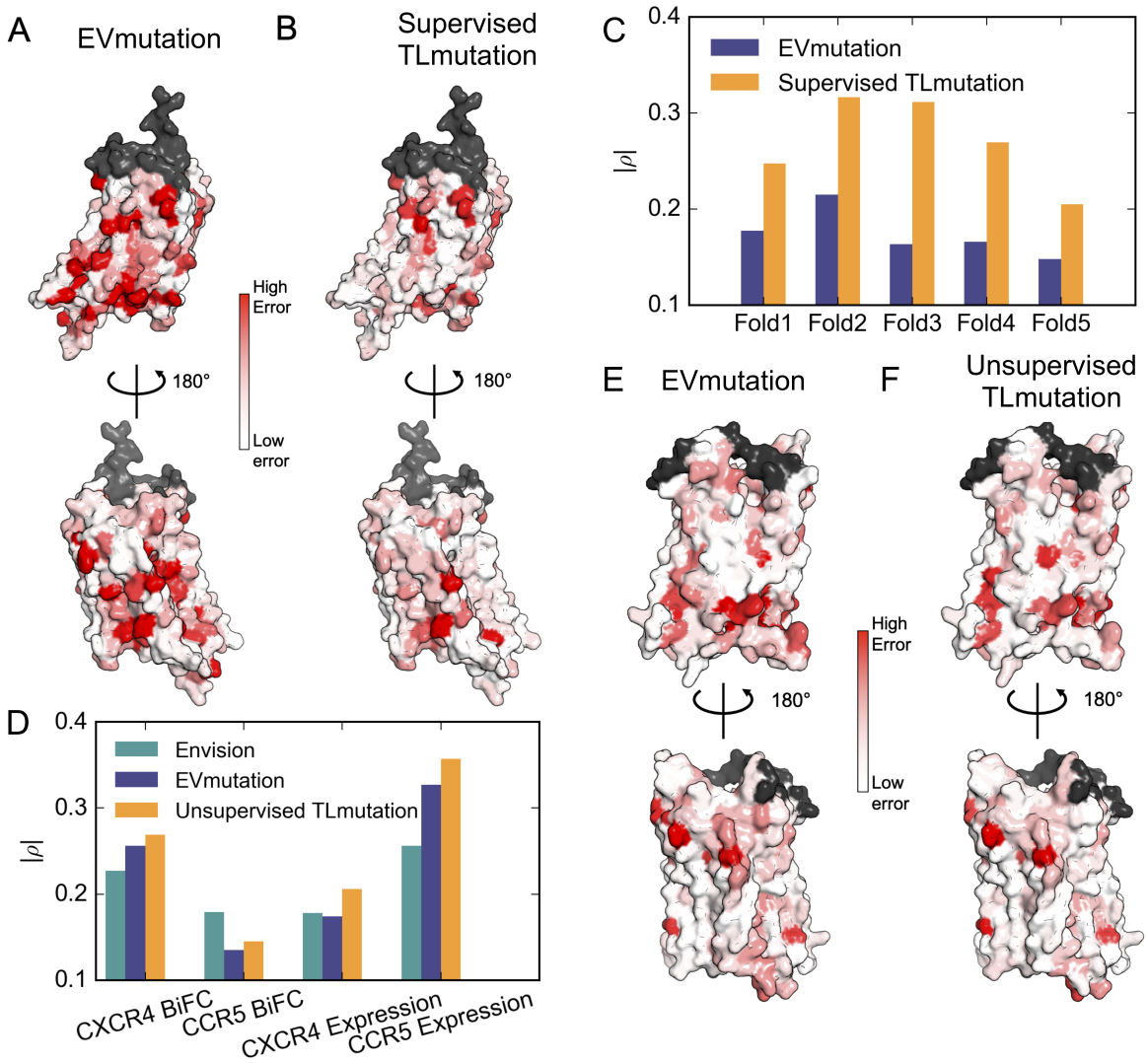

Figure 6: Comparison of the CXCR4 expression levels predicted from the EVmutation (**A**) and supervised TLmutation (**B**). Color indicates relative error with respect to experimental value, ranging from low (white) to high (red). Residues of low-confidence predictions (<20 residues) are colored in grey. (**C**) Test Spearman coefficients in 5-fold cross-validation of CXCR4 on predicting the expression levels. Correlation with experimental values is shown on the Y axis, which measures the performance of the predictive model. Higher Spearman coefficient indicates better match with the data. (**D**) Performance of unsupervised TLmutation in regards to predicting different experiments for the target protein is compared with EVmutation and Envision. For these cases, training data sets are unavailable. The knowledge is transferred to the target in order to predict mutagenesis effects. Error in prediction of CCR5 expression as compared to experimental data using (**E**) EVmutation and (**F**) proposed unsupervised TLmutation algorithm.

transfer is to CCR5, we visualized the prediction errors of each residue with respect to the crystal structure (Figure 6 **E** and **F**). Likewise, the prediction errors are lower for the proposed unsupervised transfer, as compared to the EV model.

Discussion

In this work, we aimed to extract predictive models on biological functionality of proteins from well trained predictive models on the evolutionary data of the same protein and extending it to a new protein via unsupervised transfer learning. We experimentally showed these techniques are efficient in multiple case studies, and compared the results with successful variant predictors EVmutation[16] and Envision.[15]

Another aspect of the work was focused toward understanding which subsets of datapoints yielded the most informative predictions. We showed, through our algorithm, that having few single point mutations on each site was sufficient to estimate the effects of other point mutations in the same sites. However, we believe this question can be further studied. As these experiments are experimentally expensive and the data points are not uniform in terms of information that they have, one natural followup of our work is to extend the algorithm to an active learning approach, where we can suggest to experimentalists, which mutations should be tested over few rounds. As these datasets are susceptible to noise, one approach is to implement the algorithm to enhance low-confidence predictions. By obtaining training data on the parts of the dataset which provide maximum information gained, we can reduce the the the number of experiments and train more powerful models with limited amounts of data.

