UNAAGI: ATOM-LEVEL DIFFUSION FOR GENERATING NON-CANONICAL AMINO ACID SUBSTITUTIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Proposing beneficial amino acid substitutions, whether for mutational effect prediction or protein engineering, remains a central challenge in structural biology. Recent inverse folding models, trained to reconstruct sequences from structure, have had considerable impact in identifying functional mutations. However, current approaches are constrained to designing sequences composed exclusively of natural amino acids (NAAs). The larger set of non-canonical amino acids (NCAAs), which offer greater chemical diversity, and are frequently used in in-vivo protein engineering, remain largely inaccessible for current variant effect prediction methods.

To address this gap, we introduce **UNAAGI**, a diffusion-based generative model that reconstructs residue identities from atomic-level structure using an E(3)-equivariant framework. By modeling side chains in full atomic detail rather than as discrete tokens, UNAAGI enables the exploration of both canonical and non-canonical amino acid substitutions within a unified generative paradigm. We evaluate our method on experimentally benchmarked mutation effect datasets and demonstrate that it achieves substantially improved performance on NCAA substitutions compared to the current state-of-the-art. Furthermore, our results suggest a shared methodological foundation between protein engineering and structure-based drug design, opening the door for a unified training framework across these domains.

1 INTRODUCTION

Proteins are linear chains of amino acids that fold into specific three-dimensional structures to perform a wide range of biological functions. Chemically, an amino acid consists of a central α-carbon bonded to a hydrogen atom, an amino group (NH_2), a carboxyl group (COOH), and a variable side chain denoted as R. While the R group can take on diverse chemical forms, only 20 distinct side-chain structures are commonly found in naturally occurring proteins. These are referred to as the genetically encoded, canonical, or Natural Amino Acids (NAAs) (Branden & Tooze, 1991). Several studies suggest that this set of 20 NAAs has been evolutionarily optimized to achieve functional completeness and broad chemical diversity (Ilardo et al., 2015; Doig, 2017; Ilardo et al., 2019). Nevertheless, Non-Canonical Amino Acids (NCAAs) offer side-chain chemistries and functionalities absent from the standard set, such as enhanced metal coordination, tailored electrostatics, or novel catalytic properties. Incorporating NCAAs enables protein engineers to expand the functional repertoire of proteins beyond evolutionary constraints, unlocking new possibilities in synthetic biology and therapeutics (Link et al., 2003; Chin, 2017; Rogers et al., 2018).

In recent years, machine learning – particularly deep learning – has achieved remarkable success in protein research, advancing tasks such as structure prediction, design, and mutational effect estimation (Jumper et al., 2021; Dauparas et al., 2022; Abramson et al., 2024). Variant effect prediction models are often trained in an unsupervised fashion, where amino acid propensities in a protein are predicted conditioned on a sequence (Riesselman et al., 2018) or structural (Dauparas et al., 2022) context. Such procedure generally model amino acids using a discrete distribution, typically limited to the 20 naturally occurring amino acids, based on the propensities observed in our vast databases on evolutionary data. While such models have shown impressive zero-shot predictive performance for various protein properties, the discrete modeling prohibits generalization beyond the naturally occurring amino acids, representing a fundamental limitation in the current modeling paradigm.

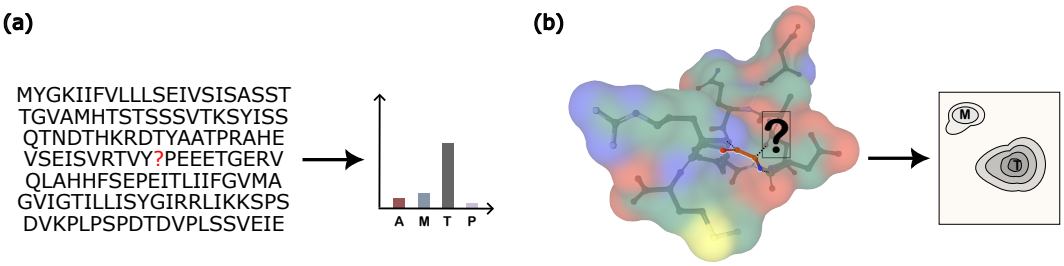

Figure 1: An illustration of the motivation behind UNAAGI. **(a)** Previous models predicted amino acid identities from a fixed vocabulary, limiting sampling to natural amino acids. **(b)** UNAAGI directly samples side chains at the atom level, enabling generation of more diverse structures, including non-canonical amino acids (NCAAs).

To address this gap, we propose a novel generative framework for residue-level sequence design based on atom-level side-chain generation via equivariant molecular diffusion. We term our approach **U**ncanonical **N**ovel **A**mino **A**cid **G**enerative **I**nference (**UNAAGI**). By modeling the atomic coordinates of side chains directly, UNAAGI implicitly infers residue identity and is capable of proposing plausible non-canonical substitutions; a visual illustration of our motivation is shown in Figure 1. We evaluate our model by comparing its generative likelihoods to experimentally measured mutational effect (Rogers et al., 2018; Notin et al., 2023), and find that the performance on non-canonical amino acids is improved considerably over the current state-of-the-art. Importantly, our model maintains consistent performance across both canonical and non-canonical residues.

Specifically, our contributions are as follows:

- We introduce a novel E(3)-equivariant diffusion framework for atom-wise side-chain generation, offering a new approach to residue identity inference.
- We investigate the zero-shot performance of UNAAGI on variant effect prediction. While our model does not fully match the state-of-the-art for naturally occurring amino acids, we are the first method to demonstrate meaningful levels of correlation for non-canonical substitutions.
- We discuss the broader implications of our approach, suggesting that it may bridge structure-based drug design (SBDD) and protein engineering, as both fields can benefit from shared modeling tools such as energy-based design.

In the following sections, we formalize the generation task at the atom level, describe our E(3)-equivariant diffusion framework UNAAGI, evaluate it on mutational benchmarks, and discuss its implications for protein design and synthetic biology.

2 Related Work

Our approach has connections to prior work in molecular diffusion, variant effect prediction and peptide design.

2.1 Molecular Diffusion

Hoogeboom et al. (2022) introduced the first framework for E(3)-equivariant diffusion, generating molecular conformations in 3D space via a diffusion process. However, their approach suffered from low generative quality and often produced molecules with disconnected atoms. Subsequent works, including Peng et al. (2023) and Vignac et al. (2023), proposed refinements that incorporated inductive biases from molecular bonding structures, leading to improved sampling quality.

Le et al. (2024) provided a comparative study of several molecular diffusion frameworks, analyzing the impact of different modeling choices on sample quality. Their work identified combinations of design choices that optimize performance for molecular generative tasks.

Beyond unconditional molecular generation, a critical application is the conditional generation of compounds that bind to specific protein targets—known as Structure-Based Drug Design (SBDD). Schneuing et al. (2024) extended the framework of Hoogeboom et al. (2022) to conditional ligand generation in 3D space, while Guan et al. (2023) proposed similar models for target-aware ligand sampling. More recently, Schneuing et al. (2025) advanced this line of research by introducing models capable of sampling ligands with variable atom counts and incorporating protein pocket side-chain conformations during generation.

2.2 Mutational Effect Prediction

Mutational effect prediction is the task of predicting the impact of amino acid substitutions in a protein on a particular property of interest (e.g. stability, affinity or function). Multiple state-of-the-art protein models—whether sequence-based, structure-based, or hybrids of the two—have demonstrated sensitivity to the mutational landscape of proteins.

Statistical models of multiple sequence alignments remain a strong baseline when many sequences are available for a given protein family. A notable example is the DeepSequence VAE model (Riesselman et al., 2018). Language-based models such as Meier et al. (2021) and Lin et al. (2023) obviate the need for alignments, leveraging conservation and coevolutionary signals learned from massive protein sequence datasets. While sequence-based model can predict mutational effect in many contexts, their performance declines when evolutionary signals are sparse, such as for shallow MSAs, antibodies, or de novo designed peptides.

Structure-based approaches replace the sequence context of a mutation with a structural context. Early models considered the structural environment around one amino acid at a time (Boomsma & Frellsen, 2017; Torng & Altman, 2017), which was later generalized to entire sequences in inverse-folding models (Ingraham et al., 2019; Dauparas et al., 2022; Hsu et al., 2022). Although primarily designed for structure reconstruction, these models also show utility in mutational effect prediction Frellsen et al. (2025). Our approach mirrors the early structure-based models by considering only single amino acids at a time, with the crucial difference that UNAAGI extends mutational effect prediction to include non-canonical amino acids.

2.3 Modeling Non-canonical Amino Acid Substitutions

PepINVENT (Geylan et al., 2025) is a transformer-based language model trained on atomic-level peptide representations using CHUCKLES (Siani et al., 1994), which converts amino acid residue tokens into SMILES strings. This enables the model to generate amino acid SMILES for masked positions. The strategy allows for generalization to non-canonical amino acids, but the original study does not validate the sample weights with experimental data. We include this method as a baseline.

NCFlow (Lee & Kim, 2025) and RareFold (Li et al., 2025) are based on AlphaFold3-style architectures. NCFlow adopts the Flow Matching framework and evaluates mutational effect prediction on the same benchmark as ours, allowing for direct comparisons (see Section 4). RareFold introduces new tokens for NCAAs and focuses on predicting protein structures containing NCAA substitutions or designing peptides with NCAAs. Their study reports on successful experimentally validated designed peptides with NCAAs, but the model itself cannot predict affinity scores or mutational effect in silico, making direct comparison on NCAA mutational effect prediction infeasible.

3 UNAAGI

We introduce UNAAGI, a diffusion-based model that bridges molecular generation and protein mutational effect prediction by constructing amino acid side chains atom by atom, thereby extending the mutational landscape beyond the natural amino acid space. By learning from the chemistry of canonical residues, UNAAGI can generalize and interpolate to non-canonical amino acids. In what follows, we outline the general idea of equivariant graph diffusion, describe the model architecture, training setup, and evaluation protocol, and finally present the baselines used for comparison.

