Investigating Mechanistic Models In Predicting Pathogenicity of Mutations

Abstract

Contents

1 Introduction

Fumarate hydratase (FH) is an essential enzyme in the citric acid cycle, playing a critical role in cellular metabolism by catalyzing the conversion of fumarate to malate. Mutations in the FH gene have been linked to severe health conditions, including Hereditary Leiomyomatosis and Renal Cell Cancer (HLRCC) [1]. The ability to accurately predict the pathogenicity of FH mutations is crucial in early diagnosis and targeted treatment, as these mutations can significantly impact enzymatic function and metabolic pathways. However, distinguishing pathogenic mutations from benign ones remains a complex challenge due to the diverse structural and functional effects that mutations can introduce.

Current computational approaches to mutation classification primarily rely on two broad methodologies: statistical meta-predictors and mechanistic biophysical models. Statistical models such as REVEL and BayesDel aggregate predictions from multiple computational tools to rank mutations based on their likelihood of pathogenicity. These models offer high accuracy and efficiency by leveraging extensive genetic datasets, evolutionary conservation scores, and machine learning algorithms [2]. However, they do not provide mechanistic insights into why a given mutation affects the FH protein structure or function. In contrast, mechanistic biophysical models aim to explain mutation effects at the molecular level by analyzing changes in protein folding, stability, and interactions. These models provide valuable mechanistic insights but currently underperform in terms of predictive accuracy when compared to statistical approaches [3].

The goal of this project is to compare mechanistic biophysical models with statistical predictors to assess their respective strengths and limitations in FH mutation classification. By refining cutoff thresholds, introducing new predictive features, and analyzing data using ROC (Receiver Operating Characteristic) curves, we aim to improve the reliability of mechanistic models while maintaining a high level of predictive accuracy. The study will help bridge the gap between explainability and predictive performance, ultimately contributing to better mutation classification in clinical and research settings.

1.1 Theory

Fumarate hydratase is a key enzyme involved in cellular respiration, particularly within the tricarboxylic acid (TCA) cycle, also known as the citric acid cycle. This cycle is fundamental to energy production, as it facilitates the oxidation of metabolic intermediates to generate ATP. The FH enzyme catalyzes the hydration of fumarate into malate, a reaction essential for the proper functioning of the cycle. Inherited or somatic mutations in FH can lead to enzyme dysfunction, resulting in an accumulation of fumarate, which has been implicated in tumorigenesis and metabolic disorders [4]. Specifically, loss-of-function mutations in FH have been associated with Hereditary Leiomyomatosis and Renal Cell Cancer (HLRCC), a highly aggressive form of cancer. Given the serious implications of FH mutations, accurate classification of pathogenic variants is crucial for early diagnosis and clinical intervention.

There are two primary computational approaches to predicting mutation effects: mechanistic biophysical models and statistical machine-learning-based models. Mechanistic biophysical models attempt to describe how mutations affect protein structure and function using physics-based simulations. These models leverage tools such as molecular dynamics simulations, which assess structural stability and conformational changes, and energy-based predictors like Rosetta and FoldX, which estimate how mutations influence protein folding and binding interactions [5]. Structural bioinformatics techniques are also employed to analyze how mutations impact critical functional regions within the protein, such as active sites and dimerization interfaces. While mechanistic models provide a detailed molecular understanding, they often require high computational resources and struggle to match the predictive performance of data-driven statistical models [6].

Statistical models such as REVEL and BayesDel take a different approach by aggregating data from multiple in silico prediction tools to classify mutations based on their likelihood of pathogenicity. These meta-predictors incorporate population genetics data, evolutionary conservation scores, and machine learning techniques to assign risk scores to mutations. REVEL, for example, integrates outputs from 13 individual predictors, including SIFT, PolyPhen-2, and MutationTaster, to provide a comprehensive pathogenicity score [2]. BayesDel, on the other hand, uses Bayesian inference to refine predictions by considering prior probabilities of pathogenicity. While these models have been shown to outperform many other computational tools, they do not provide mechanistic reasoning for their predictions [3].

Recent studies have highlighted the growing importance of integrating multiple computational approaches to enhance the accuracy of variant classification. Garcia et al. emphasize the need for combining statistical predictors with biophysical simulations to improve the interpretability of in silico mutation analysis [7]. Similarly, Aminian et al. present a case study of a patient with multiple malignancies linked to an FH germline mutation, further demonstrating the significance of accurate computational prediction in clinical settings [8]. These findings reinforce the necessity of refining mechanistic models for better diagnostic applications.