One of the questions that remains unanswered in this work, is how to define the degree of transferability between domains and tasks. This question depends on the similarities betwen the tasks and domains. In this application of proteomics, we compared the proteins based on their sequence and structural similarities, however as we do not have enough datasets to

check the transferability between multiple proteins, we have not yet defined a clear boundary to conclude whether two proteins are similar enough to allow transfer of knowledge. The example in this paper explored the transferability between CXCR4 and CCR5 proteins with $\sim 30\%$ sequence identity. While we expect systems of high sequence identity to be more transferable, we cannot sufficiently validate this claim due to the lack of mutational datasets of homologous proteins.

Overall, we anticipate supervised TLmutation will be useful in predicting the effects of multiple point mutations and filling out gaps in mutagenesis datasets. Unsupervised TLmutation will help to expand the knowledge to predict the effects of the mutation in many homologous proteins. We expect unsupervised TLmutation continue to improve as more datasets of homologous proteins become available. Furthermore, the proposed transfer learning algorithms were shown to be generalizable to all MRF models, which could be applicable to other disciplines.

References

(1) Wang, Z.; Moult, J. SNPs, protein structure, and disease. *Human mutation* **2001**, *17*, 263–270.

(2) Martinez, J.; Baquero, F. Mutation frequencies and antibiotic resistance. *Antimicrobial agents and chemotherapy* **2000**, *44*, 1771–1777.

(3) Shendure, J.; Akey, J. M. The origins, determinants, and consequences of human mutations. *Science* **2015**, *349*, 1478–1483.

(4) Mann, J. K.; Barton, J. P.; Ferguson, A. L.; Omarjee, S.; Walker, B. D.; Chakraborty, A.; Ndung'u, T. The fitness landscape of HIV-1 gag: advanced modeling approaches and validation of model predictions by in vitro testing. *PLoS computational biology* **2014**, *10*, e1003776.

(5) Gray, V. E.; Hause, R. J.; Fowler, D. M. Analysis of large-scale mutagenesis data to assess the impact of single amino acid substitutions. *Genetics* **2017**, *207*, 53–61.

(6) Stein, A.; Fowler, D. M.; Hartmann-Petersen, R.; Lindorff-Larsen, K. Biophysical and mechanistic models for disease-causing protein variants. *Trends in biochemical sciences* **2019**, *44*, 575–588.