3.1 MULTI-MODAL DIFFUSION FOR MOLECULAR GRAPHS

Diffusion models have emerged as a powerful class of generative models (Sohl-Dickstein et al., 2015; Ho et al., 2020; Song & Ermon, 2020; Song et al., 2021a;b). In the discrete-time setting, these models learn to reverse a Markovian forward process that gradually corrupts data with noise until the samples resemble a tractable prior distribution (typically Gaussian). The generative model is trained to approximate the reverse process, which recovers data via a parameterized denoising chain.

Originally developed for image generation, diffusion models have since been extended to molecular and protein structure generation (Hoogeboom et al., 2022). Let $\mathbf{x}_0 \sim q(\mathbf{x}_0)$ denote a sample from the data distribution. The forward process is defined as

$$q(\mathbf{x}_{1:T} \mid \mathbf{x}_0) = \prod_{t=1}^{T} q(\mathbf{x}_t \mid \mathbf{x}_{t-1}), \tag{1}$$

which incrementally adds noise over T steps until $\mathbf{x}_T$ becomes indistinguishable from Gaussian noise. The reverse process is then modeled as

$$p_\theta(\mathbf{x}_{0:T}) = p(\mathbf{x}_T) \prod_{t=1}^{T} p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t), \tag{2}$$

which gradually denoises $\mathbf{x}_T$ back to a clean sample.

Molecular data is inherently *multi-modal*: atomic coordinates are continuous, while atom types, bond types, and other chemical features are discrete. To handle this, diffusion over discrete variables has been explored through categorical transition kernels—discrete analogues of Gaussian noise in the forward process (Hoogeboom et al., 2021; Austin et al., 2023). Recent molecular diffusion models combine continuous and categorical processes to jointly model the heterogeneous modalities of molecules (Peng et al., 2023; Vignac et al., 2023; Guan et al., 2023; Le et al., 2024).

The forward noise process for continuous and discrete modalities can be expressed as

$$\begin{aligned} q(\mathbf{x}_t \mid \mathbf{x}_0) &= \mathcal{N}\big(\mathbf{x}_t \mid \sqrt{\bar{\alpha}_t}\mathbf{x}_0,\ (1-\bar{\alpha}_t)\mathbf{I}\big)\,, \\ q(\mathbf{c}_t \mid \mathbf{c}_0) &= \mathcal{C}(\mathbf{c}_t \mid \bar{\alpha}_t \mathbf{c}_0 + (1-\bar{\alpha}_t)\tilde{\mathbf{c}})\,, \end{aligned} \tag{3}$$

where $\bar{\alpha}_t = \prod_{k=1}^{t}(1-\beta_k) \in (0,1)$ is the cumulative noise schedule. For discrete features, $\tilde{\mathbf{c}}$ denotes the categorical prior (e.g., uniform or empirical distribution). Following D3PM (Austin et al., 2023), we adopt a *mask diffusion* scheme, which gradually converts labels into an absorbing state during the forward process.

In our setting, we perturb both atomic coordinates $\mathbf{X}$ and atom-wise categorical features $\mathbf{H}$. The features in $\mathbf{H}$ include atom types, formal charges, hybridization states, aromaticity, ring membership, and atomic degree. Each modality is diffused independently using either Gaussian or categorical noise. This chemically-aware perturbation strategy is inspired by Le et al. (2024), who showed that injecting chemically meaningful noise improves sample quality and stability.

The training objective follows the standard ELBO on the data log-likelihood:

$$\log p(\mathbf{x}_0) \geq \mathcal{L}_0 + \mathcal{L}_{\text{prior}} + \sum_{t=1}^{T-1} \mathcal{L}_t, \tag{4}$$

where $\mathcal{L}_0 = \log p(\mathbf{x}_0 \mid \mathbf{x}_1)$ is the reconstruction term and $\mathcal{L}_{\text{prior}} = -\mathrm{KL}(q(\mathbf{x}_T) \| p(\mathbf{x}_T))$ matches the prior. In practice, these terms are often omitted. The main learning signal comes from minimizing the per-timestep KL divergence:

$$\mathcal{L}_t = \mathrm{KL}[q(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{x}_0) \parallel p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t)]\,, \tag{5}$$

which has closed-form solutions under both Gaussian and categorical noise. This is typically optimized by predicting either the clean data $\hat{\mathbf{x}}_0$, the noise $\boldsymbol{\epsilon}$, or categorical logits, depending on the parameterization (Ho et al., 2020; Austin et al., 2023). We adopt the $\hat{\mathbf{x}}_0$ - parameterization to directly predict the clean data. The loss function is mean squared error for atomic coordinates and cross-entropy for other categorical features.

3.2 E(3)-EQUIVARIANT GRAPH NEURAL NETWORK

Molecular structures are inherently three-dimensional and symmetric under Euclidean transformations such as rotation and translation. To respect these symmetries, we incorporate *equivariance* as an inductive bias.

Formally, a function $f : \mathcal{X} \to \mathcal{Y}$ is equivariant to a group G if

$$f(g \cdot x) = g \cdot f(x), \quad \forall g \in G.$$

For molecular data, the relevant group is $\mathrm{E}(3)$, which includes all 3D rotations and translations. Specifically, a denoising function on atomic coordinates $\mathbf{X} \in \mathbb{R}^{N\times 3}$ should satisfy

$$f(\mathbf{X}Q + \mathbf{t}) = f(\mathbf{X})Q + \mathbf{t},$$

for any orthogonal transformation $Q \in \mathrm{O}(3)$ and translation $\mathbf{t} \in \mathbb{R}^3$.

We adopt an E(3)-equivariant Graph Neural Network, closely following the EQGAT-diff model of Le et al. (2024), to learn the score function over atom-level graphs. Each node represents an atom and carries scalar features h_i (e.g., atom type, charge) and vector features $\mathbf{v}_i \in \mathbb{R}^3$, while edges encode bond types and interatomic distances.

Message passing is performed on a fully connected graph with attention-based aggregation. The directional information between atoms i and j is encoded as a unit vector

$$\mathbf{x}_{ji,n} = \frac{\mathbf{x}_j - \mathbf{x}_i}{\|\mathbf{x}_j - \mathbf{x}_i\|},$$

which is used to construct equivariant vector features. Messages are computed via MLPs over concatenated scalar and edge features, then split into components to update scalar, vector, and positional embeddings in an equivariant manner. This design ensures that all updates preserve E(3)-equivariance and enables joint modeling of geometry, connectivity, and chemistry during denoising.

3.3 GRAPH TOPOLOGY AND VIRTUAL NODE PADDING

Unlike traditional Structure-Based Drug Design (SBDD), where generative models sample ligands conditioned on a protein pocket, our task focuses on reconstructing amino acid side chains. A key distinction is that each generated side chain is covalently bonded to a fixed protein backbone, imposing local structural constraints around the masked region. To account for this, we reformulate the graph construction. Specifically, the positions and types of backbone atoms for the residue under reconstruction are fixed, anchoring generation to the known structural context.

A further challenge is unified sampling across natural amino acids (NAA) and non-canonical amino acids (NCAA), which vary in atom counts. In standard SBDD models, the atom count must be fixed before diffusion, since the graph topology must be predetermined. We address this using a virtual node strategy inspired by DrugFlow (Schneuing et al., 2025). Unlike empirical sampling schemes (Hoogeboom et al., 2022; Schneuing et al., 2024; Guan et al., 2023) or size-estimation networks (Igashov et al., 2024), this approach is fully end-to-end. This allows sampling from a smoother distribution over amino acid identities across side chains of varying sizes.

Virtual nodes are assigned a special atom type `NOATOM` and are disconnected from all other atoms, with edges labeled `NOBOND`. During training, a random number of virtual nodes $n_{\text{virt}} \sim U(0, N_{\max})$ are added to the side chain. Following Schneuing et al. (2025), these nodes are initialized at the side-chain center of mass. The total number of nodes is capped by a predefined upper bound $N_{\max}$, while the effective side-chain size becomes variable and learnable. At sampling time, the model denoises a graph with $N_{\max}$ nodes. Nodes denoised to the virtual type are removed post hoc, allowing variable-sized side chains to be generated within a unified diffusion framework.

4 EXPERIMENTS

We now describe the experimental procedure for UNAAGI, including training details, evaluation methods, and comparisons against baselines.

4.1 Dataset Composition and Preprocessing

We use a dataset of 1,000 protein structures submitted to the Protein Data Bank (PDB) for training. For each structure, we extract every residue along with its local environment, defined as surrounding residues whose center of mass lies within a 10 Å radius of the target residue. To ensure independence between training and evaluation, we verified that none of these PDB structures overlap with the benchmark assays.

In addition to canonical amino acids, we include non-canonical amino acids (NCAAs) to enhance the model's capacity to generalize to novel chemical environments and to prevent overfitting. For NCAAs, we use datasets from Ilardo et al. (2019) and SwissSidechain (Gfeller et al., 2012), although these consist of NCAAs without associated proteomic context. To capture diverse physical interactions, we also augment the training set with protein–ligand complexes from PDBBind (Wang et al., 2005). Graph construction differs slightly between sources: - For independent NCAAs, the four backbone atoms (C, C, O, N) are fixed during training, and only side-chain atoms are diffused. - For PDBBind, we fix pocket atoms and apply noise to all ligand atoms, following the procedure of SBDD models in Le et al. (2024).

4.2 Evaluation on Deep Mutational Scanning (DMS)

During sampling, UNAAGI generates diverse side-chain conformations given a fixed environment. We interpret the sampling frequencies of side-chain identities and conformations as a proxy for the probability density learned by the diffusion model.

To evaluate predictive power, we compare this learned distribution against experimental Deep Mutational Scanning (DMS) data. Specifically, for each site in the DMS benchmark, we perform 100 sampling iterations and record frequencies of each sampled amino acid. While larger sample sizes would provide smoother frequency estimates, we adopt 100 iterations as a trade-off between statistical stability and computational feasibility. Sampled side chains are matched to known natural or non-canonical amino acids using graph isomorphism (Weisfeiler & Lehman, 1968).

From the sampling frequencies, we estimate the likelihood of each mutant or wild-type residue. We then compute the negative log-likelihood (NLL) for each sampled amino acid, calculating the differential log-likelihood as:

$$\Delta \log \mathcal{L} = -\log P(\text{mutant}) + \log P(\text{wild-type}),$$

where $P(\text{mutant})$ and $P(\text{wild-type})$ denote the estimated probabilities of mutant and wild-type residues, respectively. This log-likelihood difference is correlated with experimental $\Delta\Delta G$ values from the DMS benchmark.