2 Methods

2.1 Data and repositories

All data used in this study, along with the code, can be found at: `https://github.com/zcapbic/Group-4-FH`. This data in turn was taken from a previous study on fumarate hydratase from a publicly available repository here: `https://github.com/shorthouse-mrc/Fumarate_Hydratase`. The FH mutation database in that repository was downloaded from the Leiden Open Variation Database (LOVD).

2.2 Programming

All programs used in completing this project were made in Python.

The two datasets containing the mechanistic metrics and statistical scores were combined on their mutations. We chose to classify mutations as deleterious only if all three statistical scores exceeded the threshold value. Using this categorization, we split the combined database into benign and deleterious frames for our binary classification to take place. All of this was implemented using the Pandas package in Python.

We then tested the fit of this binary classification using receiving operator characteristic (ROC) curves, utilizing the sklearn.metrics Python package. To further understand our results, we chose to plot the distributions of our mechanistic metrics for both data frames. We used Seaborn, a Python package, to plot the kernel densities of these two metrics. The similarity between the distributions was quantified using the Kolmogorov-Smirnov (KS) statistic from the scipy.stats module.

To explain the predictive gap, we chose to look at individual residues. Using Matplotlib we plot the distributions of mechanistic metrics throughout the FH gene for the deleterious database. To obtain the Venn diagrams, we used Python's Venn package; we then found the residues corresponding to this Venn diagram using sets and unions.

3 REVEL and BayesDel comparison

3.1 Correlation analysis between REVEL and BayesDel scores

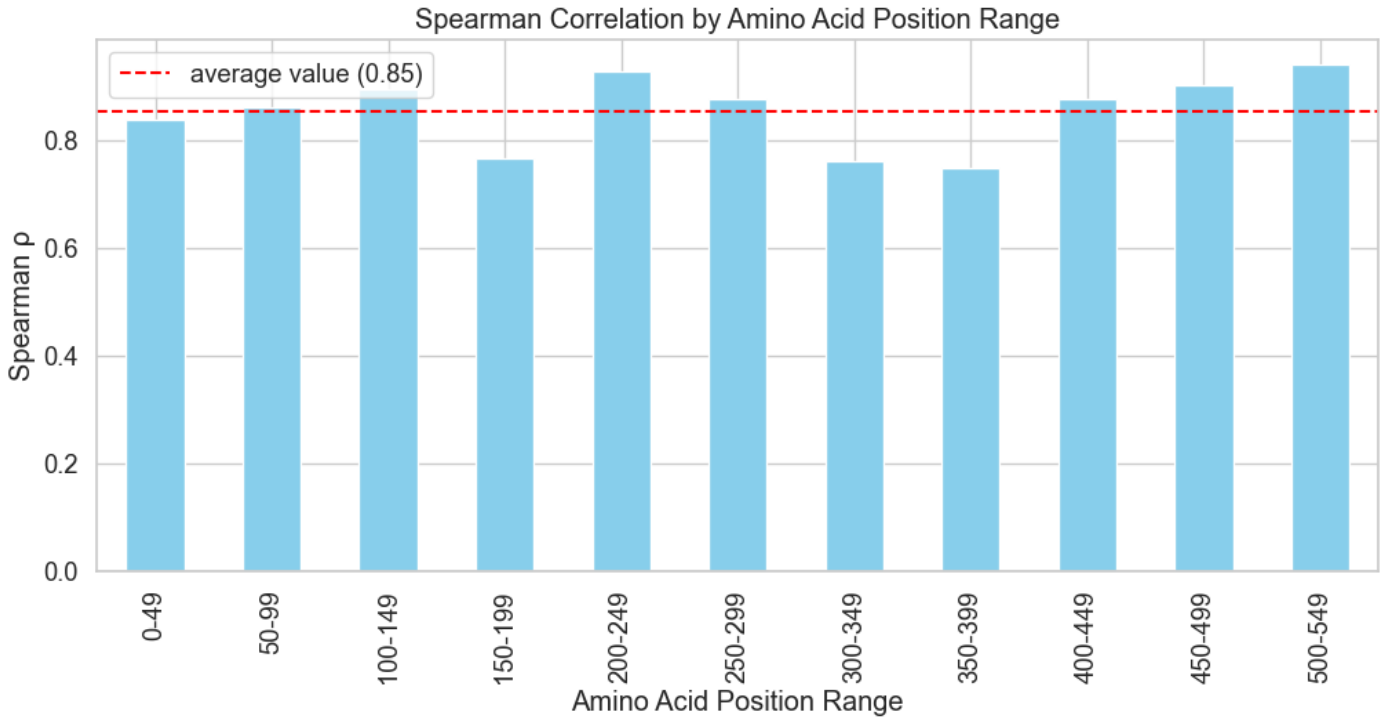

Figure 1: Enter Caption

3.2 Determination of the best classification threshold

4 Investigating mechanistic models

Of the 3048 mutations in the database, 2025 were classified as loss of function (66.4%); this is in great agreement with previous studies which suggested that this ratio was closer to 66%, showing a triumph of our classification.