(7) Fowler, D. M.; Stephany, J. J.; Fields, S. Measuring the activity of protein variants on a large scale using deep mutational scanning. *Nature protocols* **2014**, *9*, 2267.

(8) Riesselman, A. J.; Ingraham, J. B.; Marks, D. S. Deep generative models of genetic variation capture the effects of mutations. *Nature Methods* **2018**, *15*, 816–822.

(9) Fowler, D. M.; Fields, S. Deep mutational scanning: a new style of protein science. *Nature methods* **2014**, *11*, 801.

(10) Matreyek, K. A.; Stephany, J. J.; Fowler, D. M. A platform for functional assessment of large variant libraries in mammalian cells. *Nucleic acids research* **2017**, *45*, e102–e102.

(11) Adzhubei, I. A.; Schmidt, S.; Peshkin, L.; Ramensky, V. E.; Gerasimova, A.; Bork, P.; Kondrashov, A. S.; Sunyaev, S. R. A method and server for predicting damaging missense mutations. *Nature methods* **2010**, *7*, 248.

(12) Ng, P. C.; Henikoff, S. SIFT: Predicting amino acid changes that affect protein function. *Nucleic acids research* **2003**, *31*, 3812–3814.

(13) Hecht, M.; Bromberg, Y.; Rost, B. Better prediction of functional effects for sequence variants. *BMC genomics* **2015**, *16*, S1.

(14) Kircher, M.; Witten, D. M.; Jain, P.; O'roak, B. J.; Cooper, G. M.; Shendure, J. A general framework for estimating the relative pathogenicity of human genetic variants. *Nature genetics* **2014**, *46*, 310.

(15) Gray, V. E.; Hause, R. J.; Luebeck, J.; Shendure, J.; Fowler, D. M. Quantitative missense variant effect prediction using large-scale mutagenesis data. *Cell systems* **2018**, *6*, 116–124.

(16) Hopf, T. A.; Ingraham, J. B.; Poelwijk, F. J.; Schärfe, C. P. I.; Springer, M.; Sander, C.; Marks, D. S. Mutation effects predicted from sequence co-variation. *Nature Biotechnology* **2017**, *35*, 128–135.

(17) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The protein data bank. *Nucleic acids research* **2000**, *28*, 235–242.

(18) Starita, L. M.; Young, D. L.; Islam, M.; Kitzman, J. O.; Gullingsrud, J.; Hause, R. J.; Fowler, D. M.; Parvin, J. D.; Shendure, J.; Fields, S. Massively parallel functional analysis of BRCA1 RING domain variants. *Genetics* **2015**, *200*, 413–422.

(19) Starita, L. M.; Pruneda, J. N.; Lo, R. S.; Fowler, D. M.; Kim, H. J.; Hiatt, J. B.; Shendure, J.; Brzovic, P. S.; Fields, S.; Klevit, R. E. Activity-enhancing mutations in an E3 ubiquitin ligase identified by high-throughput mutagenesis. *Proceedings of the National Academy of Sciences* **2013**, *110*, E1263–E1272.

(20) Heredia, J. D.; Park, J.; Choi, H.; Gill, K. S.; Procko, E. Conformational engineering of HIV-1 Env based on mutational tolerance in the CD4 and PG16 bound states. *Journal of virology* **2019**, *93*, e00219–19.

(21) Ahler, E.; Register, A. C.; Chakraborty, S.; Fang, L.; Dieter, E. M.; Sitko, K. A.; Vidadala, R. S. R.; Trevillian, B. M.; Golkowski, M.; Gelman, H., et al. A Combined approach reveals a regulatory mechanism coupling Src's kinase activity, localization, and phosphotransferase-independent functions. *Molecular cell* **2019**, *74*, 393–408.

(22) Pan, S. J.; Yang, Q. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering* **2009**, *22*, 1345–1359.

(23) Shamsi, Z.; Moffett, A. S.; Shukla, D. Enhanced unbiased sampling of protein dynamics using evolutionary coupling information. *Scientific Reports* **2017**, *7*, 12700.