As a sanity check on the ability to recover the natural amino acids, we selected assays with fewer than 100 residues from Notin et al. (2023)[1] . The full list of assays and results is provided in Supplementary A.1.

To evaluate performance on NCAAs, we use peptide–protein complexes from Rogers et al. (2018), specifically PDB ID 5LY1 (JMJD2A/KDM4A bound to a macrocyclic peptide CP2) and PDB ID 2ROC (Mcl-1 bound to PUMA). This dataset includes DMS measurements of $\Delta\Delta G$ for substitutions across 20 canonical and 20 non-canonical amino acids. For fairness, we restrict evaluation to substitutions where both the mutant and wild-type residues appear in the sampled data.

4.3 Results

4.3.1 ProteinGym

We evaluate predictive performance by computing the Spearman correlation between likelihood differences and ground-truth experimental values, following the reporting standard of ProteinGym, with more detailed per-assay results in Supplementary A.1.

[1]Note that the computational cost of UNAAGI scales linearly with protein size (since only local structural environments are considered). Scaling to larger proteins is therefore not fundamentally problematic, but given a limited computational budget, we prioritized protein diversity over protein length in this experiment.

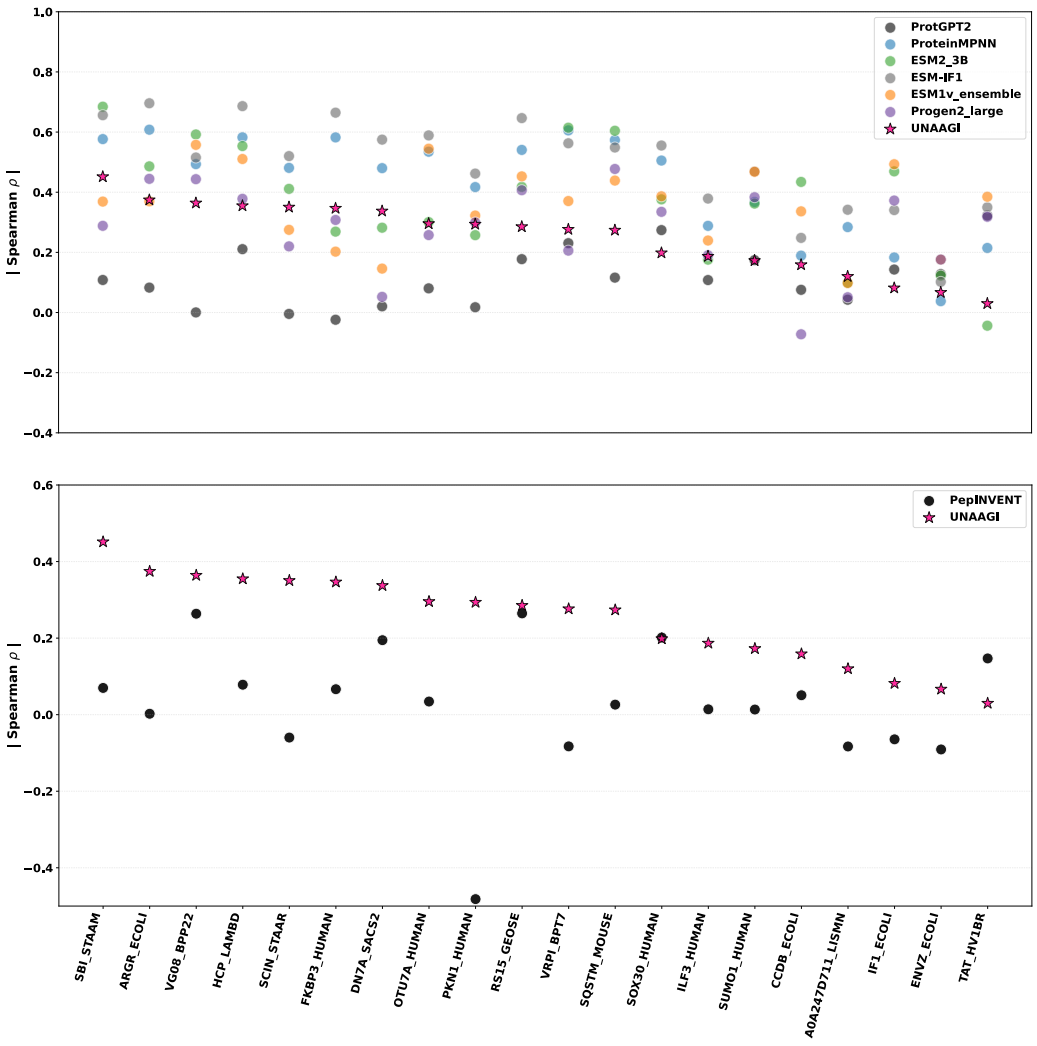

Figure 2: **Top**: Results of UNAAGI on 20 ProteinGym assays, evaluated using Spearman correlation against experimental DMS measurements. **Down**: Results of UNAAGI on ProteinGym, comparing to PepINVENT.

For reference, we also evaluate PepINVENT as a baseline in this setting , although it was not originally tested on DMS benchmarks. Following the repository guidelines, we sample by masking one residue at a time and generating candidate substitutions. This is repeated for all positions in each assay. For peptide–protein complexes, we concatenate peptide and protein sequences as input to the language model. Sampled canonical SMILES are mapped back to amino acids if they are in the benchmark to estimate frequencies.

While UNAAGI does not achieve state-of-the-art performance on any single assay, it still exhibits substantial correlation with experimental mutational effect across most assays (Figure 2). Importantly, UNAAGI is distinct from all other baselines in Figure 2, as it generates residues in an atom-wise manner rather than sampling from a fixed vocabulary. The experiment is somewhat adversarial, since we for this dataset know a priori that we are restricted to the 20 amino acids. It is encouraging that the models produces reasonable performance despite the more expressive output distribution.

For a fair comparison, we evaluate UNAAGI against its closest methodological baseline, PepINVENT, which represents sequences and samples residues atom by atom. The comparison, in terms of Spearman correlation on ProteinGym, is also shown in Figure 2. UNAAGI consistently outper-

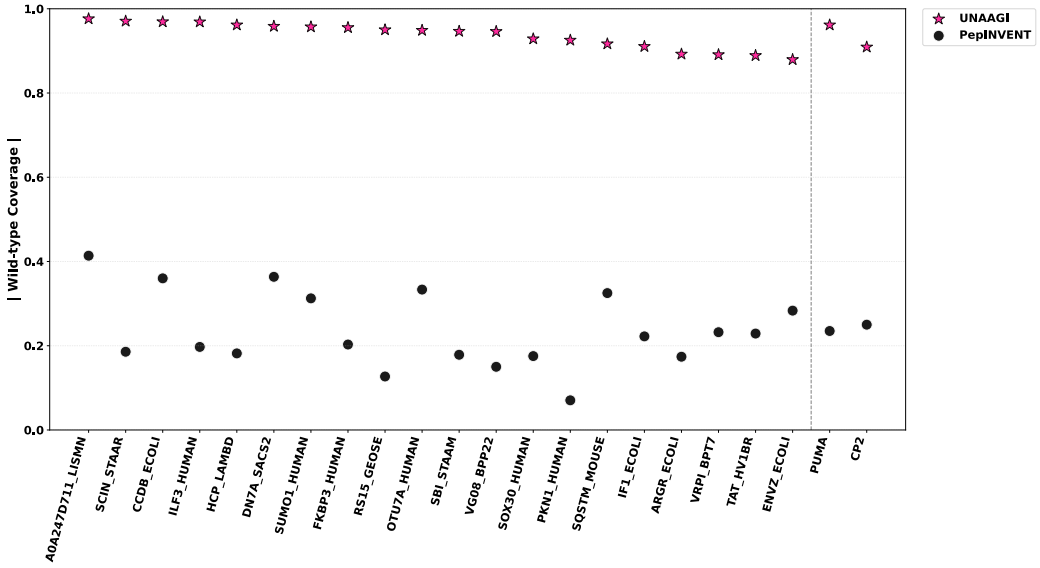

Figure 3: Wild-type coverage rate of UNAAGI and PepINVENT across benchmark assays.

forms PepINVENT on nearly all assays, with the exception of TAT_HV1BR. Notably, PepINVENT shows little to no correlation on most assays, whereas UNAAGI achieves substantial predictive performance.

Since both models operate in an atom-wise sampling regime, they must demonstrate sufficient capacity to reconstruct wild-type residues. To quantify this, we define the *wild-type coverage rate*—the frequency with which the wild-type residue is successfully sampled across all positions in an assay. Results are shown in Figure 3, where UNAAGI achieves consistently high coverage across sites, while PepINVENT rarely recovers the wild-type residue. This limitation restricts PepINVENT to evaluation on only a small subset of benchmark positions, and even under this reduced setting, its performance remains below that of UNAAGI, further emphasizing UNAAGI's advantage.

4.4 DMS FOR NCAA SUBSTITUTIONS

We further evaluate UNAAGI on a DMS benchmark containing substitutions to non-canonical amino acids (NCAAs). For comparison, we include both PepINVENT and NCFlow, as NCFlow was evaluated on the same benchmark in its original work.

As shown in Figure 4, UNAAGI achieves consistent performance across both canonical and non-canonical substitution benchmarks. The correlations observed on NCAA substitutions are comparable to those obtained on canonical amino acids, demonstrating that UNAAGI can generalize beyond the natural amino acid space. In contrast, PepINVENT shows limited signal on NCAA substitutions, and NCFlow performs poorly across the NCAA benchmark, with no or negative correlations regardless of the affinity prediction module used. These results indicate that UNAAGI is the first diffusion-based approach to provide reliable predictive power on NCAA mutational effect benchmarks.

5 CONCLUSION

In this paper, we introduced UNAAGI, a molecular diffusion model that generates amino acid side chains in an atom-wise manner and can sample across both canonical and non-canonical amino acids. We evaluated UNAAGI on Deep Mutational Scanning (DMS) benchmarks and found that it achieves meaningful correlations in variant effect prediction. Crucially, this correlation extends to benchmarks containing NCAA substitutions, making UNAAGI the first machine learning method to provide measurable signal on this problem.