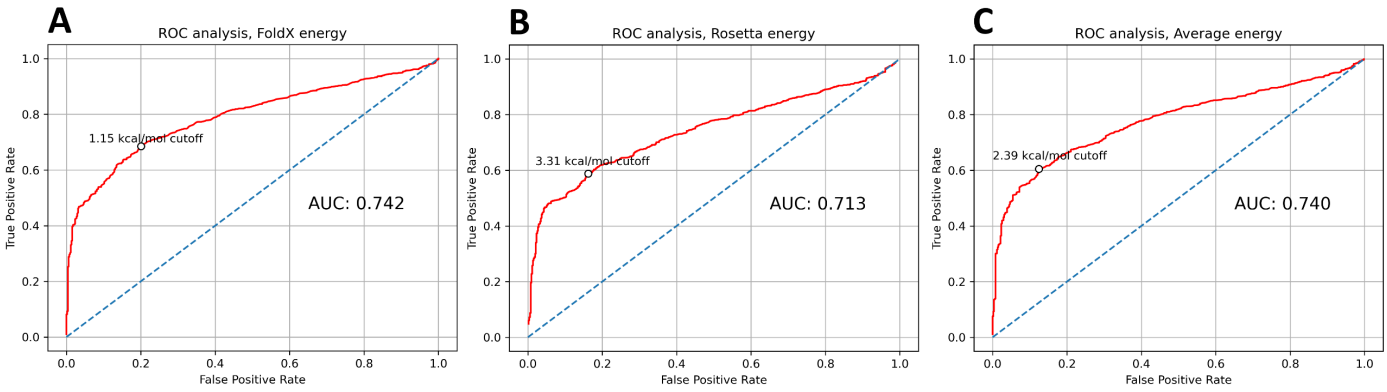

Figure 2: ROC curves for different energy metric performances in predicting LOF mutations. **(A)** FoldX energies performance. The best threshold was found to be 1.15 kcal/mol with a TPR of 68.5% and TNR of 79.1%. **(B)** Rosetta energies performance. The best threshold was found to be 3.31 kcal/mol with a TPR of 58.8% and TNR of 84.2%. **(C)** Average energy performance. The best threshold was found to be 2.39 kcal/mol with a TPR of 60.4% and TNR of 87.5%. AUC - area underneath curve.

Rosetta energies are computed using more sophisticated calculations than those of FoldX, taking into account molecular dynamics (reference). Despite this, we found that FoldX outperforms Rosetta in mutation classification; correctly identifying 69% of LOF (loss of function) mutations, in comparison to just 59% for

Rosetta, as shown in **Figure 2**. Because of its low AUC and TPR values, Rosetta was deemed to be a poor predictor of pathogenicity as a folding energy measure.

Another success of our ROC analysis is shown in the determination of the best threshold for the average energy metric, again in **Figure 2c**. This threshold was found to be 2.39 kcal/mol; in line with previous studies which use thresholds in range of 2.5 to 3 kcal/mol for this same metric. This triumph also validates our method and computed cutoffs for the different metrics in **Figure 2** and **Figure 3**.

