Assessment of Software Methods for Estimating Protein-Protein Relative Binding Affinities

Tawny R. Gonzalez[1¶], Kyle P. Martin[1,2¶], Jonathan E. Barnes[1,2], Jagdish Suresh Patel[1,3*], F. Marty Ytreberg[1,2*]

[1]Institute for Modeling Collaboration and Innovation, University of Idaho, Moscow, Idaho, United States of America

[2]Department of Physics, University of Idaho, Moscow, Idaho, United States of America

[3]Department of Biological Sciences, University of Idaho, Moscow, Idaho, United States of America

* Corresponding authors

E-mail: thejagdishpatel@gmail.com (JSP)

E-mail: ytreberg@uidaho.edu (FMY)

¶These authors contributed equally to this work.

Abstract

A growing number of computational tools have been developed to accurately and rapidly predict the impact of amino acid mutations on protein-protein relative binding affinities. Such tools have many applications, for example, designing new drugs and studying evolutionary mechanisms. In the search for accuracy, many of these methods employ expensive yet rigorous molecular dynamics simulations. By contrast, non-rigorous methods use less exhaustive statistical mechanics, allowing for more efficient calculations. However, it is unclear if such methods retain enough accuracy to replace rigorous methods in binding affinity calculations. This trade-off between accuracy and computational expense makes it difficult to determine the best method for a particular system or study. Here, eight non-rigorous computational methods were assessed using eight antibody-antigen and eight non-antibody-antigen complexes for their ability to accurately predict relative binding affinities ($\Delta\Delta G$) for 654 single mutations. In addition to assessing accuracy, we analyzed the CPU cost and performance for each method using a variety of physico-chemical structural features. This allowed us to posit scenarios in which each method may be best utilized. Most methods performed worse when applied to antibody-antigen complexes compared to non-antibody-antigen complexes. Rosetta-based JayZ and EasyE methods classified mutations as destabilizing ($\Delta\Delta G$ < -0.5 kcal/mol) with high (83-98%) accuracy and a relatively low computational cost for non-antibody-antigen complexes. Some of the most accurate results for antibody-antigen systems came from combining molecular dynamics with FoldX with a correlation coefficient (r) of 0.46, but this was also the most computationally expensive method. Overall, our results suggest these methods can be used to quickly and accurately predict stabilizing versus destabilizing mutations but are less accurate at predicting actual binding affinities. This study highlights the need for continued development of reliable, accessible, and reproducible

methods for predicting binding affinities in antibody-antigen proteins and provides a recipe for using current methods.

Introduction

Protein-protein binding is an essential physiological event that governs a large number of biological processes in the cell [1]. Amino acid mutations of these proteins can introduce diversity into genomes, and disrupt or modulate protein-protein interactions by changing the underlying binding free energy (ΔG, i.e. binding affinity), the amount of energy required to form protein complexes [2]. The binding free energy associated with a protein-protein complex determines the stability of the complex formation and the conditions for protein-protein association. Accurate prediction of binding free energies allows us to understand how these affinities can be modified, and leads to a more comprehensive understanding of protein interactions in living organisms [3].

Experimental biophysical methods can quantitatively measure change in the protein-protein binding free energy due to a mutation (i.e. relative binding affinity, $\Delta\Delta G$), but these methods are typically costly, laborious, and time-consuming since all mutant proteins must be expressed and purified. Many researchers have developed and utilized computational methods to predict $\Delta\Delta G$ values for single- or multiple-amino acid mutations (see e.g. [4-6]). Historically, the most promising in terms of accuracy are rigorous methods based on statistical mechanics that use molecular dynamics (MD) simulations and thus automatically address conformational flexibility and entropic effects [7, 8]. However, these methods are computationally expensive since they employ rigorous sampling and utilize classical mechanics [9] or quantum mechanics [10] approximations of intermolecular interactions, and require a large number of calculations per time-step. Because of the expense, rigorous methods are not well-suited to studying large sets of mutations or large proteins thus necessitating less expensive, non-rigorous methods.

Non-rigorous high-throughput methods attempt to lower the computational cost, as compared to rigorous methods, while still providing accurate $\Delta\Delta G$ predictions. They accomplish this by including precalculated physico-chemical structural information in combination with predictive algorithms. The core mechanics that drive these methods fall under numerous classification umbrellas which have been covered by review articles [11, 12]. These review articles provide a broad overview but do not provide an unbiased, rigorous, comparative analysis outside of what the original developers provide. The developers of any given method tend to provide comparisons with other methods of the same general class to define where their method fits in the current landscape. BindProfX, for example, is available as a web server and standalone and utilizes structure-based interface profiles with pseudo counts. Upon release, it was most notably compared to FoldX (a semi-empirical trained method [13]) and DCOMPLEX (a physics-based method [14]) [15, 16]. iSEE, a statistically trained method based on 31 structure, evolution, and energy-based terms was tested against FoldX, BindProfX, and BeAtMuSiC (a machine learning-based approach [17]). Mutabind [18] and some other methods not explored in this work follow a similar testing methodology [19-21]. While these comparisons are beneficial in providing context for how a given model fits in the existing research landscape, they are not very robust, since only a narrow subset of methodologies are included. Conversely for folding stability, Kroncke et al. compared a large number of available software methods on a small dataset of transmembrane proteins providing a general overview of performance [6]. Despite the narrow dataset, this study provides a diverse, useful collection of evaluation metrics between multiple classes of methods. Our intent in this study is to provide a similar robust comparison of methods for non-rigorous binding affinity estimation.

In this work, we evaluate the ability of eight non-rigorous methods to predict relative binding affinities due to single amino acid mutations. We restrict our study to cases where both an experimental structure of the complex, and experimentally determined binding affinity values are available. To investigate the trade-off between speed and accuracy, we chose 16 protein-protein test complexes with empirical $\Delta\Delta G$ values for observed mutations. We calculated the $\Delta\Delta G$ values for each mutation using all eight methods and compared the results against empirical $\Delta\Delta G$ values. The goal of this study was to determine whether software methods that use (most costly) energy functions with a wider variety of physico-chemical structural features would provide more accurate binding affinity and interface destabilization predictions compared to those that rely on a single descriptive (less costly) energy function. We have determined scenarios in which some of these methods may be better or worse than traditional computational methods in predicting $\Delta\Delta G$ values.

Methods

Compilation of Experimental ΔΔG Values

To assess the performance of a range of protein-protein binding affinity prediction methods, we first assembled a dataset containing single amino acid mutations with known experimental $\Delta\Delta G$ values. This list was assembled from Structural Kinetic and Energetic database of Mutant Protein Interaction (SKEMPI) version 2.0 [22]. While generating this list, we considered four aspects: (i) type of protein-protein complex; (ii) availability of quality 3-D structural information; (iii) range of experimental $\Delta\Delta G$ values; and (iv) the type of mutations at differing sites on the complex. Our final dataset contained 654 mutations from 16 protein-protein complexes and their respective experimental $\Delta\Delta G$ values. We further categorized these 16 complexes as either non-antibody-antigen (non-Ab) or antibody-antigen (Ab). Table 1 shows the complexes in our dataset with their respective non-Ab and Ab categories and the number of mutations associated with each complex. The dataset contains a total of 401 non-Ab mutations and 253 Ab mutations.

Non-Ab			Ab		
PDB ID	**# Mutations**	**# Residues**	**PDB ID**	**# Mutations**	**# Residues**
1a4y [23]	32	583	1bj1[24]	10	547
1brs [25]	30	199	1jrh [26]	42	540
1cbw [27]	31	299	1mlc [28]	11	561
1iar [29]	36	336	1vfb [30]	48	352
1jtg [31]	37	428	1yy9 [32]	16	1058
1lfd [33]	19	254	2jel [34]	43	520
1ppf [35]	190	274	3hfm [36]	71	558
2wpt [37]	26	220	4i77 [38]	12	549

Table 1. Dataset used in our study containing 16 protein complexes. For both non-Ab (left) and Ab (right) categories, columns show PDB IDs, total number of residues in a complex, and number of experimental mutants per complex.

Selection of Protein-Protein Binding Affinity Methods

Binding affinity prediction methods were chosen to have both a distinct approach to binding affinity calculation that utilized 3-D structural information and had functional standalone software in September 2020, available either online or upon request to the author. Table 2 summarizes the methods selected in this study, their approaches, and their type of scoring functions. For simplicity, we categorized scoring functions (mathematical functions to calculate $\Delta\Delta G$ values) as semi-empirical, statistical, or physics-based. Semi-empirical methods replace as many calculations as possible with pre-calculated data and are trained using existing crystal structures and known binding affinity measurements for mutations [39]. Statistical methods use pre-calculated data and consider changes in coarse structural features such as the change in overall volume [40]. Physics-based methods use molecular mechanics based-energy functions to estimate enthalpic binding contributions [14]. In general, statistical or semi-empirical scoring functions involve a training step where existing datasets are leveraged to determine the weight of input parameters. MD, JayZ, and EasyE were not developed by training the methods against experimental data designed to improve predictive power while all other methods utilized this step.