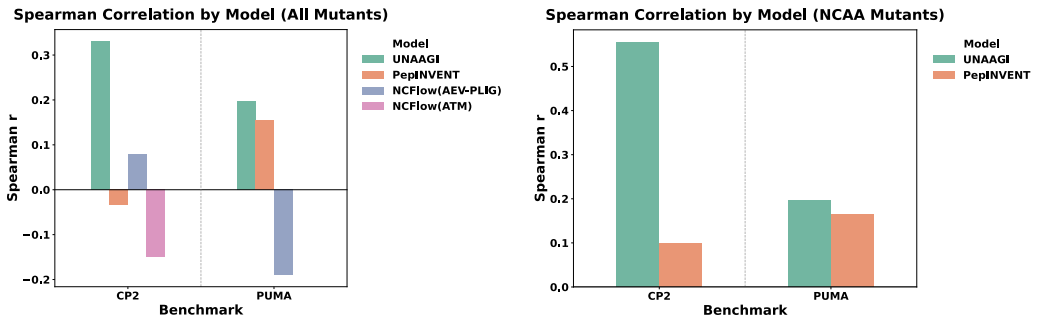

Figure 4: Results of UNAAGI on NCAA DMS benchmarks. NCFlow (ATM) was reported in the original work to fail entirely on the PUMA assay. **Left**: Spearman correlations compared to PepINVENT and NCFlow variants with different affinity modules. **Right**: Performance on NCAA substitutions specifically.

Methodologically, UNAAGI follows the molecular diffusion paradigm developed for Structure-Based Drug Discovery (SBDD), learning the density of feasible chemical combinations in protein structures. Unlike protein language models (Lin et al., 2023), UNAAGI does not rely on coevolutionary information from protein sequences, yet still shows predictive ability on mutational effect tasks. This suggests a promising connection between SBDD and variant effect prediction, as both are governed by the same non-covalent interaction principles. Building on this insight, future work could integrate additional SBDD techniques—such as pharmacophore-based conditioning (Peng et al., 2025) or flexible side-chain modeling (Schneuing et al., 2025)—to further improve variant effect prediction.

At the same time, UNAAGI has important limitations. It remains far from solving the problem of NCAA variant effect prediction and cannot yet propose high-fitness NCAA substitutions efficiently. As shown in Supplementary A.1, although UNAAGI achieves substantial correlations on the NCAA mutants it successfully samples, coverage remains limited to a small subset of the 20 NCAAs in the benchmark. This limitation is largely driven by the scarcity of relevant NCAA data. Moreover, UNAAGI tends to interpolate between canonical-like structures, while sampling chemically distinct NCAAs remains an out-of-distribution challenge under the current setup.

We also observe substantial variability in predictive performance across different assays, a phenomenon shared by all methods. This highlights the intrinsic complexity of the variant effect prediction problem: no single model performs universally well, and outcomes depend strongly on model design, training strategy, and data availability.

Looking forward, UNAAGI opens several avenues for exploration. A natural next step is scaling the model, both in parameter count (beyond the current 3.6M) and in training data (from the 1,000 PDB structures used here toward the full database). Incorporating protein structures with NCAA substitutions—although sparse in the PDB—could further expand coverage of the NCAA chemical space. Finally, applying guidance or auxiliary correctors may enhance the generation of chemically diverse NCAAs and enable biasing toward desirable properties, such as synthetic accessibility.

REFERENCES

Josh Abramson, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, Lindsay Willmore, Andrew J. Ballard, Joshua Bambrick, Sebastian W. Bodenstein, David A. Evans, Chia-Chun Hung, Michael O'Neill, David Reiman, Kathryn Tunyasuvunakool, Zachary Wu, Akvilė Žemgulytė, Eirini Arvaniti, Charles Beattie, Ottavia Bertolli, Alex Bridgland, Alexey Cherepanov, Miles Congreve, Alexander I. Cowen-Rivers, Andrew Cowie, Michael Figurnov, Fabian B. Fuchs, Hannah Gladman, Rishub Jain, Yousuf A. Khan, Caroline M. R. Low, Kuba Perlin, Anna Potapenko, Pascal Savy, Sukhdeep Singh, Adrian Stecula, Ashok Thillaisundaram, Catherine Tong, Sergei Yakneen, Ellen D. Zhong, Michal Zielinski, Augustin Žídek, Victor Bapst, Pushmeet Kohli, Max Jaderberg, Demis Hassabis, and John M. Jumper.

Accurate structure prediction of biomolecular interactions with alphafold 3. *Nature*, 630(8016): 493–500, 2024. doi: 10.1038/s41586-024-07487-w. URL https://doi.org/10.1038/s41586-024-07487-w.

Jacob Austin, Daniel D. Johnson, Jonathan Ho, Daniel Tarlow, and Rianne van den Berg. Structured denoising diffusion models in discrete state-spaces, 2023. URL https://arxiv.org/abs/2107.03006.

Wouter Boomsma and Jes Frellsen. Spherical convolutions and their application in molecular modelling. *Advances in neural information processing systems*, 30, 2017.

Carl Branden and John Tooze. *Introduction to Protein Structure*. Garland Publishing, New York, 1991.

Jason W. Chin. Expanding and reprogramming the genetic code. *Nature*, 550:53–60, 2017. doi: 10.1038/nature24031. URL https://doi.org/10.1038/nature24031.

J. Dauparas, I. Anishchenko, N. Bennett, H. Bai, R. J. Ragotte, L. F. Milles, B. I. M. Wicky, A. Courbet, R. J. de Haas, N. Bethel, P. J. Y. Leung, T. F. Huddy, S. Pellock, D. Tischer, F. Chan, B. Koepnick, H. Nguyen, A. Kang, B. Sankaran, A. K. Bera, N. P. King, and D. Baker. Robust deep learning–based protein sequence design using proteinmpnn. *Science*, 378(6615):49–56, 2022. doi: 10.1126/science.add2187. URL https://www.science.org/doi/abs/10.1126/science.add2187.

Alan J. Doig. Frozen, but no accident – why the 20 standard amino acids were selected. *FEBS Journal*, 284(9):1296–1305, 2017. doi: 10.1111/febs.13982. URL https://doi.org/10.1111/febs.13982.

Jes Frellsen, Maher M Kassem, Tone Bengtsen, Lars Olsen, Kresten Lindorff-Larsen, Jesper Ferkinghoff-Borg, and Wouter Boomsma. Zero-shot protein stability prediction by inverse folding models: a free energy interpretation. *arXiv preprint arXiv:2506.05596*, 2025.

G. Geylan, J. P. Janet, A. Tibo, J. He, A. Patronov, M. Kabeshov, W. Czechtizky, F. David, O. Engkvist, and L. De Maria. Pepinvent: Generative peptide design beyond natural amino acids. *Chemical Science*, 16(20):8682–8696, Apr 2025. doi: 10.1039/d4sc07642g.

David Gfeller, Olivier Michielin, and Vincent Zoete. Expanding molecular modeling and design tools to non-natural sidechains. *Journal of Computational Chemistry*, 33(19):1525–1535, 2012. doi: 10.1002/jcc.22982.

Jiaqi Guan, Wesley Wei Qian, Xingang Peng, Yufeng Su, Jian Peng, and Jianzhu Ma. 3d equivariant diffusion for target-aware molecule generation and affinity prediction, 2023. URL https://arxiv.org/abs/2303.03543.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models, 2020. URL https://arxiv.org/abs/2006.11239.

Emiel Hoogeboom, Didrik Nielsen, Priyank Jaini, Patrick Forré, and Max Welling. Argmax flows and multinomial diffusion: Learning categorical distributions, 2021. URL https://arxiv.org/abs/2102.05379.

Emiel Hoogeboom, Victor Garcia Satorras, Clément Vignac, and Max Welling. Equivariant diffusion for molecule generation in 3d, 2022. URL https://arxiv.org/abs/2203.17003.

Chloe Hsu, Robert Verkuil, Jason Liu, Zeming Lin, Brian Hie, Tom Sercu, Adam Lerer, and Alexander Rives. Learning inverse folding from millions of predicted structures. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 8946–8970. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/hsu22a.html.

Ilia Igashov, Hannes Stärk, Clément Vignac, Arne Schneuing, Victor Garcia Satorras, Pascal Frossard, Max Welling, Michael Bronstein, and Bruno Correia. Equivariant 3d-conditional diffusion model for molecular linker design. *Nature Machine Intelligence*, 6(4):417–427, 2024. ISSN 2522-5839. doi: 10.1038/s42256-024-00815-9. URL https://doi.org/10.1038/s42256-024-00815-9.

Melissa Ilardo, Markus Meringer, Stephen Freeland, Bakhtiyor Rasulev, and H. James Cleaves. Extraordinarily adaptive properties of the genetically encoded amino acids. *Scientific Reports*, 5: 9414, 2015. doi: 10.1038/srep09414. URL https://doi.org/10.1038/srep09414.

Melissa Ilardo, Raghav Bose, Markus Meringer, Bakhtiyor Rasulev, Oleksandr Zhulyn, Ian Carrick, and H. James Cleaves. Adaptive properties of the genetically encoded amino acid alphabet are inherited from its subsets. *Scientific Reports*, 9:12468, 2019. doi: 10.1038/s41598-019-47574-x. URL https://doi.org/10.1038/s41598-019-47574-x.

John Ingraham, Vikas Garg, Regina Barzilay, and Tommi Jaakkola. Generative models for graph-based protein design. *Advances in neural information processing systems*, 32, 2019.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A. A. Kohl, Andrew J. Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W. Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, 2021. doi: 10.1038/s41586-021-03819-2. URL https://doi.org/10.1038/s41586-021-03819-2.

Tuan Le, Julian Cremer, Frank Noe, Djork-Arné Clevert, and Kristof T Schütt. Navigating the design space of equivariant diffusion-based generative models for de novo 3d molecule generation. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=kzGuiRXZrQ.

Jin Sub Lee and Philip M. Kim. Design of peptides with non-canonical amino acids using flow matching. *bioRxiv*, 2025. doi: 10.1101/2025.07.31.667780. URL https://www.biorxiv.org/content/early/2025/07/31/2025.07.31.667780.