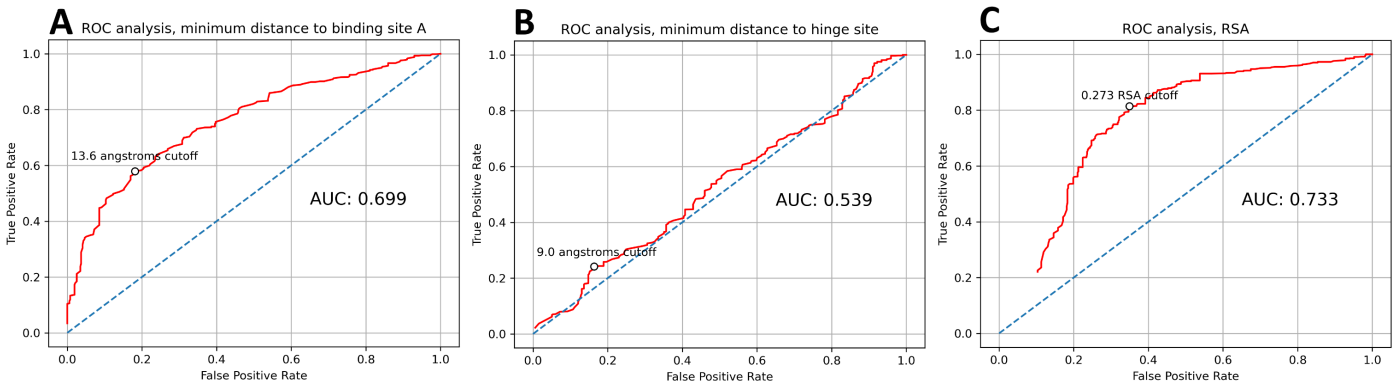

Figure 3: ROC curves for other mechanistic metric performances in predicting LOF mutations. **(A)** Distance to binding site performance. The best threshold was found to be 13.6 angstroms with a TPR of 57.9% and TNR of 81.9%. **(B)** Distance to hinge site performance. The best threshold was found to be 9.0 angstroms with a TPR of 24.2% and TNR of 83.3%. **(C)** Relative solvent accessibility (RSA) performance. The best threshold was found to be 0.273 with a TPR of 81.4% and TNR of 65.1%. AUC - area underneath curve.

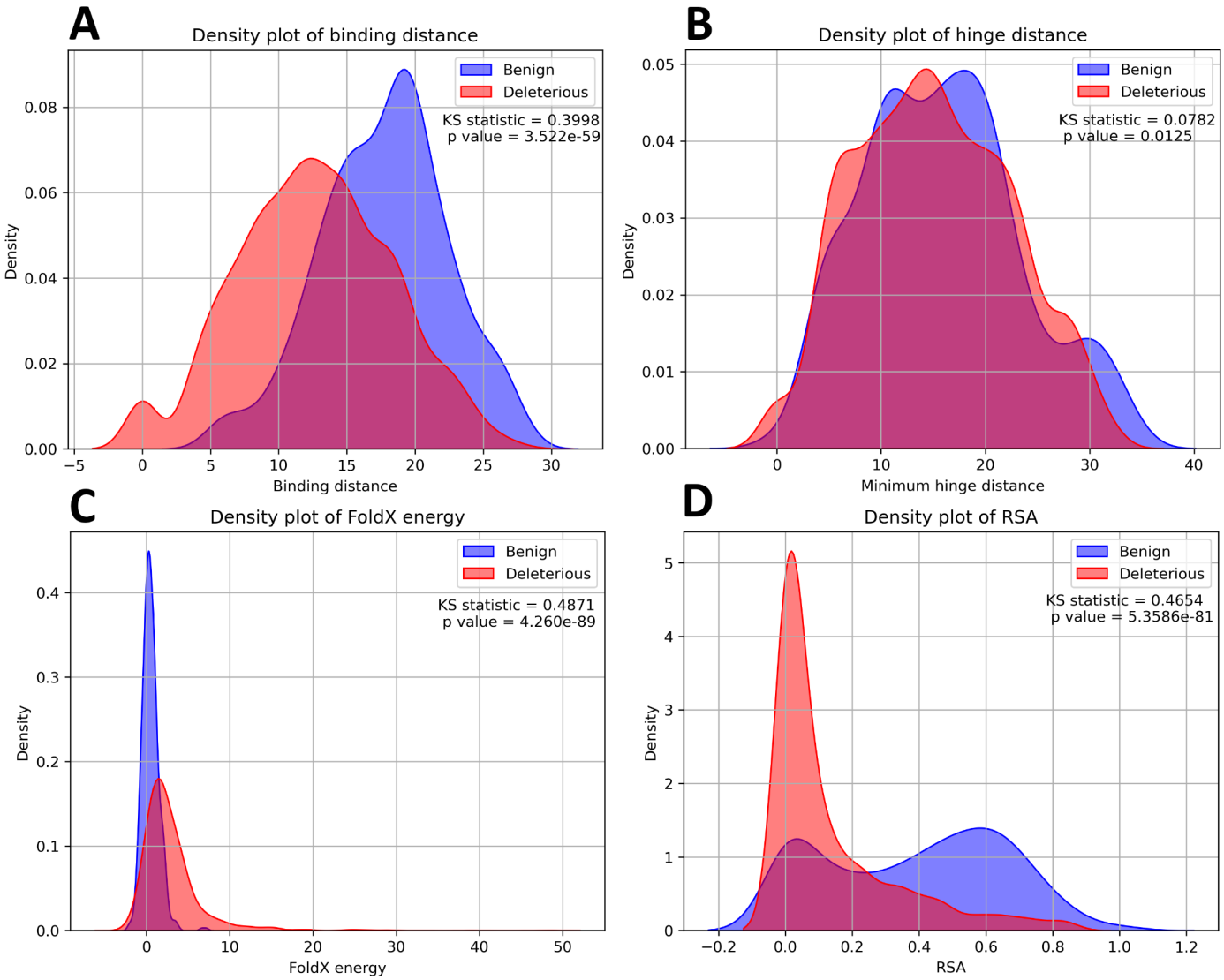

Figure 4: Kernel Density Estimate (KDE) plots of four metrics: **(A)** binding site distance, **(B)** hinge distance, **(C)** FoldX energy and **(D)** relative solvent accessibility (RSA), for their benign and deleterious distributions. Similar distributions have a lower Kolmogorov-Smirnov (KS) statistic.