Name	Brief Description	Scoring Function	Runtime (CPU hours)
BindProfX [15, 16]	Interface profile score based on conservation of homologous interfaces	Semi-Empirical	1ppf = 0.57 CPUh 1yy9 = 0.73 CPUh
BindProfX(BPX)+FoldX v4 [15, 16]	Profile score weighted and combined with FoldX energy potential	Semi-Empirical	1ppf = 0.62 CPUh 1yy9 = 0.71 CPUh
iSEE [41]	Random forest model using structural, evolutionary, and energy-based features	Statistical	1ppf < 0.01 CPUh 1yy9 < 0.01 CPUh
DCOMPLEX v2 [14]	Structural ideal-gas reference state potential	Physics-Based	1ppf = 0.013 CPUh 1yy9 = 0.001 CPUh
EasyE v1.0 [40, 42]	GMEC-based method utilizing the Rosetta [43, 44] energy function	Statistical	1ppf = 0.48 CPUh 1yy9 = 0.09 CPUh
JayZ v1.0 [40, 42]	Partition-function method utilizing Rosetta energy function	Statistical	1ppf = 0.14 CPUh 1yy9 = 0.21 CPUh
FoldX v4 [13, 39]	Empirical energy score based on various energy parameters (e.g. van der Waals, solvation, electrostatics, hydrogen bonding)	Semi-Empirical	1ppf = 0.42 CPUh 1yy9 = 0.16 CPUh
MD+FoldX v4 [45-47]	Molecular dynamics used to explore conformation space and generate snapshots; FoldX score calculated for each snapshot and averaged	Semi-Empirical	1ppf = 941 CPUh 1yy9 = 4093 CPUh

Table 2. Methods used for comparison in study with a short summary of their approach and scoring function. Columns (left to right) indicate the method, a brief description of the method, the type of scoring function used, and runtimes. Runtimes are the amount of CPU hours for estimating the $\Delta\Delta G$ for a representative protein complex for Ab (1yy9, 1058 residues) and Non-Ab (1ppf, 274 residues) categories. Although 1yy9 is roughly four times bigger than 1ppf, the total runtime may or may not be affected depending on the method used.

Calculation and Comparison of Computational Speed

The methods in Table 2 were used to predict $\Delta\Delta G$ values for each mutation on our experimental list shown in Table 1. Detailed protocols for predicting $\Delta\Delta G$ values using each selected method are provided in the Supplemental Information (see S1 File). Runtimes were determined by using a representative protein complex from each category: 1ppf, a non-Ab complex with 274 total amino acids, and 1yy9, an Ab complex with 1058 total amino acids (see Table 2). These runtimes are estimates provided to give a point of comparison between the speeds of different methods. Specific runtimes will be determined by hardware specifications, method parameters, the number of mutations being computed, and overall protein size. For MD+FoldX, computational runtime was the length of time of the MD simulation plus the FoldX runtime for a single mutation. Reporting runtime in this fashion highlights the large CPUh requirement needed in order to add the sampling of MD into FoldX calculations. We note that, in contrast to the other methods tested here, the MD simulations that must be performed for MD+FoldX can be completed very quickly on modern GPUs, significantly offsetting the high initial cost of the MD+FoldX method. For all other methods, the algorithms rely either on various pre-calculated data or limited conformational sampling to calculate $\Delta\Delta G$ values rapidly.

Comparing Experimental and Predicted ΔΔG Values

To carry out statistical analysis of our results we built an in-house Python script (see S2 File) that uses a combination of libraries including matplotlib, numpy, pandas, statistics, scipy, and sklearn. Using this script, we compared predicted values to experimental $\Delta\Delta G$ values for each method.

To evaluate the predictive ability of each method tested, we compared the following correlation coefficients using our script: concordance (ρ_c), Pearson (r), Kendall (τ), and Spearman (ρ) (see Table 3). We distinguish between methods that were trained to predict $\Delta\Delta G$ values from methods that compute metrics that are expected to linearly correlate with $\Delta\Delta G$ values. This distinction is important since for optimal performance we expect a regression line that passes through the coordinate origin and has a slope of 1, leading to a correlation coefficient equal to 1.

Correlation	Brief Description	Type
Concordance	The concordance correlation coefficient (ρ_c) measures the degree to which the predicted $\Delta\Delta G$ value equals the actual experimental value (0 indicates no agreement and 1 perfect agreement).	Linear
Pearson	The Pearson correlation coefficient (r) measures the degree to which a uniform linear transformation of the predicted $\Delta\Delta G$ values (i.e., a shift and scale change) would yield the actual experimental values (0 indicates no agreement after transformation, 1 perfect agreement, and −1 perfect inverse agreement).	Linear
Kendall and Spearman	The rank correlation coefficient measures the degree to which the rank ordering of the predicted $\Delta\Delta G$ values matches the rank ordering of the actual experimental values (0 indicates no agreement after transformation, 1 perfect agreement, and −1 perfect inverse agreement). In a normal case, the Kendall correlation (τ) is considered more robust than the Spearman correlation (ρ) because of a smaller gross error sensitivity and more efficient due to a smaller asymptotic variance [48].	Rank
AUC and ROC	The receiver operating characteristic (ROC) curve tests several cutoff values for binning mutations as neutral or destabilizing between the most negative calculated $\Delta\Delta G$ value and the most positive calculated $\Delta\Delta G$ value, with true positive rates (sensitivity) calculated at each point. As the true positive rate is calculated, the classifier is moved to less extreme values; this yields the ROC curve. The area under curve (AUC) is a summary statistic that approximates how well the predictor actually discriminates between the two classifications.	N/A

Table 3. Statistical measures used to test the performance of each method in predicting $\Delta\Delta G$ values.

To compare the discriminating power of the methods, we generated receiver operating characteristic (ROC) curves (see Table 3). These curves quantify the ability of a method to correctly classify point mutations as destabilizing ($\Delta\Delta G < -0.5$ kcal/mol) or neutral/stabilizing ($\Delta\Delta G > -0.5$ kcal/mol). ROC curves that are skewed toward a higher true positive rate (sensitivity) classify mutations more accurately, as quantified by area under curve (AUC, ranging between 1.0 and 0.5 for perfect and chance classification, respectively).

We also used our script to parse the results on the basis of several physico-chemical and structural features to allow us to evaluate the methods based on these characteristics: wild type amino acid type, mutant amino acid type, protein-protein interacting versus antibody-antigen, secondary structure classification of the mutation [49, 50], coordination number [51], Sneath index [52], mostly α-helical proteins versus mostly β-sheet proteins versus a mix of both α-helical and β-sheet proteins, percent exposure, location of the mutation, change in charge, change in polarity, change in volume, and whether or not the mutation location is predicted as an active or passive residue [53-55]. The script uses data from S3 File as an input and outputs scatter plots, correlation plots, receiver operating characteristic (ROC) curves, and box plots to visualize the data, as well as correlations and standard deviations for each method. All plots in this manuscript were generated using this script.

Results

The purpose of our study was to assess the ability of eight different relative binding affinity calculation methods (see Table 2) to estimate $\Delta\Delta G$ values. We selected 16 different protein complexes (eight Ab, eight non-Ab, see Table 1) with a total of 654 single amino acid mutations. Each method was then used to estimate $\Delta\Delta G$ values of 654 mutations and a variety of statistical measures were employed to assess their predictive ability. We also examined the computational speed of each method in the context of accuracy to determine its efficiency.

Non-Antibody-Antigen (non-Ab) Results

Our dataset of eight non-Ab test protein complexes contains 401 total mutations and are mainly classified as protein-protein systems formed by inhibitors and receptors that range from 199 to 583 residues in size. The distribution and our classification of experimental $\Delta\Delta G$ values for all non-Ab test complexes is as follows: 13% of point mutations resulted in $\Delta\Delta G$ values less than -0.5 kcal/mol (classified as destabilizing); 31% between -0.5 and 0.5 kcal/mol (neutral); and 56% greater than 0.5 kcal/mol (stabilizing).

Figures 1 (blue data points and values) and 2 show various performance metrics for each method to assess their ability to predict the non-Ab $\Delta\Delta G$ values. Overall, EasyE has the highest correlation coefficient, $r = 0.62$, and iSEE has the lowest, $r = 0.17$ (see Figures 1 and 2). JayZ and EasyE, both of which utilize Rosetta's conformational sampling algorithms, consistently have the best r values for non-Ab mutations.

Figure 1. **Calculated ΔΔ*G* values (x-axis) compared to experimental ΔΔ*G* values (y-axis) for each method tested in this study**. Black, red, and blue lines are simple linear regressions from which r are derived. The red points are a scatter for Ab complexes and the blue points are for non-Ab complexes. The dashed line is the $y = x$ line measuring perfect agreement between predicted and experimental ΔΔ*G* values. The solid black, red, and blue lines indicate a linear relationship between calculated and experimental observations for all data points, Ab complexes, and non-Ab complexes respectively. The top values in black, red, and blue match the root-mean-square error and the bottom values indicate r for all values, Ab values, and non-Ab values respectively.

Figure 2. **Performance of each method for non-Ab complexes (401 total mutations) in predicting true ΔΔ*G* values (ρ_c), linearly correlated ΔΔ*G* values (r), and rank order (ρ and τ).** The error for each method is reported under the correlation points.