Qiuzhen Li, Diandra Daumiller, and Patrick Bryant. Rarefold: Structure prediction and design of proteins with noncanonical amino acids. *bioRxiv*, 2025. doi: 10.1101/2025.05.19.654846. URL https://www.biorxiv.org/content/early/2025/05/23/2025.05.19.654846.

Zeming Lin, Haydar Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wei Lu, Tom Sercu, Sergio Candido, and Alexander Rives. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379:1123–1130, 2023. doi: 10.1126/science.ade2574.

A. J. Link, M. L. Mock, and D. A. Tirrell. Non-canonical amino acids in protein engineering. *Current Opinion in Biotechnology*, 14(6):603–609, 2003. doi: 10.1016/j.copbio.2003.10.011. URL https://doi.org/10.1016/j.copbio.2003.10.011.

Joshua Meier, Roshan Rao, Robert Verkuil, Jason Liu, Tom Sercu, and Alexander Rives. Language models enable zero-shot prediction of the effects of mutations on protein function. *bioRxiv*, 2021. doi: 10.1101/2021.07.09.450648. URL https://doi.org/10.1101/2021.07.09.450648.

Pascal Notin, Aaron Kollasch, Daniel Ritter, Lood van Niekerk, Steffanie Paul, Han Spinner, Nathan Rollins, Ada Shaw, Rose Orenbuch, Ruben Weitzman, Jonathan Frazer, Mafalda Dias, Dinko Franceschi, Yarin Gal, and Debora Marks. Proteingym: Large-scale benchmarks for protein fitness prediction and design. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 64331–64379. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/

cac723e5ff29f65e3fcbb0739ae91bee-Paper-Datasets_and_Benchmarks.pdf.

Jian Peng, Jun-Lin Yu, Zeng-Bao Yang, Yi-Ting Chen, Si-Qi Wei, Fan-Bo Meng, Yao-Geng Wang, Xiao-Tian Huang, and Guo-Bo Li. Pharmacophore-oriented 3d molecular generation toward efficient feature-customized drug discovery. *Nature Computational Science*, Aug 2025. ISSN 2662-8457. doi: 10.1038/s43588-025-00850-5. URL https://doi.org/10.1038/s43588-025-00850-5.

Xingang Peng, Jiaqi Guan, Qiang Liu, and Jianzhu Ma. Moldiff: Addressing the atom-bond inconsistency problem in 3d molecule diffusion generation, 2023. URL https://arxiv.org/abs/2305.07508.

Adam J Riesselman, John B Ingraham, and Debora S Marks. Deep generative models of genetic variation capture the effects of mutations. *Nature methods*, 15(10):816–822, 2018.

J. M. Rogers, T. Passioura, and H. Suga. Nonproteinogenic deep mutational scanning of linear and cyclic peptides. *Proceedings of the National Academy of Sciences of the United States of America*, 115(43):10959–10964, 2018. doi: 10.1073/pnas.1809901115. URL https://doi.org/10.1073/pnas.1809901115.

Arne Schneuing, Charles Harris, Yuanqi Du, Kieran Didi, Arian Jamasb, Ilia Igashov, Weitao Du, Carla Gomes, Tom L. Blundell, Pietro Liò, Max Welling, Michael Bronstein, and Bruno Correia. Structure-based drug design with equivariant diffusion models. *Nature Computational Science*, 4(12):899–909, 2024. ISSN 2662-8457. doi: 10.1038/s43588-024-00737-x. URL https://doi.org/10.1038/s43588-024-00737-x.

Arne Schneuing, Ilia Igashov, Adrian W. Dobbelstein, Thomas Castiglione, Michael M. Bronstein, and Bruno Correia. Multi-domain distribution learning for de novo drug design. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=g3VCIM94ke.

M. A. Siani, D. Weininger, and J. M. Blaney. CHUCKLES: a method for representing and searching peptide and peptoid sequences on both monomer and atomic levels. *Journal of Chemical Information and Computer Sciences*, 34(3):588–593, May 1994. ISSN 0095-2338. doi: 10.1021/ci00019a017.

Jascha Sohl-Dickstein, Eric A. Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics, 2015. URL https://arxiv.org/abs/1503.03585.

Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution, 2020. URL https://arxiv.org/abs/1907.05600.

Yang Song, Conor Durkan, Iain Murray, and Stefano Ermon. Maximum likelihood training of score-based diffusion models. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, 2021a. URL https://openreview.net/forum?id=AklttWFnxS9.

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations, 2021b. URL https://arxiv.org/abs/2011.13456.

Wen Torng and Russ B Altman. 3D deep convolutional neural networks for amino acid environment similarity analysis. *BMC Bioinformatics*, 18(1):302, 2017.

Clement Vignac, Nagham Osman, Laura Toni, and Pascal Frossard. Midi: Mixed graph and 3d denoising diffusion for molecule generation, 2023. URL https://arxiv.org/abs/2302.09048.

Renxiao Wang, Xueguang Fang, Ying Lu, Chuangye Yang, and Shaomeng Wang. The pdbbind database: methodologies and updates. *Journal of Medicinal Chemistry*, 48(12):4111–4119, 2005. doi: 10.1021/jm048957q.

Boris Weisfeiler and A. A. Lehman. A Reduction of a Graph to a Canonical Form and an Algebra Arising During This Reduction. *Nauchno-Technicheskaya Informatsia*, Ser. 2(N9):12–16, 1968.

A Appendix

A.1 Detailed results on DMS Benchmarks

In this section, we present the detailed performance of UNAAGI on the DMS benchmarks. Section A.1.1 reports results on ProteinGym, while Section A.1.2 reports results on DMS assays containing NCAAs.