Interestingly, it was observed that the hinge distance is actually a very poor predictor of pathogenicity, giving a low AUC of 0.539 as indicated in **Figure 3b**. Intuitively, one might expect that mutations close to a hinge will contribute to a loss of function. However, a kernel density plot of hinge distance as shown in **Figure 4b** shows that the distributions of the minimum hinge distance between benign and deleterious mutations are very similar, with a KS statistic of 0.0782. This is in contrast to the other three mechanistic models in **Figures 4a**, **c** and **d** which all show a decent separation between the benign and pathogenic distributions.

In particular, the relative solvent accessibility (RSA) was observed as a very strong predictor of pathogenicity, achieving a TPR of 81.4% for a threshold of 0.273. This cutoff is well in line with the ≈ 0.2 used in many other papers; validating our methods once again.

Despite our data matching the information used in different studies, these predictors are still failing to capture about 20 to 30% of deleterious mutations. As was observed in **Figure 1**, different residues have different performances in correlations between REVEL and BayesDel; this leads to the possibility that the gap between mechanistic models and statistical scores can be explained by this. **Figure 5** was one of the plots created in attempting this challenge. Whilst all folding energies below 1.15 kcal/mol would be classed as benign, **Figure 5a** shows that 33 out of these 160 residues (20.6%) marked as benign have an average folding energy above this cutoff. A similar story is seen in **Figure 5b**, where 94 out of 415 residues (22.7%) are labeled as deleterious, despite having a folding energy below 1.15 kcal/mol. These insights can combine to suggest that whilst the folding energy is a decent means of predicting pathogenicity, it does not capture the full picture; other mechanistic metrics must be considered if we are attempting to predict deleteriousness.

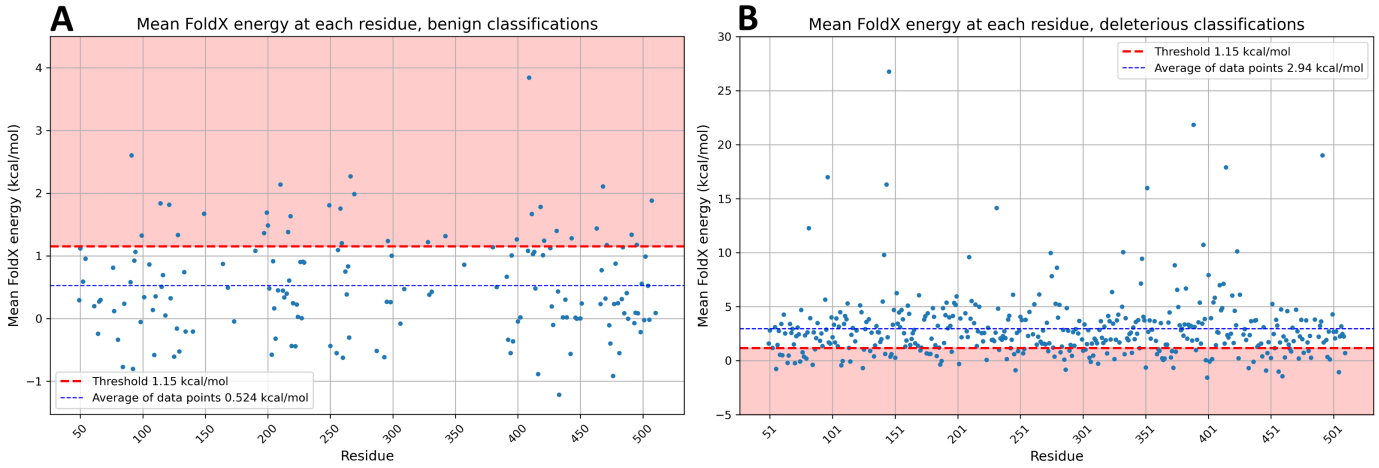

Figure 5: Mean FoldX energy of each residue for **(A)** benign mutations and **(B)** deleterious mutations. Energies above the cutoff line in **(A)** are falsely labeled as deleterious; energies below the cutoff line in **(B)** are falsely labeled as benign.