Figure 3 shows the ROC plot for all the tested methods. These ROC plots highlight how well a method can discriminate between stabilizing and destabilizing mutations. JayZ (0.84), EasyE (0.83), DCOMPLEX (0.82), FoldX (0.79), and MD+FoldX (0.76) have the highest AUC. Combined with the results from Figures 1 and 2, for the systems studied here, JayZ and EasyE methods are the best overall performers in terms of accuracy, discriminating stabilizing mutations from destabilizing, and ranking mutations based on their experimental ΔΔ*G* values.

Figure 3. **Receiver operating characteristic (ROC) curves for non-Ab complexes of the classification of variants as stabilizing (ΔΔ*G* < -0.5 kcal/mol) or destabilizing (ΔΔ*G* > 0.5 kcal/mol)**. The values in the legend represent the area-under-curve (AUC). The higher the value, the better method is at discriminating between destabilizing and destabilizing mutations.

Table 2 reports CPUh required (i.e. runtimes) for each method to calculate ΔΔ*G* for the entire list of mutations for a representative non-Ab protein complex. BindProfX, BindProfX(BPX)+FoldX, JayZ, and EasyE allow users to specify a list of mutations that the method is then able to calculate in one setting. This list can be optimized based on the available hardware to achieve efficiency.

iSEE requires significant preparatory work (see File S1) prior to calculation, but once completed, it calculates the $\Delta\Delta G$ values for the entire list of mutations nearly instantly. DCOMPLEX is not as flexible out of the box but can handle large numbers of mutations through an automated script. For MD+FoldX, 1yy9 (roughly four times larger than 1ppf) requires considerably more CPUh to calculate. All other methods calculate 1yy9 in a shorter time frame than 1ppf. This may seem counterintuitive. However, MD must statistically sample the conformational energy of the entire complex, while all other methods use algorithms that are likely impacted more by the number of residues involved in the interaction rather than the protein size. Overall, DCOMPLEX has a much faster runtime compared to other methods, and if the goal is to determine stabilizing and destabilizing non-Ab mutations, it offers similar discriminating power to JayZ and EasyE, at a fraction of the computational cost. JayZ estimates $\Delta\Delta G$ value of one mutation in ~2.7 s, EasyE in ~9.1 s, but DCOMPLEX requires just ~0.25 s. Overall, EasyE appears to be the best option for balancing accuracy and speed and DCOMPLEX is recommended for discriminating between stability and destabilizing mutations.

A method might not be a good overall performer in predicting $\Delta\Delta G$ values but could still perform well for mutations with certain physico-chemical and structural features. Therefore, we calculated various statistical measures to assess each method on unique subsets of mutations (see Table 4 and SI Figs S1-4). This table shows eight different data subsets with two r per method. EasyE has the highest r for non-Ab for five out of eight subsets (wild type non-gly or non-pro, alpha helix, beta sheet, surface exposure, and large volume changes). Where this method did not have the highest r, it had either the second or third highest r. JayZ mirrors the performance of EasyE in all the same categories and performs better than Easy in the neutral charge subset. These results further highlight the versatility of EasyE's and JayZ's performance in estimating the effects of non-Ab mutations compared to the other methods tested in this study. All methods apart from iSEE and BindProfX perform surprisingly well in the WT Gly or Pro subset. iSEE's performance in this subset, while still mediocre compared to the other tested methods, is substantially better than in all other subsets.

Method	WT Gly or Pro	WT Non-Gly or Non-Pro	Alpha Helix	Beta Sheet	Surface Exposure	Neutral Charge	Hydrophobic to Polar	Large Vol Changes
BindProfX	Non-Ab: 0.11 Ab: -0.03	Non-Ab: 0.33 Ab: 0.23	Non-Ab: 0.29 Ab: 0.16	Non-Ab: 0.29 **Ab: 0.52**	Non-Ab: 0.22 Ab: 0.09	**Non-Ab: 0.37** Ab: 0.28	Non-Ab: 0.33 Ab: 0.17	Non-Ab: 0.13 Ab: 0.42
BPX+FoldX	**Non-Ab: 0.81** Ab: 0.09	Non-Ab: 0.45 Ab: 0.34	Non-Ab: 0.43 Ab: 0.39	Non-Ab: 0.43 **Ab: 0.54**	Non-Ab: 0.32 Ab: 0.21	Non-Ab: 0.52 Ab: 0.41	Non-Ab: 0.41 Ab: 0.26	Non-Ab: 0.71 **Ab: 0.50**
FoldX	**Non-Ab: 0.85** Ab: -0.11	Non-Ab: 0.45 Ab: 0.25	Non-Ab: 0.39 Ab: 0.25	Non-Ab: 0.39 Ab: 0.31	Non-Ab: 0.50 Ab: 0.26	Non-Ab: 0.42 **Ab: 0.41**	Non-Ab: 0.41 Ab: 0.11	Non-Ab: 0.63 Ab: -0.32
MD+FoldX	**Non-Ab: 0.83** **Ab: 0.71**	Non-Ab: 0.49 **Ab: 0.42**	Non-Ab: 0.44 **Ab: 0.54**	Non-Ab: 0.44 Ab: 0.49	Non-Ab: 0.47 Ab: 0.35	Non-Ab: 0.46 Ab: 0.46	**Non-Ab: 0.46** **Ab: 0.31**	Non-Ab: 0.71 Ab: 0.35
DCOMPLEX	**Non-Ab: 0.65** **Ab: 0.89**	Non-Ab: 0.34 Ab: 0.37	Non-Ab: 0.33 Ab: 0.31	Non-Ab: 0.33 Ab: 0.30	Non-Ab: 0.52 Ab: 0.27	Non-Ab: 0.36 **Ab: 0.56**	Non-Ab: 0.38 Ab: 0.16	Non-Ab: 0.62 Ab: 0.28
JayZ	Non-Ab: 0.80 **Ab: 0.54**	Non-Ab: 0.49 Ab: 0.24	Non-Ab: 0.44 Ab: -0.06	Non-Ab: 0.44 Ab: 0.16	Non-Ab: 0.59 **Ab: 0.36**	**Non-Ab: 0.62** Ab: 0.26	Non-Ab: 0.41 Ab: 0.01	**Non-Ab: 0.83** Ab: 0.19
EasyE	Non-Ab: 0.80 Ab: 0.29	**Non-Ab: 0.51** Ab: 0.22	**Non-Ab: 0.51** Ab: 0.06	**Non-Ab: 0.51** Ab: 0.03	**Non-Ab: 0.60** **Ab: 0.35**	Non-Ab: 0.61 Ab: 0.23	Non-Ab: 0.45 Ab: 0.02	**Non-Ab: 0.84** Ab: 0.18
iSEE	**Non-Ab: 0.43** Ab: -0.43	Non-Ab: 0.28 Ab: -0.16	Non-Ab: 0.05 Ab: -0.04	Non-Ab: 0.05 Ab: -0.24	Non-Ab: 0.15 **Ab: 0.11**	Non-Ab: 0.15 Ab: -0.11	Non-Ab: 0.14 Ab: -0.02	Non-Ab: 0.24 Ab: -0.44

Table 4. **All methods *r* with respect to certain subsets**. "WT Gly or Pro" are wild type amino acids that are either glycine or proline. "WT Non-Gly or Non-Pro" are wild type amino acids that are neither glycine nor proline. "Alpha Helix" are mutations that occur in a helix structure. "Beta Sheet" are mutations that occur in a beta structure. "Surface Exposure" are mutations that occur in an amino acid that have relative solvent accessibility values between 0 and 10%. "Neutral Charge" is a neutrally charged wild type amino acid mutating to a neutrally charged mutant amino acid. "Hydrophobic to Polar" is a hydrophobic or polar wild type amino acid mutating to a polar or hydrophobic mutant amino acid, respectively. "Larger Vol Changes" is a mutant amino acid that is greater than 40% larger than the wild type amino acid. Values that are bolded are the highest *r* for each method and protein type. Values that are red or blue are the highest *r* for each subset, blue for non-Ab and red for Ab.

Antibody-Antigen (Ab) Results

Our dataset of eight Ab test protein complexes contains 253 mutations and the proteins range in size from 352 to 1058 residues. The distribution and our classification of experimental $\Delta\Delta G$ values for all Ab test complexes is as follows: 5% of point mutations resulted in $\Delta\Delta G$ values less than -

0.5 kcal/mol (classified as destabilizing); 40% between -0.5 and 0.5 kcal/mol (neutral); and 55% greater than 0.5 kcal/mol (stabilizing).

Figures 1 (data points and values in red), 4, and 5 show the performance of each method in predicting the $\Delta\Delta G$ values of Ab mutations. Overall, the highest correlation is for MD+FoldX with r = 0.39 and the lowest is iSEE with r = -0.09 (see Figures 1 and 4). An interesting trend is that the methods with the highest r values for non-Ab complexes do not have the highest r for Ab complexes.

Figure 4. **Performance of each evaluated method for Ab complexes (253 total mutations) in predicting true $\Delta\Delta G$ values (ρ_c), linearly correlated $\Delta\Delta G$ values (r), and rank order (ρ and τ)**. The error for each method is reported under the correlation points.

Figure 5. **Receiver operating characteristic curves of the classification of variants that are more destabilized or less destabilized than 0.5 kcal/mol**. The values in the legend represent the area-under-curve (AUC). The higher the value, the better the prediction capability of the method.