A.1.1 Detailed results on ProteinGym

A0A247D711_LISMN

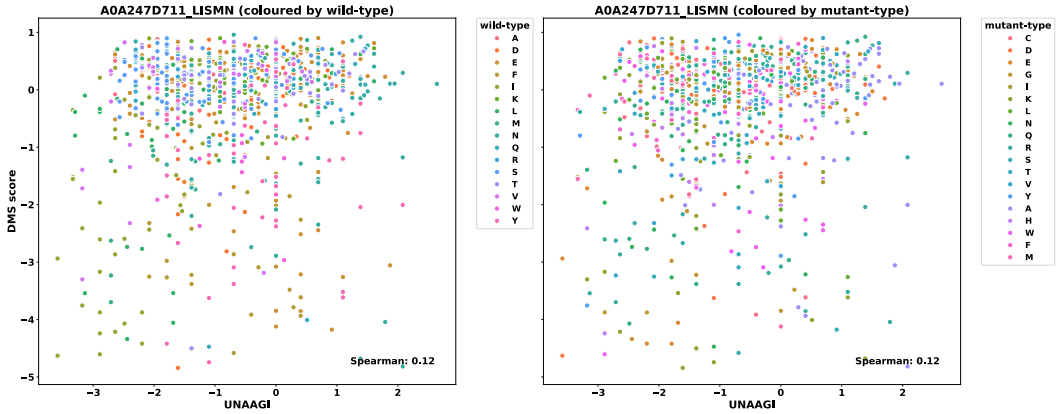

Figure 5: Scatterplot of UNAAGI vs. DMS score

ARGR_ECOLI

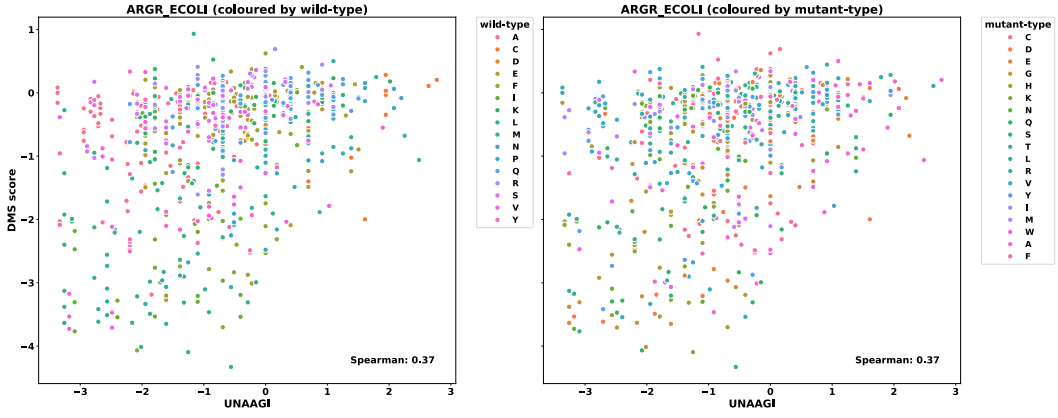

Figure 6: Scatterplot of UNAAGI vs. DMS score

CCDB_ECOLI

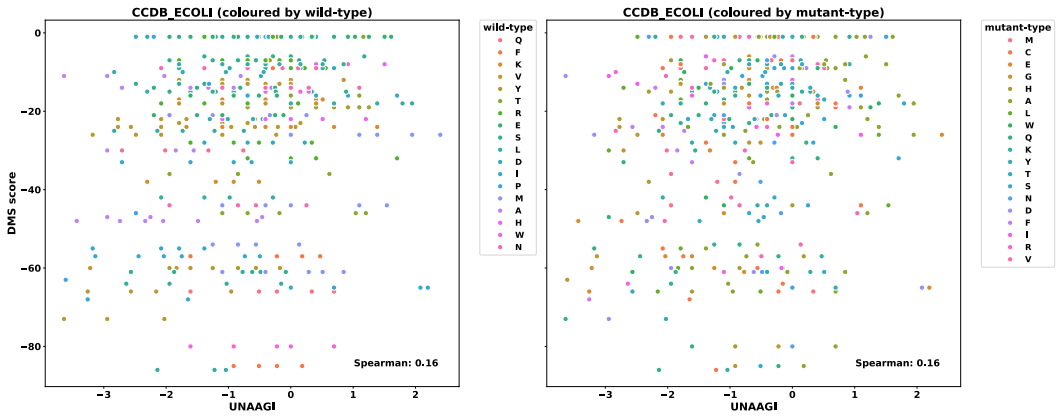

Figure 7: Scatterplot of UNAAGI vs. DMS score

DN7A_SACS2

Figure 8: Scatterplot of UNAAGI vs. DMS score

ENVZ_ECOLI

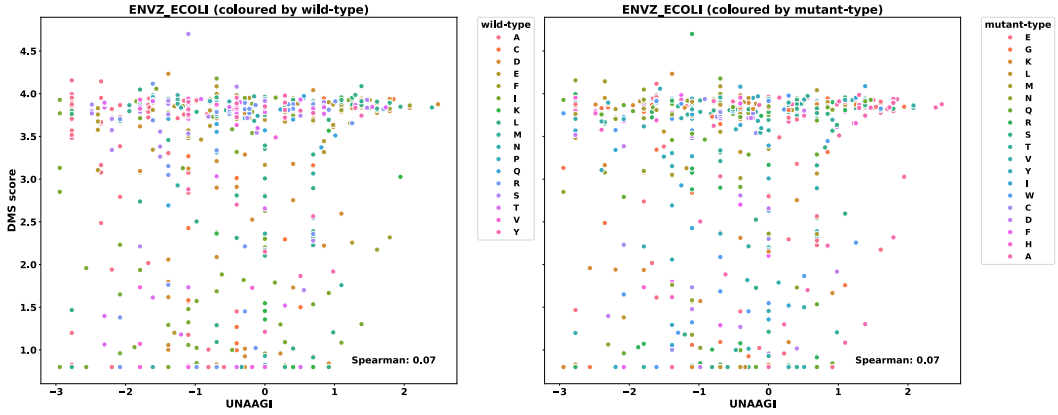

Figure 9: Scatterplot of UNAAGI vs. DMS score

FKBP3_HUMAN

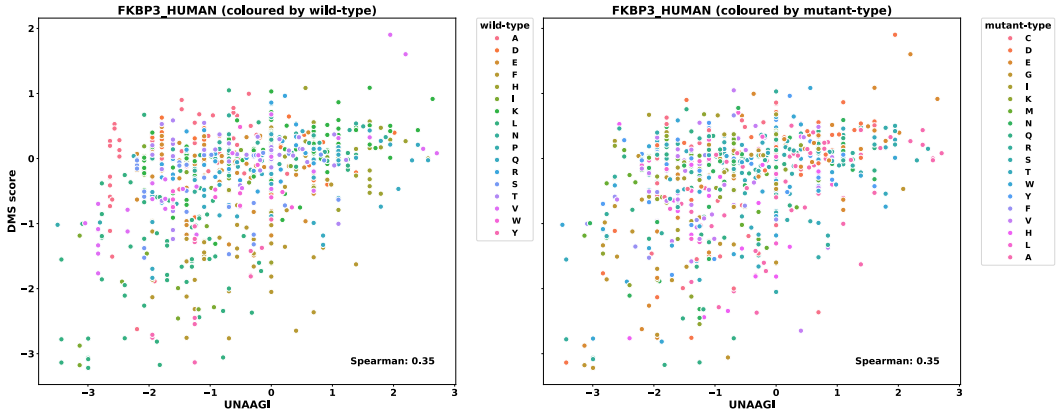

Figure 10: Scatterplot of UNAAGI vs. DMS score

HCP_LAMBD

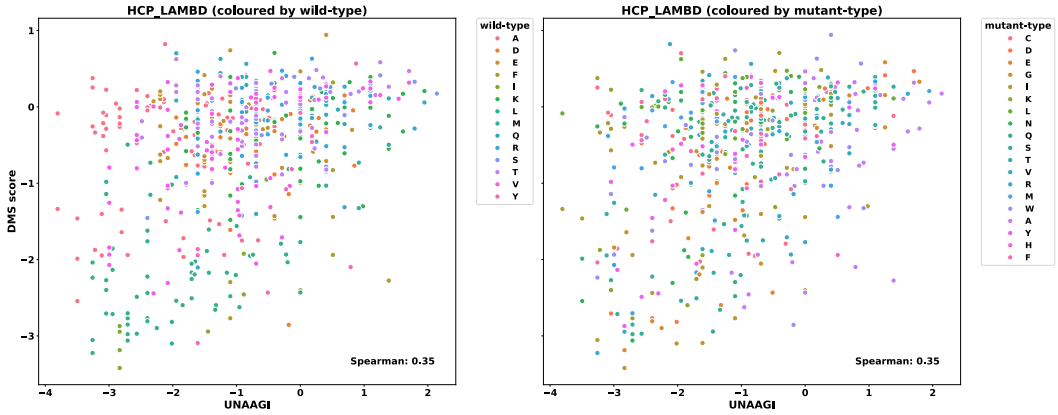

Figure 11: Scatterplot of UNAAGI vs. DMS score

IF1_ECOLI

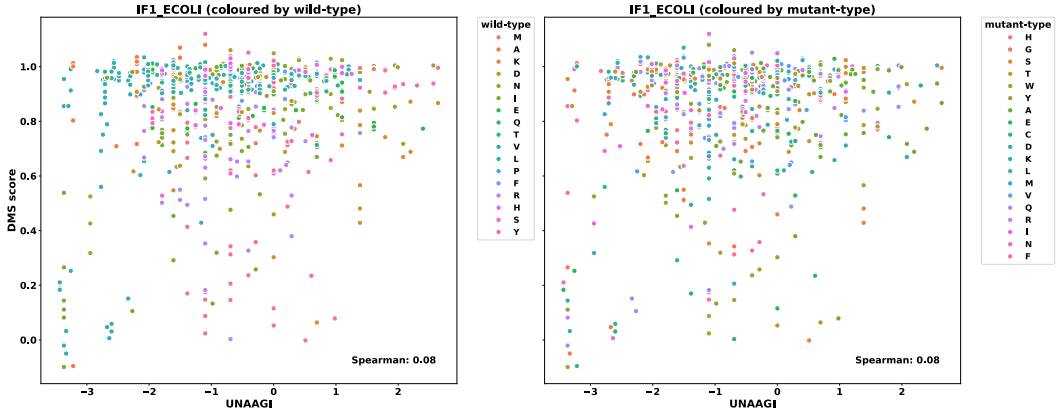

Figure 12: Scatterplot of UNAAGI vs. DMS score

ILF3_HUMAN

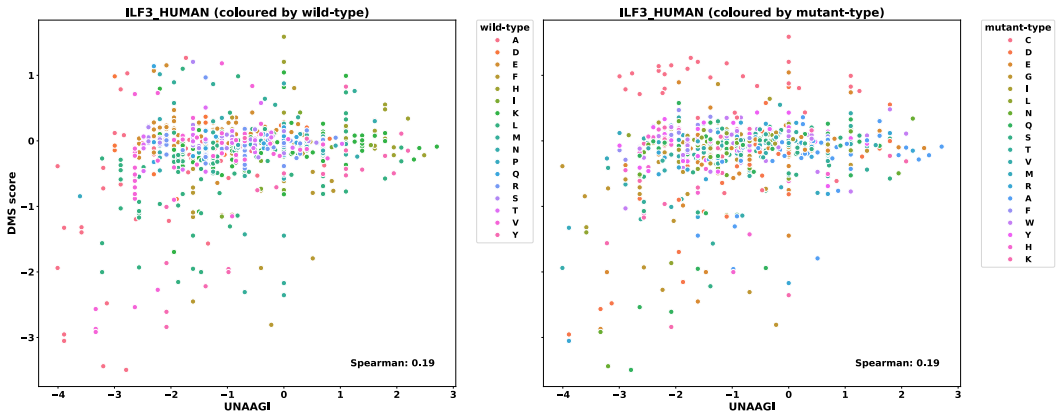

Figure 13: Scatterplot of UNAAGI vs. DMS score

OTU7A_HUMAN

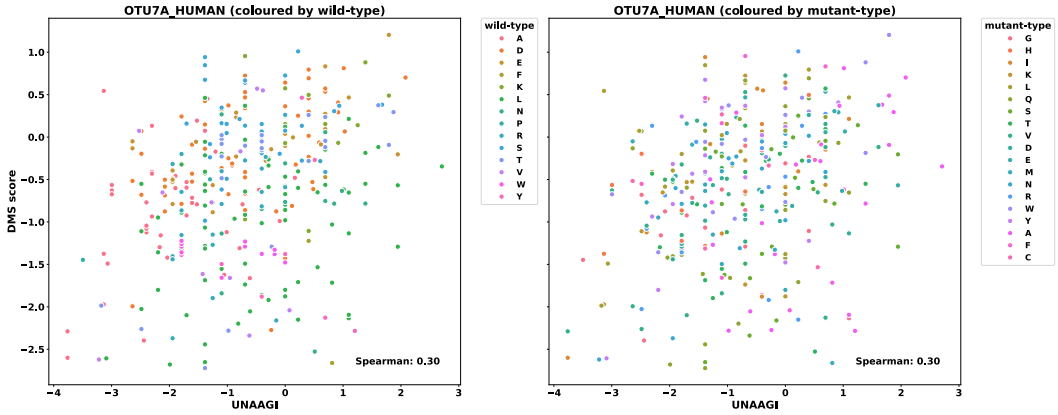

Figure 14: Scatterplot of UNAAGI vs. DMS score