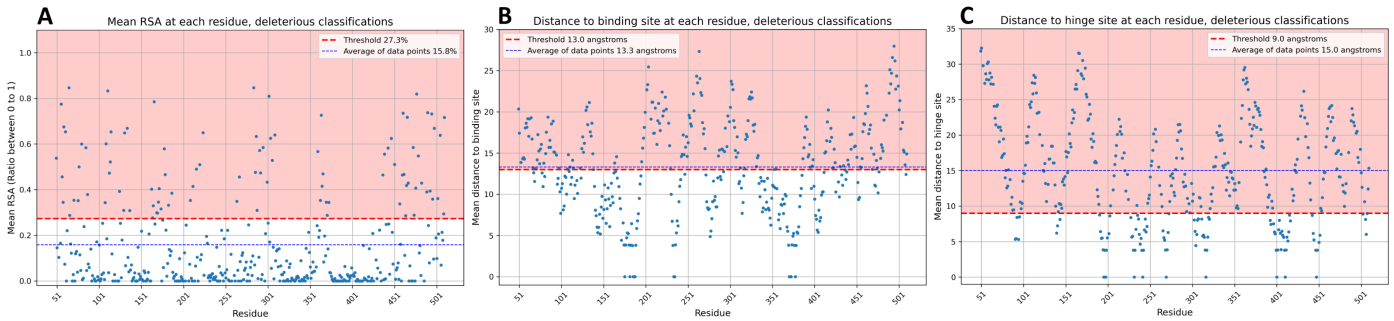

Figure 6: Mechanistic metrics **(A)** relative solvent accessibility (RSA), **(B)** distance to binding site and **(C)** distance to hinge, plot for each residue. Metrics above the cutoff line, in the red region, are falsely labeled as benign. Notice that in **(B)** and **(C)** the average metric calculated across residues lie in the false positive region.

To investigate this further, we extended this procedure out to the RSA, hinge distance and binding distance metrics, as shown in **Figure 6**. It was found that the average distances to important sites exceeded the threshold for the deleterious dataset, reinforcing our belief that these two metrics are poor predictors of pathogenicity.

Furthermore, we again observed that 94 out of 415 residues (22.7%) were falsely labeled as deleterious by the RSA predictor; further analysis showed that not all 94 residues were the same as the ones incorrectly predicted by FoldX. In particular, exactly half of these residues were different.

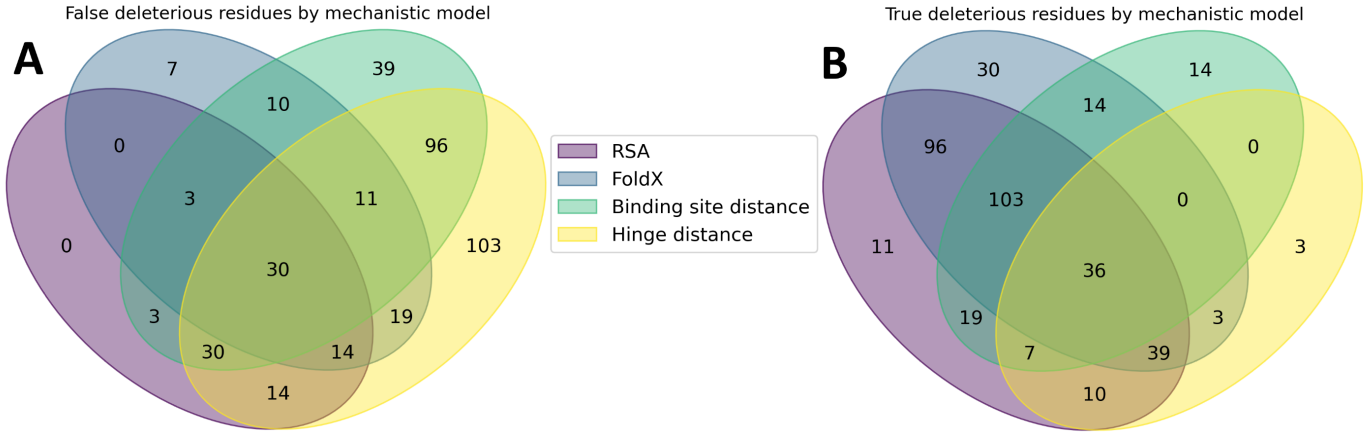

Figure 7: Four-way Venn diagram for the mechanistic metrics used as predictors of pathogenicity. **(A)** represents the deleterious mutations which have not been flagged by mechanistic models, **(B)** represents those deleterious mutations which have been picked up by the mechanistic models.

We decided to explore further in this vein, seeing **Figure 7** as an important tool to identify the residues which are more difficult to classify. Although the hinge distance was found to be a subpar method of classifying

pathogenic mutations, we observed in **Figure 7b** that there are 3 residues which can only be explained by this mechanistic metric, showing that the hinge distance still can pick up some deleterious mutations.