Figure 5 shows the ROC plot for all the tested Ab methods. These ROC plots highlight how well a method is actually able to discriminate between stabilizing and destabilizing mutations. Compared to non-Ab complexes, all methods performed better for antibody-antigen complexes except for FoldX and DCOMPLEX which were marginally worse. JayZ (0.97), EasyE (0.98), FoldX (0.85), and MD+FoldX (0.82) had the highest AUC values. Combined with the results from Figures 1 and 4, at least for the systems studied here, it appears that the MD+FoldX method is the best overall performer in terms of accuracy, discriminating stabilizing mutations from destabilizing, and ranking mutations based on their experimental $\Delta\Delta G$ values.

Compared to other methods, EasyE has a much faster runtime and is recommended if the goal is to discriminate between stabilizing and destabilizing ($\Delta\Delta G$ for one mutation takes ~21 s, see Table 2). By comparison, MD+FoldX cost ~941 CPUh for one mutation of 1yy9. DCOMPLEX provides a slightly lower r (0.31) and computational cost (~0.35 s) for one mutation of 1yy9. Overall, MD+FoldX appears to be the best option for accuracy and EasyE or JayZ are the best options for discriminating between destabilizing and stabilizing mutations.

Table 4 summarizes the ability of each method to predict $\Delta\Delta G$ values for subsets of Ab mutations. Most methods had mediocre r values less than 0.60. The exceptions to this are MD+FoldX and DCOMPLEX within the WT Gly or Pro subset with $r = 0.71$ and 0.89, respectively. MD+FoldX has the highest r for Ab complexes for three of the eight subsets (WT nonGly or nonPro, alpha helix, and hydrophobic to polar). BPX+FoldX has the highest r in two of the eight subsets (beta sheet and large volume changes). For the beta sheet subset, BindProfX had the second highest r. DCOMPLEX had the highest r for two different subsets (WT Gly or Pro and neutral charge). In the surface exposure subset, JayZ and EasyE both had nearly identical r (0.36 and 0.35 respectively), the highest for this subset, but substantially worse than they did for non-Ab complexes.

Discussion

We assessed the performance of eight distinct protein-protein binding affinity calculation methods that use 3-D structural information. To test the performance of these methods, we selected 16 different protein complexes (see Table 1) with a total of 654 single amino acid mutations: eight antigen-antibody complexes (Ab, 253 mutations) and eight non-antigen-antibody (Non-Ab, 401 mutations) complexes. Each method was used to estimate $\Delta\Delta G$ values of the 654 mutations, a variety of statistical measures, CPU cost, and physico-chemical structural features to assess the performance. Our results suggest each method has both strengths and weaknesses depending on the properties of the protein system. Most methods did not perform well when applied to mutations in Ab complexes compared to non-Ab complexes. Rosetta-based JayZ and EasyE were able to classify mutations as destabilizing ($\Delta\Delta G$ < -0.5 kcal/mol) with high (83-98%) accuracy at relatively low computational cost. Some of the best results for Ab systems came from combining MD simulations with FoldX with a r coefficient of 0.39, but at the highest computational cost of all the tested methods.

Figure 1 summarizes the performance of each method in terms of its ability to estimate $\Delta\Delta G$ values for all (non-Ab + Ab) single mutations. None of the test methods show a very high r between experimental and predicted $\Delta\Delta G$ values. Two of the best performing methods, JayZ and EasyE, both have an r of 0.49 for all mutations, with a higher r of 0.61 and 0.62 respectively for non-Ab complexes. These results agree with published results from the authors of JayZ and EasyE. Our results agree moderately with published results from iSEE (they obtained r = 0.25, we obtained r = 0.17) and BindProfX (they used a much larger dataset). Published results for DCOMPLEX show a very good correlation of r = 0.87; much larger than what we obtained here. This difference is very likely due to the dataset size and compilation; DCOMPLEX was originally tested against 69 experimental data points, compared to the 654 values used here. MD+FoldX has an r of 0.39 for Ab complexes and appears to perform well for larger systems, which could indicate the importance of conformational sampling for antibody-antigen systems. Other methods used in this study have little to no conformational sampling which could explain their poor performance on Ab complexes. By contrast, these same methods perform well for non-Ab complexes, suggesting that conformational sampling is not the limiting factor to achieve accurate results for these protein complexes. For example, FoldX has a trained scoring function derived using a dataset of mostly non-Ab complexes and performs poorly for Ab complexes when using a single structure (see Table 2). However, when used with snapshots from an MD simulation, this same method outperforms all other methods selected in this study. This highlights the need for conformational sampling for reliable and efficient predictions of binding affinity for some systems. In our previous study, we combined coarse-grained forcefield with umbrella sampling to calculate $\Delta\Delta G$ values for eight mutations of 3hfm Ab complex (one of the test systems in this study) and obtained better predictions than FoldX and MD+FoldX [56]. This study further emphasizes the need for better conformational strategies for some systems.

Statistical measures used to analyze performance are listed and defined in Table 3. For Ab, BPX+FoldX, MD+FoldX, and DCOMPLEX have the highest r values of the methods in our study (see Figure 4). MD+FoldX appears to be the most accurate method for Ab complexes. BindProfX, FoldX, JayZ, EasyE, and iSEE have low r and ρ_c indicating that affinities estimated using these methods do not correlate well with experimental $\Delta\Delta G$ values using a linear transformation. Also, the τ and ρ were lower compared to MD+FoldX, indicating these methods do poorly at calculating a rank order that matches experimental data.

The ROC curves allow us to determine each method's ability to classify mutations as either destabilizing or neutral/stabilizing (Figures 3 and 5). For non-Ab complexes, JayZ (0.84 AUC) and EasyE (0.83 AUC) have the best true positive rate followed by DCOMPLEX (0.82 AUC). For Ab complexes, JayZ (0.97 AUC) and EasyE (0.98 AUC) have better true positive rates than MD+FoldX, the method with the highest r value. If classification of destabilizing vs stabilizing is the primary need, then JayZ or EasyE are both recommended over the other methods tested here due to their high accuracy and fast runtime.

While accuracy is generally the main reason for choosing a particular method, computational efficiency is also an important consideration, especially when predicting the effects of a large number of mutations. Here, we discuss the performance of each method in terms of its trade-off between speed and accuracy for predicting $\Delta\Delta G$ values. For all single mutations and our non-Ab subset, EasyE and JayZ performed well; JayZ is the faster method of the two with EasyE at a similar speed to FoldX. DCOMPLEX is more accurate than FoldX for all single mutations and has similar accuracy as FoldX for non-Ab mutations, but at much lower cost. MD+FoldX has similar accuracy to DCOMPLEX for all single mutations and has similar accuracy to FoldX in non-Ab mutations but is by far the most computationally expensive method we tested. Although a synergistic combination of BPX+FoldX implements several structural and physico-chemical interaction terms in its algorithm, computation time was longer than all but MD+FoldX without a concomitant improvement in r. We note that this method is perhaps the most accessible of those tested, due to the easy-to-use online server interface and accuracy that is similar to FoldX for most systems. BindProfX utilizes the same scoring profile as BPX+FoldX without the FoldX calculations. In this case, accuracy decreased while calculation speed remained similar to BPX+FoldX. iSEE, the least correlating method, employs the widest variety of information to obtain relative binding affinity predictions and is the fastest of all methods (not including the non-trivial preparation time). For Ab complexes, MD+FoldX, the slowest of all the methods, had the highest accuracy, followed by DCOMPLEX. iSEE is again the fastest of all methods but also the least accurate. BindProfX utilizes several pre-calculated physico-chemical structural data in its scoring function while, JayZ and EasyE each layer an additional predictive calculating feature on top of Rosetta's backbone sampling, adding complexity to the predictive algorithms. However, all three have similar *r* yet they do not achieve the accuracy of MD+FoldX. Overall, for non-Ab complexes, EasyE and JayZ appear to have the best balance between speed and accuracy of the methods we tested. For Ab complexes, DCOMPLEX appears to have the best balance.

We have demonstrated that all the tested methods have specific strengths and weaknesses; some perform better in specific contexts (Table 4), and some have longer runtimes to obtain similar predictive power to comparably faster methods. This highlights the complexity of the physico-chemical properties and structural features that drive, and limit, these predictive models. Our results can be used to make informed decisions for methods that may be preferable for a particular study or system. Table 4 suggests that if the goal is to estimate only the order of magnitude or sign of relative binding affinities, then the preferred method will likely be very different than if the goal is to obtain the best possible accuracy for antibody-antigen systems. To improve accessibility, we have generated an in-house Python script (provided in the supplement with the full dataset used in this work) that can be used to parse any of the parameters used in this study and provide tailored information. This information in combination with the runtime and other details provided in this study can be used to inform users on methods that can provide the best accuracy and efficiency for a given protein-protein complex type, set of physico-chemical features or structural parameters, and set of mutations. Additionally, the script can be extended to other methods and feature-sets, potentially elucidating specific problems or areas of improvement to existing and future methods.