(24) Wei, Z.; Li, H. A Markov random field model for network-based analysis of genomic data. *Bioinformatics* **2007**, *23*, 1537–1544.

(25) Chen, W.; Jin, X.; Li, Z.; Zhang, X.; Hong, L. Clock Synchronization for Distributed Multi-hop Wireless Networks Using Markov Random Field. Journal of Physics: Conference Series. 2018; p 052008.

(26) Jernite, Y.; Rush, A.; Sontag, D. A fast variational approach for learning Markov random field language models. International Conference on Machine Learning. 2015; pp 2209–2217.

(27) Li, S. Z. Markov random field models in computer vision. European conference on computer vision. 1994; pp 361–370.

(28) Kindermann, R.; Snell, J. L. *Markov Random Fields and Their Applications*; American Mathematical Society., 1980.

(29) Myers, L.; Sirois, M. J. Spearman correlation coefficients, differences between. *Encyclopedia of statistical sciences* **2004**, *12*.

(30) Zhang, W.; Dourado, D. F.; Fernandes, P. A.; Ramos, M. J.; Mannervik, B. Multidimensional epistasis and fitness landscapes in enzyme evolution. *Biochemical Journal* **2012**, *445*, 39–46.

(31) Jonson, P. H.; Petersen, S. B. A critical view on conservative mutations. *Protein engineering* **2001**, *14*, 397–402.

(32) Proudfoot, A. E. Chemokine receptors: multifaceted therapeutic targets. *Nature Reviews Immunology* **2002**, *2*, 106.

(33) Hughes, C. E.; Nibbs, R. J. A guide to chemokines and their receptors. *The FEBS journal* **2018**, *285*, 2944–2971.

(34) Wang, Z.; Shang, H.; Jiang, Y. Chemokines and Chemokine Receptors: Accomplices for Human immunodeficiency virus infection and Latency. *Frontiers in immunology* **2017**, *8*, 1274.

(35) Weitzenfeld, P.; Ben-Baruch, A. The chemokine system, and its CCR5 and CXCR4 receptors, as potential targets for personalized therapy in cancer. *Cancer letters* **2014**, *352*, 36–53.

(36) Chow, M. T.; Luster, A. D. Chemokines in cancer. *Cancer immunology research* **2014**, *2*, 1125–1131.

(37) Donzella, G. A.; Schols, D.; Lin, S. W.; Esté, J. A.; Nagashima, K. A.; Maddon, P. J.; Allaway, G. P.; Sakmar, T. P.; Henson, G.; DeClercq, E., et al. AMD3100, a small molecule inhibitor of HIV-1 entry via the CXCR4 co-receptor. *Nature medicine* **1998**, *4*, 72.

(38) Andtbacka, R. H.; Pierce, R. H.; Campbell, J. S.; Yushak, M.; Milhem, M.; Ross, M. I.; Grossmann, K. F.; Sprott, K.; Tsiroyannis, E.; Niland, K., et al. X4P-001, an orally bioavailable CXCR4 antagonist, enhances immune cell infiltration and activation in the tumor microenvironment of melanoma. *RNA (Normalized Counts)* **2018**, *400*, 200.

(39) Ostrowski, M. A.; Justement, S. J.; Catanzaro, A.; Hallahan, C. A.; Ehler, L. A.; Mizell, S. B.; Kumar, P. N.; Mican, J. A.; Chun, T.-W.; Fauci, A. S. Expression of chemokine receptors CXCR4 and CCR5 in HIV-1-infected and uninfected individuals. *The Journal of Immunology* **1998**, *161*, 3195–3201.

(40) Griffith, J. W.; Sokol, C. L.; Luster, A. D. Chemokines and chemokine receptors: positioning cells for host defense and immunity. *Annual review of immunology* **2014**, *32*, 659–702.

(41) Heredia, J. D.; Park, J.; Brubaker, R. J.; Szymanski, S. K.; Gill, K. S.; Procko, E. Mapping Interaction Sites on Human Chemokine Receptors by Deep Mutational Scanning. *The Journal of Immunology* **2018**, *200*, 3825–3839.