It was also seen in **Figure 7a** that there were LOF mutations on 30 residues which all mechanistic models fail to pick up; these residues are listed in **Table 1**, found in the appendix. Our belief is that mutations on these residues involve another mechanistic model which was not looked at in this study; charge distribution is one such example.

We attempted to exclude these 30 residues from our database and reran our ROC analysis, producing the results shown in **Figure 8**. Removing these 30 residues corresponded to removing 190 out of 3048 mutations (6.23%). In contrast to **Figures 2** and **3**, we saw that FoldX and RSA were capturing approximately 4% more deleterious mutations, at the expensive of falsely classifying about 2% to 4% more benign mutations as pathogenic. In general, we observed a slight increase in the AUC values for these two metrics. However, for removing 6% of the "difficult to classify" mutations, these mechanistic models have not had too much of an improved performance; once again suggesting that different mechanistic metrics are required to explain some deleterious mutations.

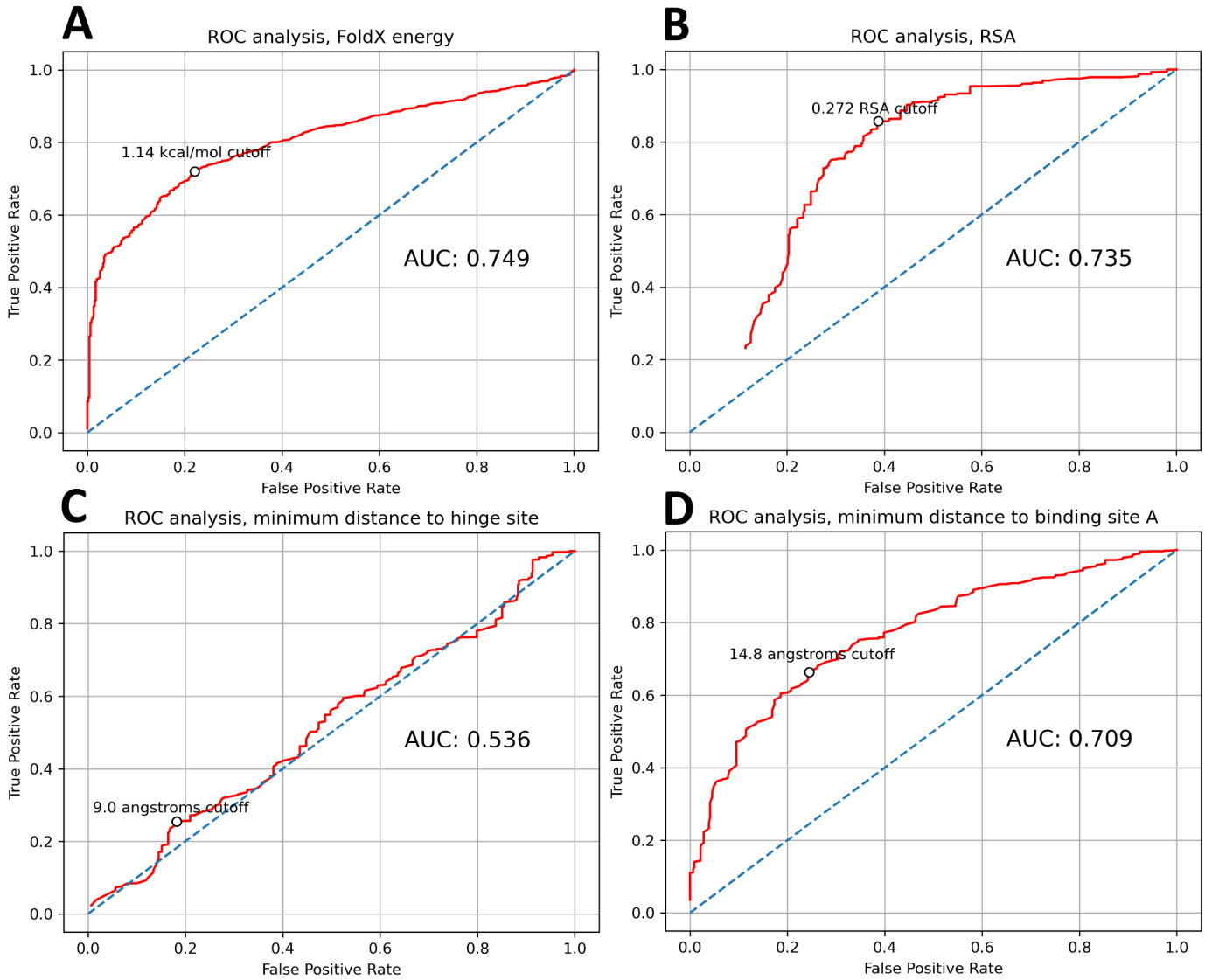

Figure 8: ROC curves for mechanistic metric performances in predicting LOF mutations, after having removed the 30 so-called difficult to classify residues. **(A)** FoldX energy performance. The best threshold was found to be 1.14 kcal/mol with a TPR of 72.0% and TNR of 77.9%. **(B)** Relative solvent accessibility (RSA) performance. The best threshold was found to be 0.272 with a TPR of 85.8% and TNR of 61.3%. **(C)** Distance to hinge site performance. The best threshold was found to be 9.0 angstroms with a TPR of 25.5% and TNR of 82.1%. **(D)** Distance to binding site performance. The best threshold was found to be 14.8 angstroms with a TPR of 66.4% and TNR of 75.5%.