Conclusions

In this study, we have assessed the accuracy and efficiency of eight computational methods on predicting binding affinity changes due to single amino acid mutations. Methods were tested on 16 different protein complexes: eight antigen-antibody (Ab) and eight non-antigen-antibody (Non-Ab) complexes. While some methods perform consistently better than others, how well each performs depends on the physico-chemical and structural components of each complex. EasyE was the most accurate for non-Ab complexes, and MD+FoldX was most accurate for Ab complexes. JayZ and EasyE were better able to distinguish between destabilizing ($\Delta\Delta G > 0.5$ kcal/mol) and stabilizing ($\Delta\Delta G < -0.5$ kcal/mol) as compared to any other method. Future work could include more systems or different methods, including those that are solely web server-based in order to expand and better refine our conclusions on their predictive capability.

References

1. Jones S, Thornton JM. Principles of protein-protein interactions. Proceedings of the National Academy of Sciences. 1996;93(1):13-20.
2. Yates CM, Sternberg MJE. The Effects of Non-Synonymous Single Nucleotide Polymorphisms (nsSNPs) on Protein–Protein Interactions. Journal of Molecular Biology. 2013;425(21):3949-63.
3. Baaden M, Marrink SJ. Coarse-grain modelling of protein-protein interactions. Curr Opin Struct Biol. 2013;23(6):878-86. PubMed PMID: 24172141.
4. Ezkurdia I, Bartoli L, Fariselli P, Casadio R, Valencia A, Tress ML. Progress and challenges in predicting protein-protein interaction sites. Briefings in bioinformatics. 2009;10(3):233-46. PubMed PMID: 19346321.
5. Kastritis PL, Bonvin AM. Are scoring functions in protein-protein docking ready to predict interactomes? Clues from a novel binding affinity benchmark. J Proteome Res. 2010;9(5):2216-25. PubMed PMID: 20329755.
6. Kroncke BM, Duran AM, Mendenhall JL, Meiler J, Blume JD, Sanders CR. Documentation of an Imperative To Improve Methods for Predicting Membrane Protein Stability. Biochemistry. 2016;55(36):5002-9.
7. Bernardi RC, Melo MC, Schulten K. Enhanced sampling techniques in molecular dynamics simulations of biological systems. Biochim Biophys Acta. 2015;1850(5):872-7. PubMed PMID: 25450171; PubMed Central PMCID: PMCPMC4339458.
8. Spiwok V, Sucur Z, Hosek P. Enhanced sampling techniques in biomolecular simulations. Biotechnol Adv. 2015;33(6 Pt 2):1130-40. PubMed PMID: 25482668.
9. Gumbart JC, Roux B, Chipot C. Efficient Determination of Protein–Protein Standard Binding Free Energies from First Principles. J Chem Theory Comput. 2013;9(8):3789-98.
10. Pokorná P, Kruse H, Krepl M, Šponer J. QM/MM Calculations on Protein–RNA Complexes: Understanding Limitations of Classical MD Simulations and Search for Reliable Cost-Effective QM Methods. J Chem Theory Comput. 2018;14(10):5419-33.
11. Geng C, Xue LC, Roel-Touris J, Bonvin AMJJ. Finding the ΔΔG spot: Are predictors of binding affinity changes upon mutations in protein–protein interactions ready for it? WIREs Computational Molecular Science. 2019;9(5):e1410.
12. Gromiha MM, Yugandhar K, Jemimah S. Protein-protein interactions: scoring schemes and binding affinity. Curr Opin Struct Biol. 2016;44:31-8. PubMed PMID: 27866112.
13. Schymkowitz J, Borg J, Stricher F, Nys R, Rousseau F, Serrano L. The FoldX web server: an online force field. Nucleic acids research. 2005;33(Web Server issue):W382-8. PubMed PMID: 15980494; PubMed Central PMCID: PMCPMC1160148.
14. Liu S, Zhang C, Zhou H, Zhou Y. A physical reference state unifies the structure-derived potential of mean force for protein folding and binding. Proteins: Structure, Function, and Bioinformatics. 2004;56(1):93-101.
15. Xiong P, Zhang C, Zheng W, Zhang Y. BindProfX: Assessing Mutation-Induced Binding Affinity Change by Protein Interface Profiles with Pseudo-Counts. J Mol Biol. 2017;429(3):426-34.
16. Brender JR, Zhang Y. Predicting the Effect of Mutations on Protein-Protein Binding Interactions through Structure-Based Interface Profiles. PLoS Comp Biol. 2015;11(10):e1004494. PubMed PMID: 26506533; PubMed Central PMCID: PMCPMC4624718.
17. Dehouck Y, Kwasigroch JM, Rooman M, Gilis D. BeAtMuSiC: Prediction of changes in protein-protein binding affinity on mutations. Nucleic Acids Res. 2013;41(Web Server issue):W333-9. PubMed PMID: 23723246; PubMed Central PMCID: PMCPMC3692068.
18. Li M, Simonetti FL, Goncearenco A, Panchenko AR. MutaBind estimates and interprets the effects of sequence variants on protein–protein interactions. Nucleic Acids Res. 2016;44(W1):W494-W501.
19. Vreven T, Hwang H, Pierce BG, Weng Z. Prediction of protein–protein binding free energies. Protein Sci. 2012;21(3):396-404.

20. Rodrigues CHM, Myung Y, Pires DEV, Ascher DB. mCSM-PPI2: predicting the effects of mutations on protein–protein interactions. Nucleic Acids Res. 2019;47(W1):W338-W44.

21. Jemimah S, Sekijima M, Gromiha MM. ProAffiMuSeq: sequence-based method to predict the binding free energy change of protein–protein complexes upon mutation using functional classification. Bioinformatics. 2019;36(6):1725-30.

22. Jankauskaitė J, Jiménez-García B, Dapkūnas J, Fernández-Recio J, Moal IH. SKEMPI 2.0: an updated benchmark of changes in protein–protein binding energy, kinetics and thermodynamics upon mutation. Bioinformatics. 2018;35(3):462-9.

23. Papageorgiou AC, Shapiro R, Acharya KR. Molecular recognition of human angiogenin by placental ribonuclease inhibitor—an X-ray crystallographic study at 2.0 Å resolution. The EMBO Journal. 1997;16(17):5162-77.

24. Muller YA, Chen Y, Christinger HW, Li B, Cunningham BC, Lowman HB, et al. VEGF and the Fab fragment of a humanized neutralizing antibody: crystal structure of the complex at 2.4 å resolution and mutational analysis of the interface. Structure. 1998;6(9):1153-67.

25. Buckle AM, Schreiber G, Fersht AR. Protein-protein recognition: crystal structural analysis of a barnase-barstar complex at 2.0-A resolution. Biochemistry. 1994;33(30):8878-89. PubMed PMID: 8043575.

26. Sogabe S, Stuart F, Henke C, Bridges A, Williams G, Birch A, et al. Neutralizing epitopes on the extracellular interferon γ receptor (IFNγR) α-chain characterized by homolog scanning mutagenesis and X-ray crystal structure of the A6 Fab-IFNγR1-108 complex11Edited by R. Huber. J Mol Biol. 1997;273(4):882-97.

27. Scheidig AJ, Hynes TR, Pelletier LA, Wells JA, Kossiakoff AA. Crystal structures of bovine chymotrypsin and trypsin complexed to the inhibitor domain of alzheimer's amyloid β-protein precursor (APPI) and basic pancreatic trypsin inhibitor (BPTI): Engineering of inhibitors with altered specificities. Protein Sci. 1997;6(9):1806-24.

28. Braden BC, Souchon H, Eiselé J-L, Bentley GA, Bhat TN, Navaza J, et al. Three-dimensional structures of the free and the antigen-complexed Fab from monoclonal anti-lysozyme antibody D44.1. J Mol Biol. 1994;243(4):767-81.

29. Hage T, Sebald W, Reinemer P. Crystal Structure of the Interleukin-4/Receptor α Chain Complex Reveals a Mosaic Binding Interface. Cell. 1999;97(2):271-81.

30. Bhat TN, Bentley GA, Boulot G, Greene MI, Tello D, Dall'Acqua W, et al. Bound water molecules and conformational stabilization help mediate an antigen-antibody association. Proceedings of the National Academy of Sciences. 1994;91(3):1089-93.

31. Lim D, Park HU, De Castro L, Kang SG, Lee HS, Jensen S, et al. Crystal structure and kinetic analysis of β-lactamase inhibitor protein-II in complex with TEM-1 β-lactamase. Nat Struct Biol. 2001;8(10):848-52.

32. Li S, Schmitz KR, Jeffrey PD, Wiltzius JJW, Kussie P, Ferguson KM. Structural basis for inhibition of the epidermal growth factor receptor by cetuximab. Cancer Cell. 2005;7(4):301-11.

33. Huang L, Hofer F, Martin GS, Kim S-H. Structural basis for the interaction of Ras with RaIGDS. Nat Struct Biol. 1998;5(6):422-6.

34. Prasad L, Waygood EB, Lee JS, Delbaere LTJ. The 2.5 Å resolution structure of the Jel42 Fab fragment/HPr complex11Edited by I. A. Wilson. J Mol Biol. 1998;280(5):829-45.

35. Bode W, Wei AZ, Huber R, Meyer E, Travis J, Neumann S. X-ray crystal structure of the complex of human leukocyte elastase (PMN elastase) and the third domain of the turkey ovomucoid inhibitor. EMBO J. 1986;5(10):2453-8. PubMed PMID: 3640709; PubMed Central PMCID: PMCPMC1167139.