PKN1_HUMAN

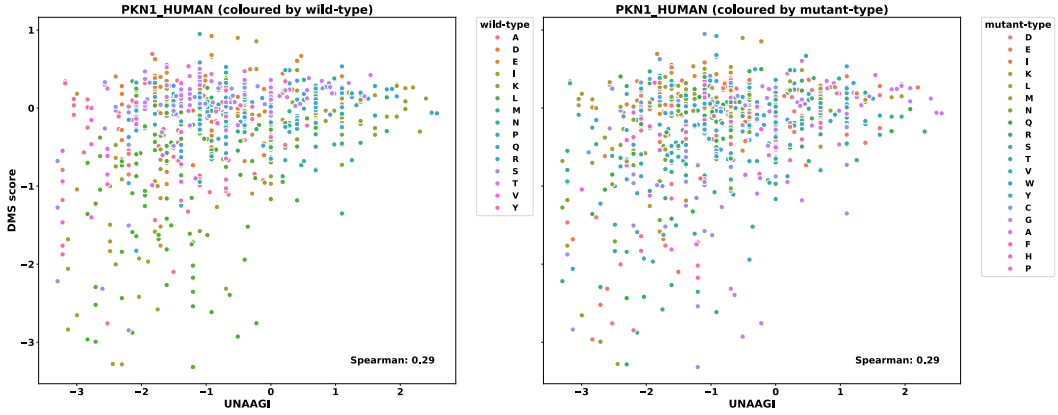

Figure 15: Scatterplot of UNAAGI vs. DMS score

RS15_GEOSE

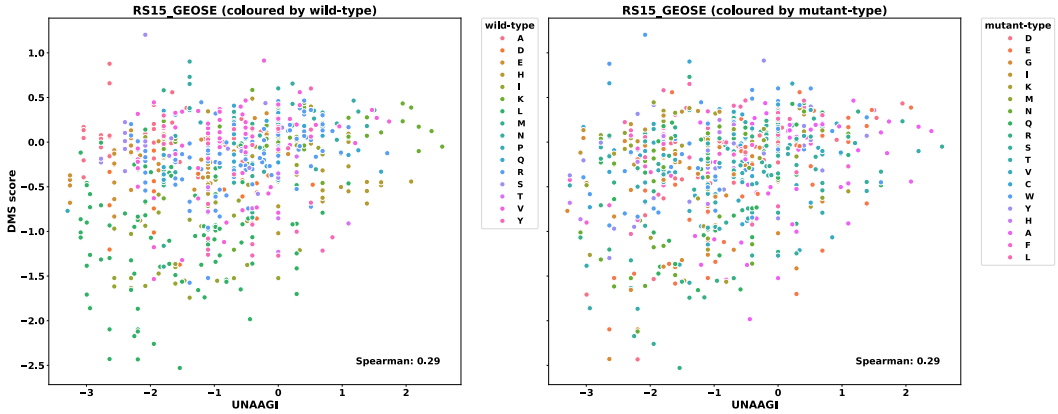

Figure 16: Scatterplot of UNAAGI vs. DMS score

SBI_ STAAM

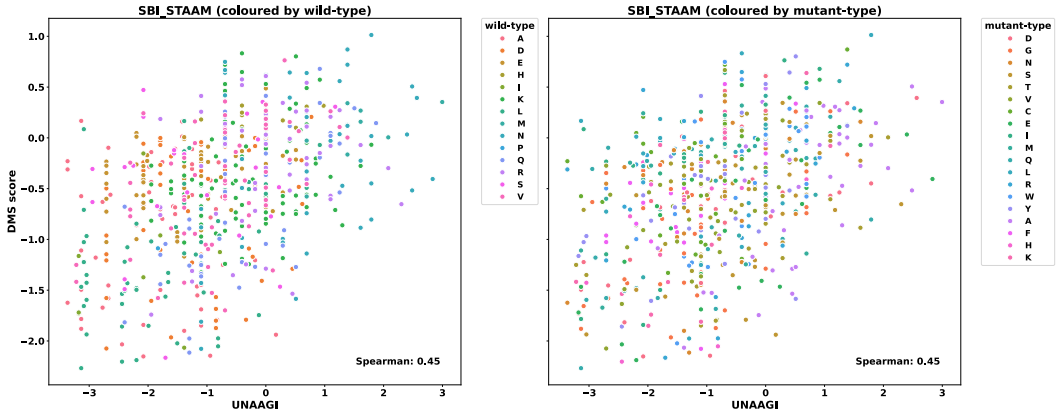

Figure 17: Scatterplot of UNAAGI vs. DMS score

SCIN_ STAAR

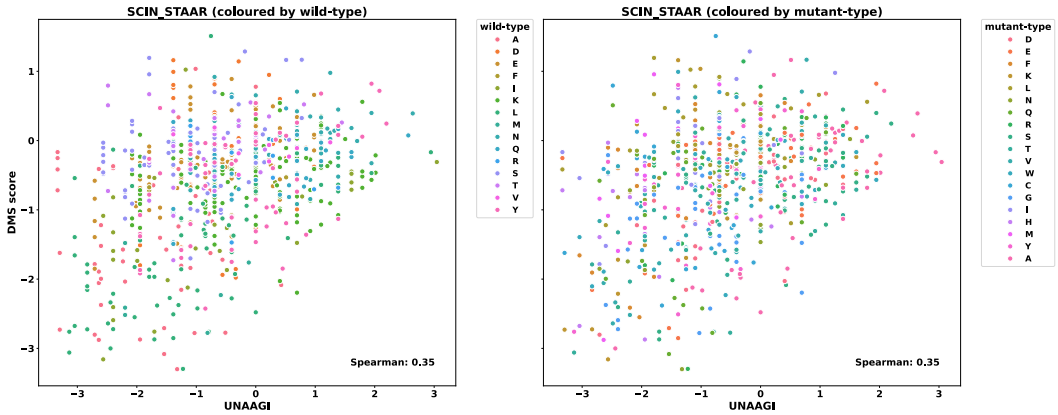

Figure 18: Scatterplot of UNAAGI vs. DMS score

SOX30_ HUMAN

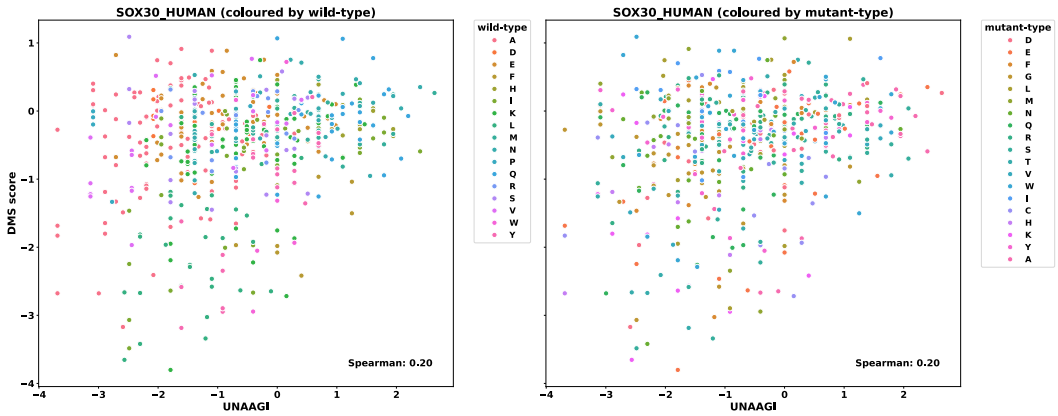

Figure 19: Scatterplot of UNAAGI vs. DMS score

SQSTM_ MOUSE

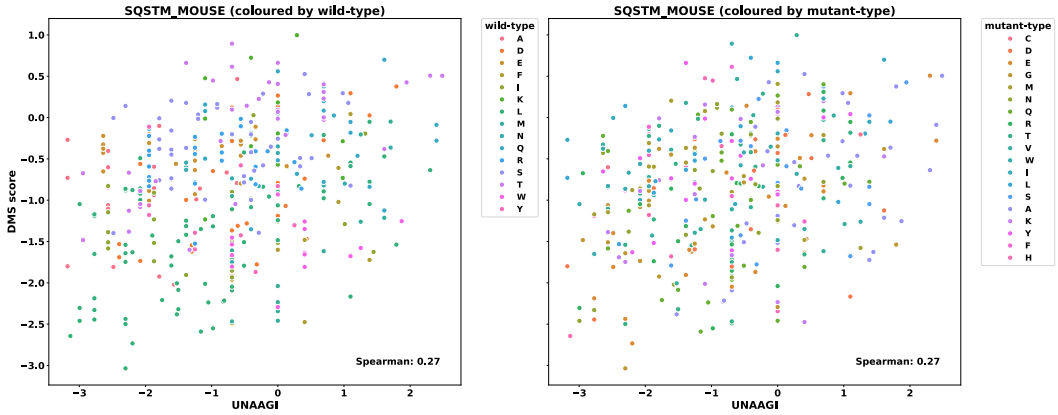

Figure 20: Scatterplot of UNAAGI vs. DMS score

SUMO1_ HUMAN

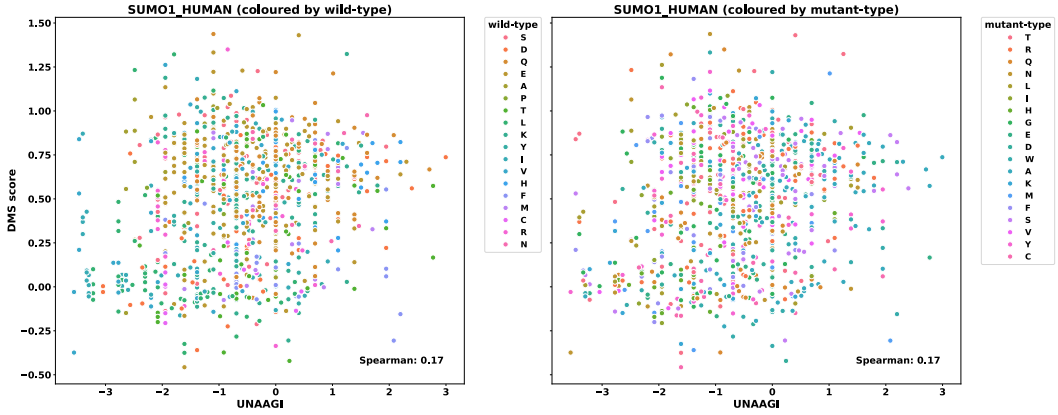

Figure 21: Scatterplot of UNAAGI vs. DMS score

TAT_ HV1BR

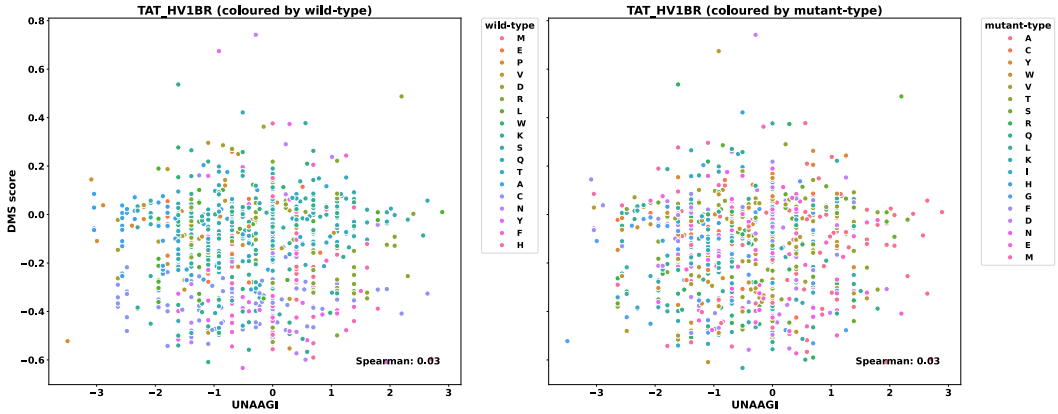

Figure 22: Scatterplot of UNAAGI vs. DMS score

VG08_ BPP22

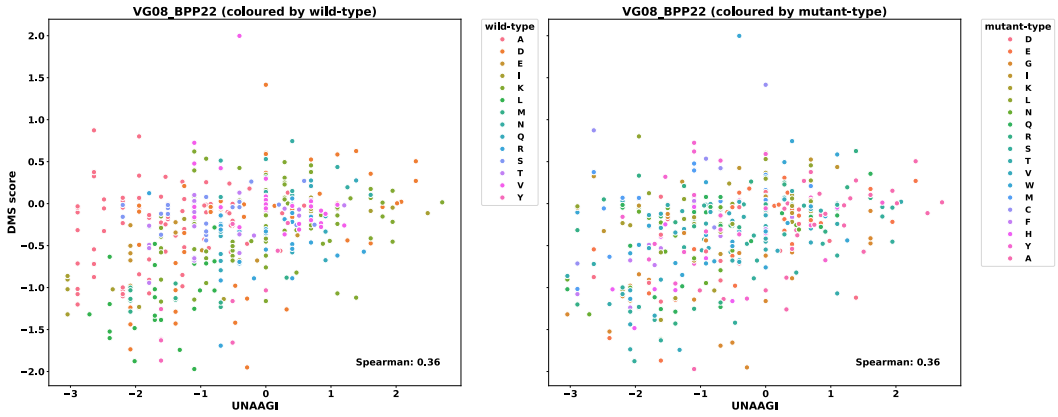

Figure 23: Scatterplot of UNAAGI vs. DMS score

VRPI_ BPT7

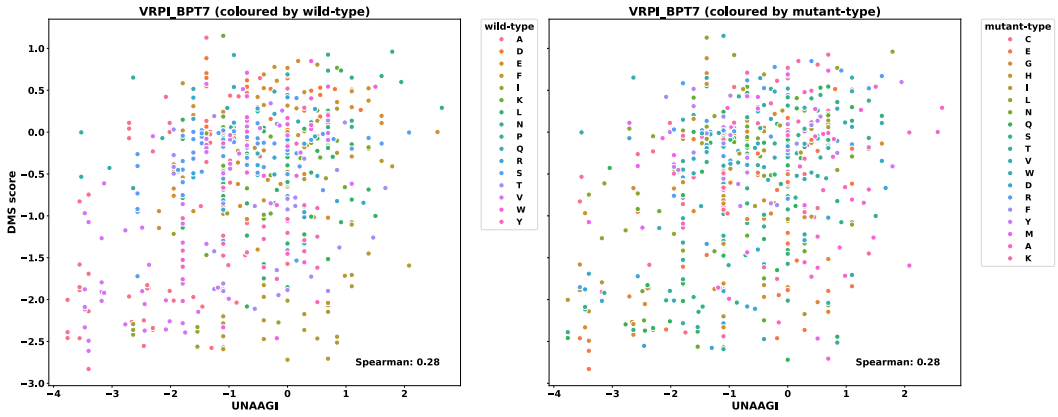