5 Conclusion

A Supplementary information

Overlap category	Residues
4 overlaps: 36 residues	96, 140, 141, 144, 191, 192, 193, 194, 195, 231, 232, 233, 236, 237, 238, 239, 241, 242, 272, 312, 314, 315, 316, 317, 397, 401, 403, 405, 406, 407, 408, 409, 410, 411, 412, 452
3 overlaps: 149 residues	68, 69, 79, 81, 91, 92, 93, 95, 97, 99, 100, 101, 102, 103, 104, 105, 106, 107, 110, 112, 116, 117, 119, 120, 124, 139, 142, 145, 146, 149, 150, 152, 153, 154, 155, 156, 157, 158, 159, 160, 161, 162, 167, 175, 177, 180, 181, 182, 185, 186, 189, 190, 197, 198, 199, 200, 203, 228, 229, 230, 234, 235, 240, 243, 244, 245, 247, 266, 268, 269, 270, 271, 273, 274, 275, 276, 277, 279, 280, 281, 282, 285, 289, 293, 305, 308, 311, 318, 319, 321, 335, 336, 337, 338, 339, 341, 342, 343, 344, 345, 347, 348, 349, 350, 352, 354, 356, 358, 359, 367, 368, 369, 374, 376, 377, 378, 379, 380, 381, 382, 383, 384, 385, 386, 393, 394, 398, 400, 402, 404, 413, 414, 415, 416, 417, 426, 429, 430, 434, 438, 445, 449, 453, 454, 472, 503, 506, 507, 508
2 overlaps: 142 residues	51, 53, 55, 60, 62, 63, 67, 70, 72, 73, 74, 78, 82, 83, 86, 88, 89, 98, 113, 123, 127, 128, 132, 135, 136, 138, 143, 147, 148, 151, 163, 165, 169, 170, 172, 174, 176, 183, 184, 188, 196, 202, 204, 207, 208, 210, 211, 214, 215, 217, 218, 221, 222, 225, 246, 248, 250, 251, 252, 255, 257, 258, 261, 262, 265, 267, 278, 283, 284, 290, 294, 297, 298, 303, 304, 307, 310, 313, 320, 322, 323, 324, 325, 327, 328, 329, 330, 332, 333, 340, 346, 351, 353, 355, 361, 362, 363, 370, 372, 375, 387, 388, 389, 390, 392, 395, 396, 399, 419, 420, 421, 422, 423, 424, 431, 433, 435, 439, 440, 441, 443, 444, 448, 455, 456, 458, 459, 463, 468, 469, 471, 475, 476, 480, 482, 486, 491, 492, 497, 500, 501, 509
No overlaps: 58 residues	50, 57, 58, 75, 76, 111, 114, 126, 130, 137, 164, 166, 171, 178, 179, 187, 209, 212, 216, 220, 249, 253, 254, 264, 286, 288, 291, 295, 300, 301, 302, 306, 326, 334, 360, 364, 365, 366, 371, 373, 391, 425, 437, 447, 451, 461, 464, 465, 473, 474, 477, 479, 481, 489, 490, 493, 495, 502
Not in any set: 30 residues	56, 59, 61, 65, 66, 71, 77, 80, 85, 87, 108, 109, 115, 125, 131, 134, 224, 287, 292, 309, 457, 460, 462, 466, 467, 483, 496, 498, 505, 510

Table 1: Residues which appear in the overlaps for the four-way Venn diagram in **Figure 7b**. Note that because **Figure 7a** are the false statistics, the residue table for that one is reversed.

B Meeting Minutes

B.1 Meeting 1

Tuesday 21 January 2025
Start time: 13:00
End time: 14:06
Duration: 66 minutes
Apologies: Elina

B.2 Meeting 2

Tuesday 4 February 2025
Apologies: Nobody

B.3 Meeting 3

Tuesday 18 February 2025
Apologies: Nobody

B.4 Meeting 4

Tuesday 4 March 2025

Start time: 15:30
End time: 16:04
Duration: 34 minutes
Apologies: Nobody

B.5 Meeting 5

Tuesday 18 March 2025