36. Padlan EA, Silverton EW, Sheriff S, Cohen GH, Smith-Gill SJ, Davies DR. Structure of an antibody-antigen complex: crystal structure of the HyHEL-10 Fab-lysozyme complex. Proc Natl Acad Sci U S A. 1989;86(15):5938-42. PubMed PMID: 2762305; PubMed Central PMCID: PMCPMC297746.

37. Meenan NAG, Sharma A, Fleishman SJ, MacDonald CJ, Morel B, Boetzel R, et al. The structural and energetic basis for high selectivity in a high-affinity protein-protein interaction. Proceedings of the National Academy of Sciences. 2010;107(22):10080-5.

38. Ultsch M, Bevers J, Nakamura G, Vandlen R, Kelley RF, Wu LC, et al. Structural Basis of Signaling Blockade by Anti-IL-13 Antibody Lebrikizumab. J Mol Biol. 2013;425(8):1330-9.
39. Guerois R, Nielsen JE, Serrano L. Predicting Changes in the Stability of Proteins and Protein Complexes: A Study of More Than 1000 Mutations. Journal of Molecular Biology. 2002;320(2):369-87.
40. Viricel C, de Givry S, Schiex T, Barbe S. Cost function network-based design of protein–protein interactions: predicting changes in binding affinity. Bioinformatics. 2018;34(15):2581-9.
41. Geng C, Vangone A, Folkers GE, Xue LC, Bonvin AMJJ. iSEE: Interface structure, evolution, and energy-based machine learning predictor of binding affinity changes upon mutations. Proteins: Structure, Function, and Bioinformatics. 2019;87(2):110-9.
42. Hurley B, O'Sullivan B, Allouche D, Katsirelos G, Schiex T, Zytnicki M, et al. Multi-language evaluation of exact solvers in graphical model discrete optimization. Constraints. 2016;21(3):413-34.
43. Alford RF, Leaver-Fay A, Jeliazkov JR, O'Meara MJ, DiMaio FP, Park H, et al. The Rosetta All-Atom Energy Function for Macromolecular Modeling and Design. J Chem Theory Comput. 2017;13(6):3031-48.
44. Park H, Bradley P, Greisen P, Liu Y, Mulligan VK, Kim DE, et al. Simultaneous Optimization of Biomolecular Energy Functions on Features from Small Molecules and Macromolecules. J Chem Theory Comput. 2016;12(12):6201-12.
45. Miller CR, Johnson EL, Burke AZ, Martin KP, Miura TA, Wichman HA, et al. Initiating a watch list for Ebola virus antibody escape mutations. PeerJ. 2016;4:e1674.
46. Patel JS, Quates CJ, Johnson EL, Ytreberg FM. Expanding the watch list for potential Ebola virus antibody escape mutations. PLOS ONE. 2019;14(3):e0211093.
47. Yang J, Naik N, Patel JS, Wylie CS, Gu W, Huang J, et al. Predicting the viability of beta-lactamase: How folding and binding free energies correlate with beta-lactamase fitness. PloS one. 2020;15(5):e0233509.
48. Croux C, Dehon C. Influence functions of the Spearman and Kendall correlation measures. Statistical Methods & Applications. 2010;19(4):497-515.
49. Tien MZ, Meyer AG, Sydykova DK, Spielman SJ, Wilke CO. Maximum Allowed Solvent Accessibilites of Residues in Proteins. PloS one. 2013;8(11):e80635.
50. Kabsch W, Sander C. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. Biopolymers. 1983;22(12):2577-637.
51. Tribello GA, Bonomi M, Branduardi D, Camilloni C, Bussi G. PLUMED 2: New feathers for an old bird. Computer Physics Communications. 2014;185(2):604-13.
52. Sneath PHA. Relations between chemical structure and biological activity in peptides. J Theor Biol. 1966;12(2):157-95.
53. van Zundert GCP, Rodrigues JPGLM, Trellet M, Schmitz C, Kastritis PL, Karaca E, et al. The HADDOCK2.2 Web Server: User-Friendly Integrative Modeling of Biomolecular Complexes. J Mol Biol. 2016;428(4):720-5.
54. Dominguez C, Boelens R, Bonvin AMJJ. HADDOCK: A Protein−Protein Docking Approach Based on Biochemical or Biophysical Information. J Am Chem Soc. 2003;125(7):1731-7.
55. Wassenaar TA, van Dijk M, Loureiro-Ferreira N, van der Schot G, de Vries SJ, Schmitz C, et al. WeNMR: Structural Biology on the Grid. Journal of Grid Computing. 2012;10(4):743-67.
56. Patel JS, Ytreberg FM. Fast Calculation of Protein–Protein Binding Free Energies Using Umbrella Sampling with a Coarse-Grained Model. J Chem Theory Comput. 2018;14(2):991-7.

Supporting information captions

S1 File. A word document with detailed protocols for calculating ΔΔ*G* values using each of the eight methods used in this study.

S2 File. An in-house Python script that can be used to parse any of the parameters used in this study and provide tailored information.

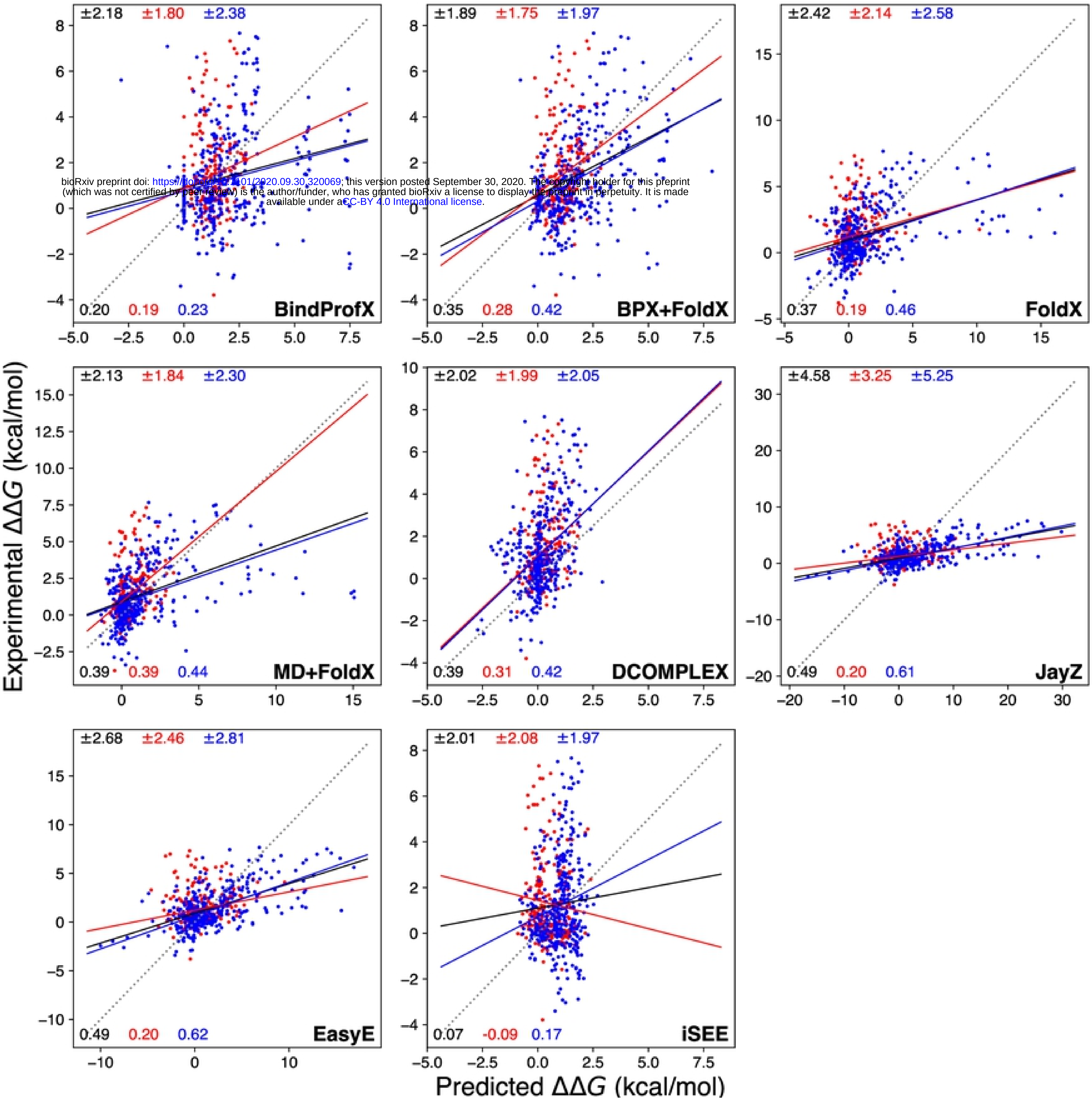
S3 File. **A CSV file with full dataset used in this work and predicted ΔΔ*G* values for each mutation using eight methods**.

S1 Figure. Performance of each evaluated method for Ab and non-Ab complexes in predicting true ΔΔ*G* values (ρ_c), linearly correlated ΔΔ*G* values (r), and rank order (ρ and τ) for a select subset of mutations that occur in beta sheet. The error for each method is reported under the correlation points.