Figure 24: Scatterplot of UNAAGI vs. DMS score

A.1.2 DETAILED RESULTS ON DMS OF NCAAS

For results on DMS of NCAAs, we report the spearman correlation over all the mutants, and for the NCAA mutants specifically, as depicted in Figure 4.

CP2

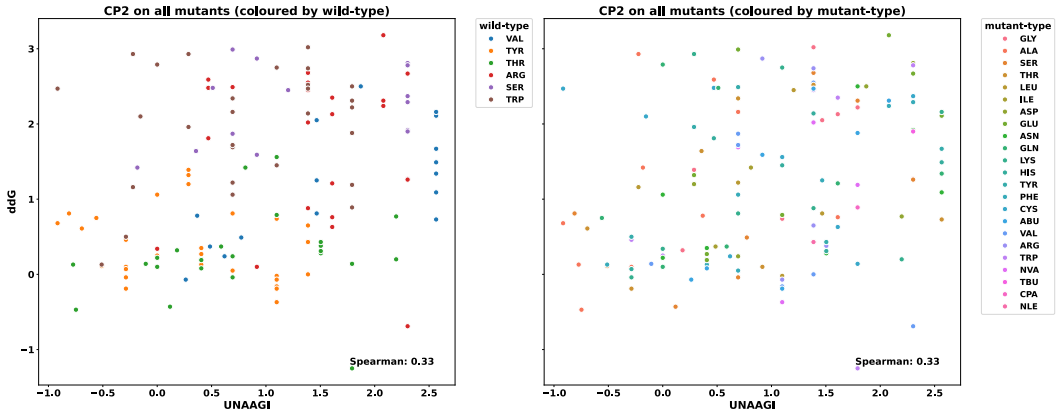

Figure 25: Scatterplot of UNAAGI vs. ddG, for all mutants

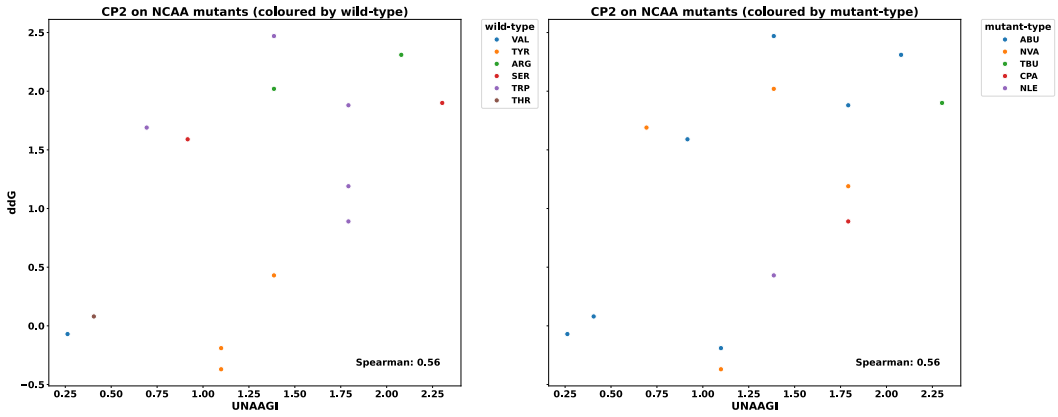

Figure 26: Scatterplot of UNAAGI vs. ddG, for NCAA mutants

PUMA

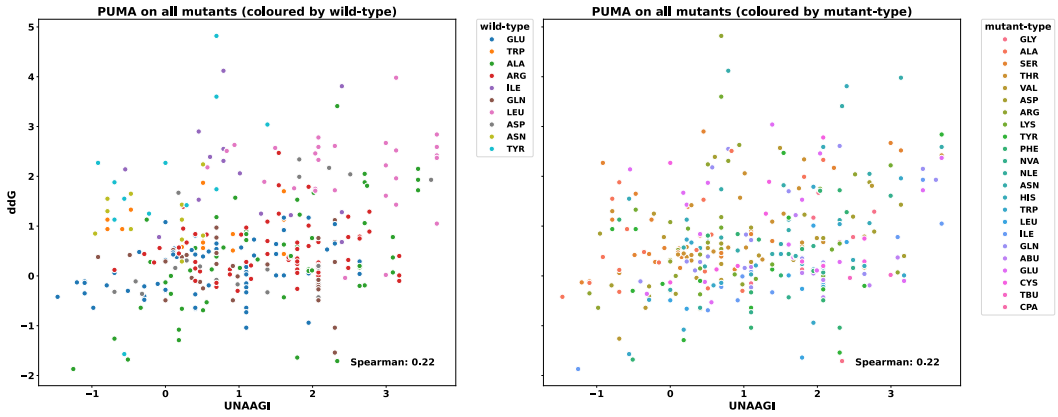

Figure 27: Scatterplot of UNAAGI vs. ddG, for all mutants

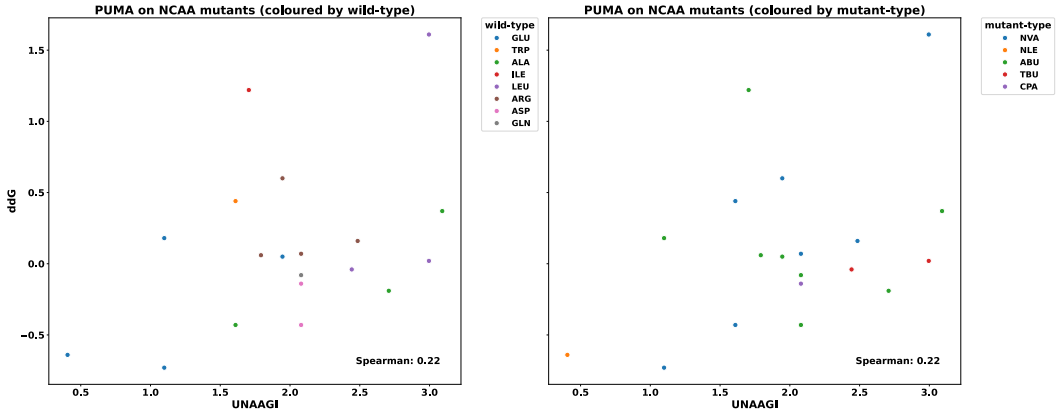

Figure 28: Scatterplot of UNAAGI vs. ddG, for NCAA mutants

A.2 THE USE OF LARGE LANGUAGE MODELS (LLMS)

In accordance with ICLR requirements, we disclose the role of LLMs in preparing this work. Large language models were not used for any experimental design, data analysis, or implementation. Their use was limited to assisting with minor language editing, such as correcting grammar and polishing phrasing.