S2 Figure. Performance of each evaluated method for Ab and non-Ab complexes in predicting true ΔΔ*G* values (ρ_c), linearly correlated ΔΔ*G* values (r), and rank order (ρ and τ) for a select subset of mutations that occur in alpha helix. The error for each method is reported under the correlation points.

S3 Figure. Performance of each evaluated method for Ab and non-Ab complexes in predicting true ΔΔ*G* values (ρ_c), linearly correlated ΔΔ*G* values (r), and rank order (ρ and τ) for a select subset of mutations with wild type amino acids that are either glycine or proline. The error for each method is reported under the correlation points.

S4 Figure. Performance of each evaluated method for Ab and non-Ab complexes in predicting true ΔΔ*G* values (ρ_c), linearly correlated ΔΔ*G* values (r), and rank order (ρ and τ) for a select subset of mutations with wild type amino acids that are neither glycine nor proline. The error for each method is reported under the correlation points.



±2.18 ±1.80 ±2.38
0.20 0.19 0.23
BindProfX
±1.89 ±1.75 ±1.97
0.35 0.28 0.42
BPX+FoldX
±2.42 ±2.14 ±2.58
0.37 0.19 0.46
FoldX
±2.13 ±1.84 ±2.30
0.39 0.39 0.44
MD+FoldX
±2.02 ±1.99 ±2.05
0.39 0.31 0.42
DCOMPLEX
±4.58 ±3.25 ±5.25
0.49 0.20 0.61
JayZ
±2.68 ±2.46 ±2.81
0.49 0.20 0.62
EasyE
±2.01 ±2.08 ±1.97
0.07 -0.09 0.17
iSEE
Experimental ΔΔG (kcal/mol)
Predicted ΔΔG (kcal/mol)

Figure 1

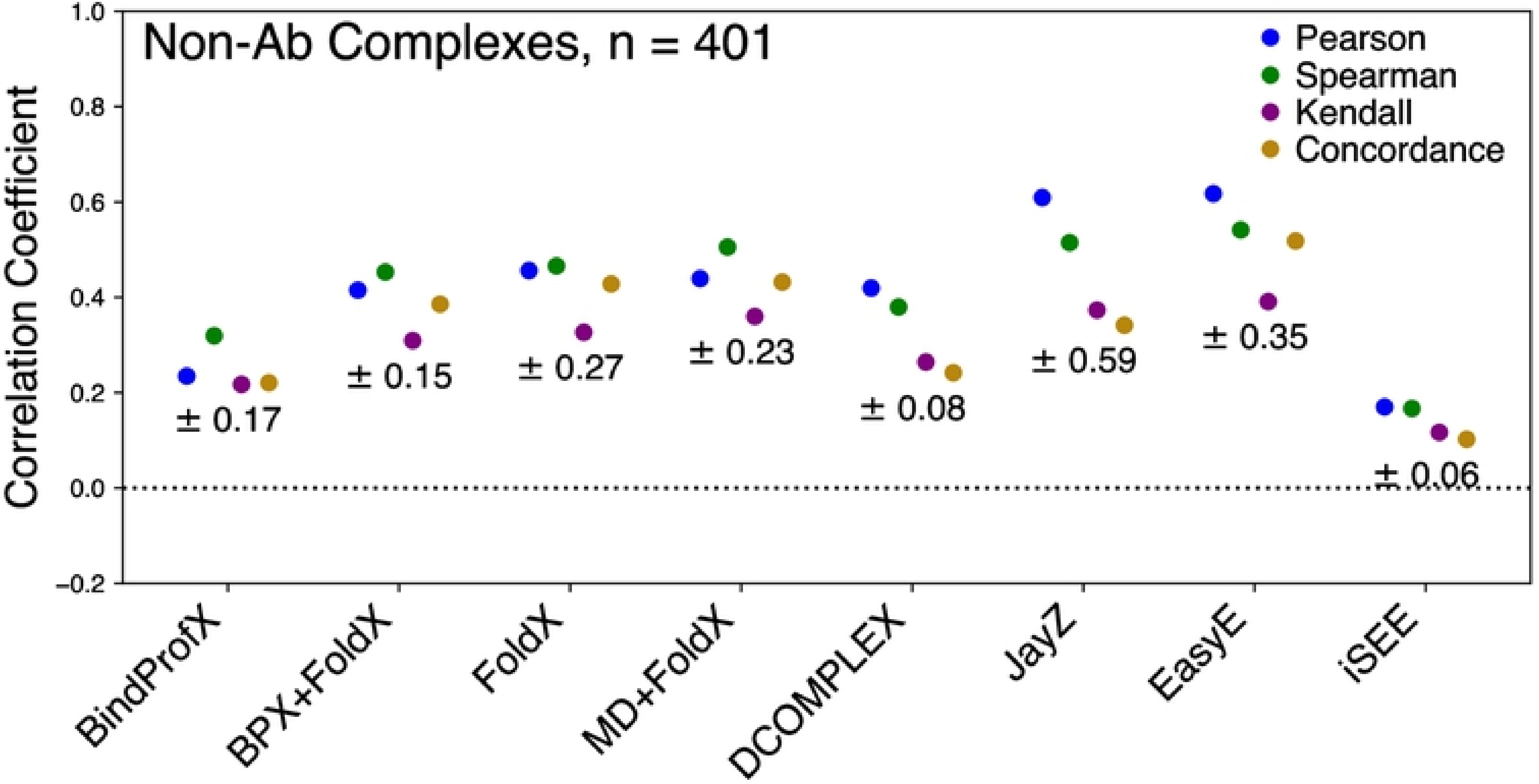
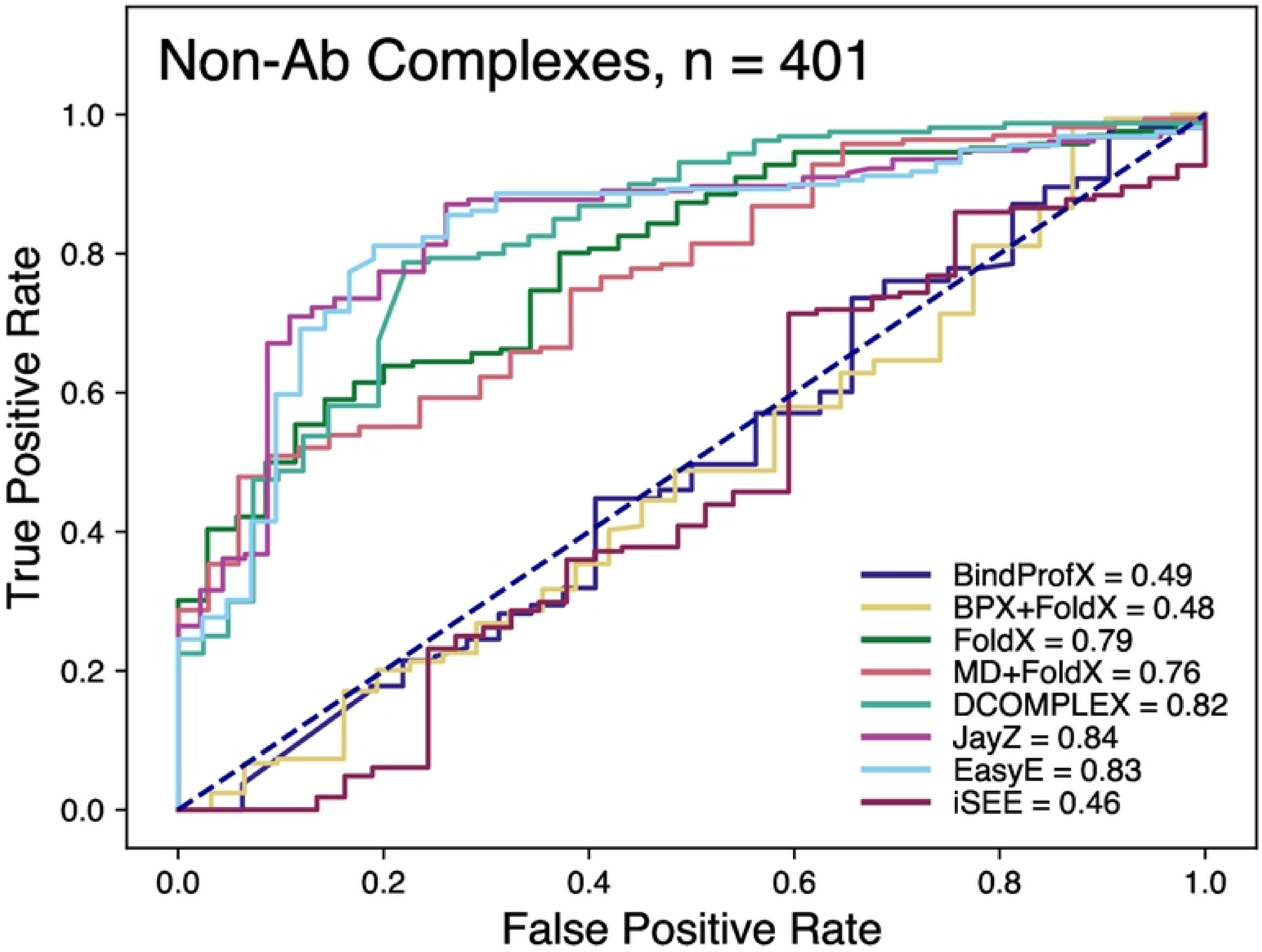
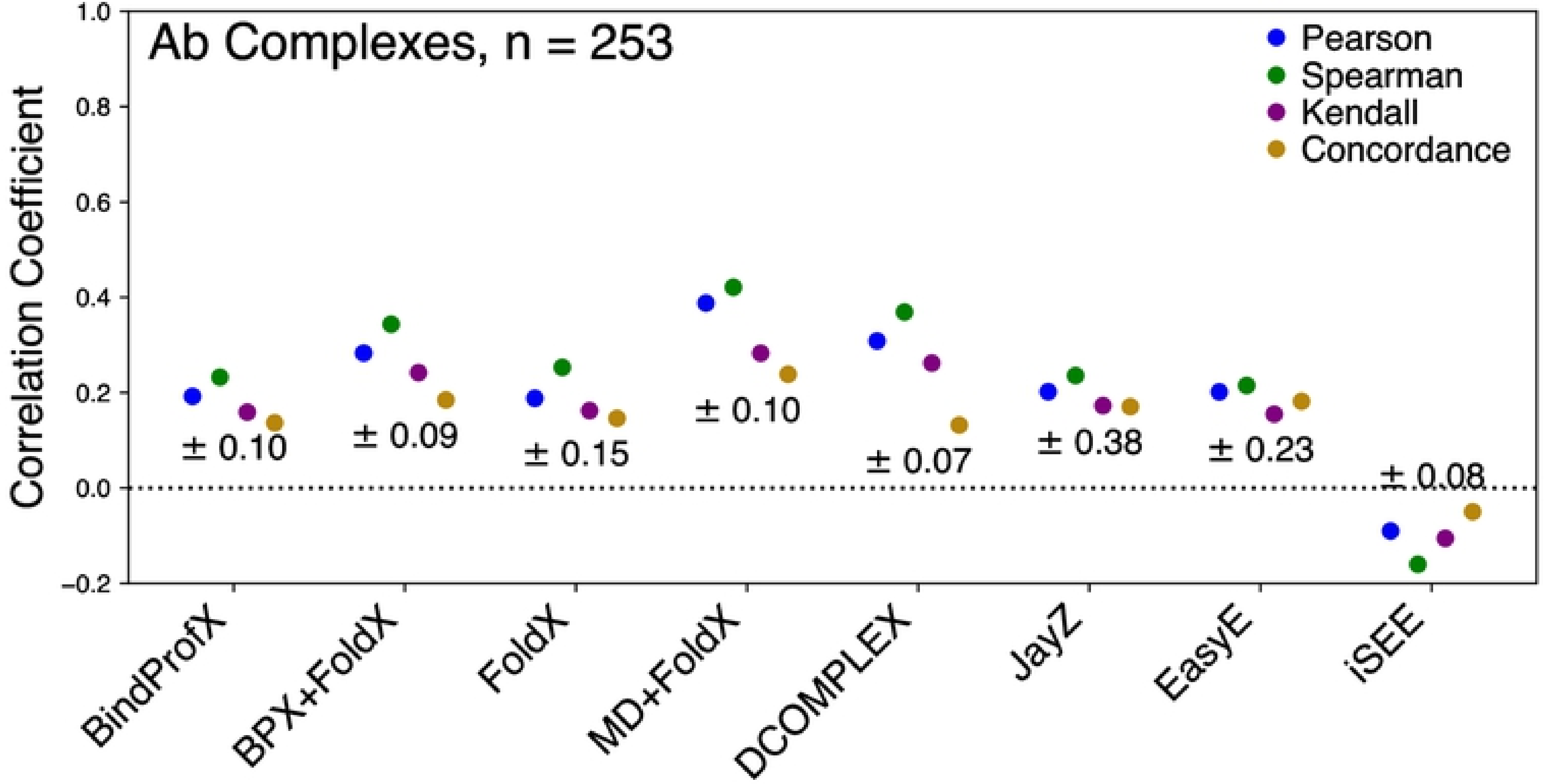

Non-Ab Complexes, n = 401
Pearson
Spearman
Kendall
Concordance
Correlation Coefficient
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
± 0.17
± 0.15
± 0.27
± 0.23
± 0.08
± 0.59
± 0.35
± 0.06
BindProfX
BPX+FoldX
FoldX
MD+FoldX
DCOMPLEX
JayZ
EasyE
iSEE

Figure 2



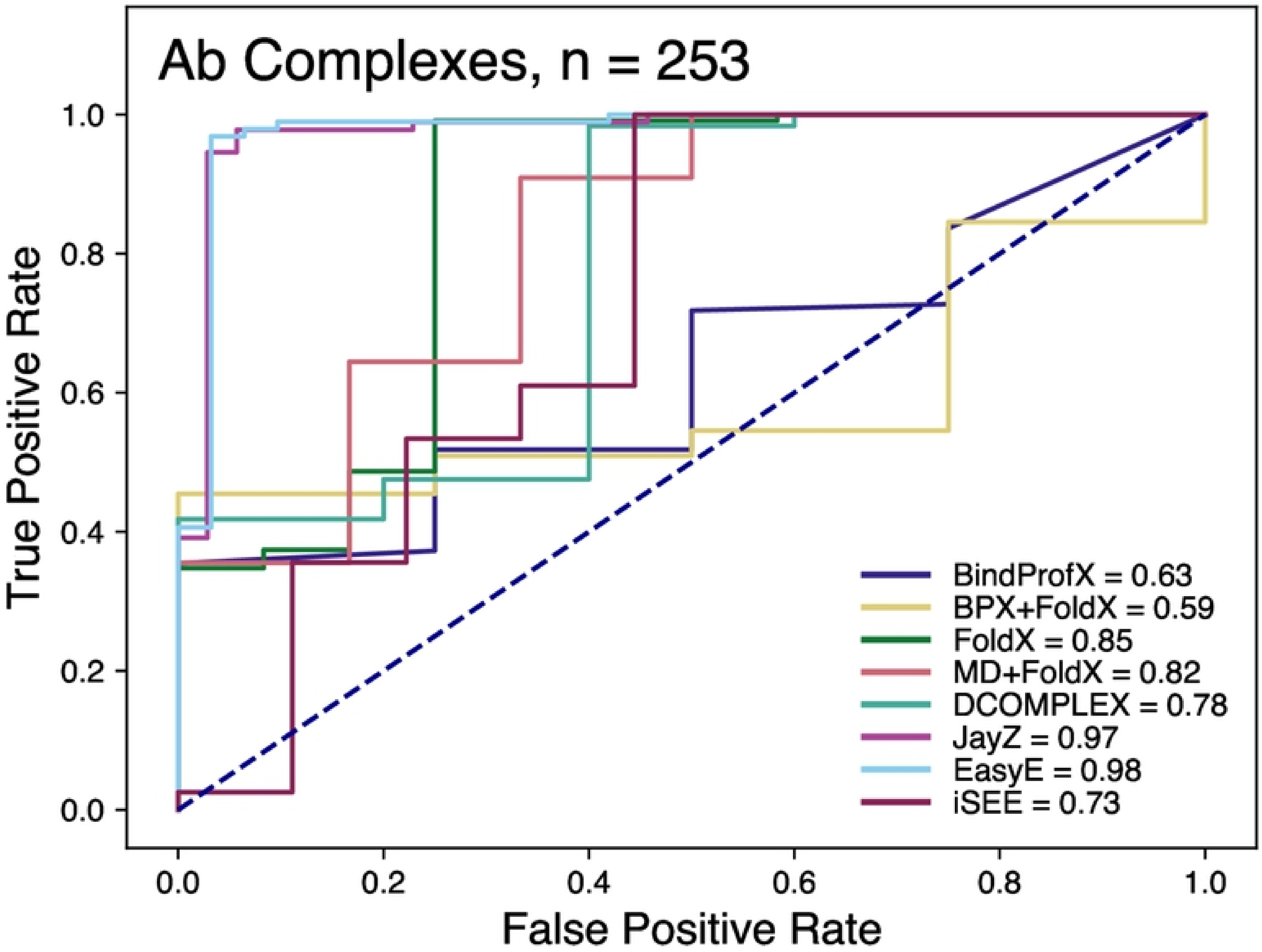
Non-Ab Complexes, n = 401
1.0
0.8
0.6
0.4
0.2
0.0
True Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
False Positive Rate
BindProfX = 0.49
BPX+FoldX = 0.48
FoldX = 0.79
MD+FoldX = 0.76
DCOMPLEX = 0.82
JayZ = 0.84
EasyE = 0.83
iSEE = 0.46

Figure 3



Ab Complexes, n = 253
Pearson
Spearman
Kendall
Concordance
Correlation Coefficient
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
± 0.10
± 0.09
± 0.15
± 0.10
± 0.07
± 0.38
± 0.23
± 0.08
BindProfX
BPX+FoldX
FoldX
MD+FoldX
DCOMPLEX
JayZ
EasyE
iSEE

Figure 4



Ab Complexes, n = 253
1.0
0.8
0.6
0.4
0.2
0.0
True Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
False Positive Rate
BindProfX = 0.63
BPX+FoldX = 0.59
FoldX = 0.85
MD+FoldX = 0.82
DCOMPLEX = 0.78
JayZ = 0.97
EasyE = 0.98
iSEE = 0.73

Figure 5