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

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

A central challenge in structural biology is identifying beneficial amino acid substitutions, a problem that underpins both mutational effect prediction and protein engineering. Recent inverse-folding models, trained to reconstruct sequences from structure, have shown considerable promise for identifying functional mutations. However, current approaches are constrained to designing sequences composed exclusively of canonical amino acids (CAAs). Non-canonical amino acids (NCAAs) offer greater chemical diversity and are frequently used for protein engineering in vivo, yet they remain largely inaccessible to 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 canonical and non-canonical amino acid substitutions within a unified generative paradigm. We evaluate UNAAGI zero-shot, without mutational-effect supervision or task-specific fine-tuning, on experimental mutational-effect benchmarks and demonstrate substantially improved performance on NCAA substitutions relative to existing methods.

## 1 INTRODUCTION

Proteins perform diverse biological functions through three-dimensional structures formed by linear chains of amino acids. Although amino acid side chains can, in principle, span broad chemical diversity, naturally occurring proteins primarily use a set of 20 genetically encoded canonical amino acids (CAAs). This set is thought to be evolutionarily optimized for functional completeness and broad chemical coverage (Ilardo et al., 2015; Doig, 2017; Ilardo et al., 2019). Nevertheless, non-canonical amino acids (NCAAs) provide side-chain chemistries absent from the standard set, including enhanced metal coordination, tailored electrostatics, and novel catalytic properties. Incorporating NCAAs therefore expands the functional repertoire of proteins beyond evolutionary constraints, opening 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 manner by predicting amino acid propensities conditioned on sequence (Riesselman et al., 2018) or structural (Dauparas et al., 2022) context. These methods generally represent amino acids with a discrete distribution over the 20 canonical residues, learned from large evolutionary databases. Although such models have shown impressive zero-shot performance across various protein properties, their discrete representation prevents generalization beyond the canonical amino acid vocabulary, creating a fundamental limitation in the current modeling paradigm.

A central question is whether inverse-folding models can be generalized beyond canonical amino acids. This problem presents two challenges: (i) the vast chemical space of NCAAs makes categorical modeling impractical, and (ii) the predictive capabilities of inverse-folding models arise from evolutionary constraints encoded in millions of naturally occurring protein sequences, a source of information that is not available for NCAAs.



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

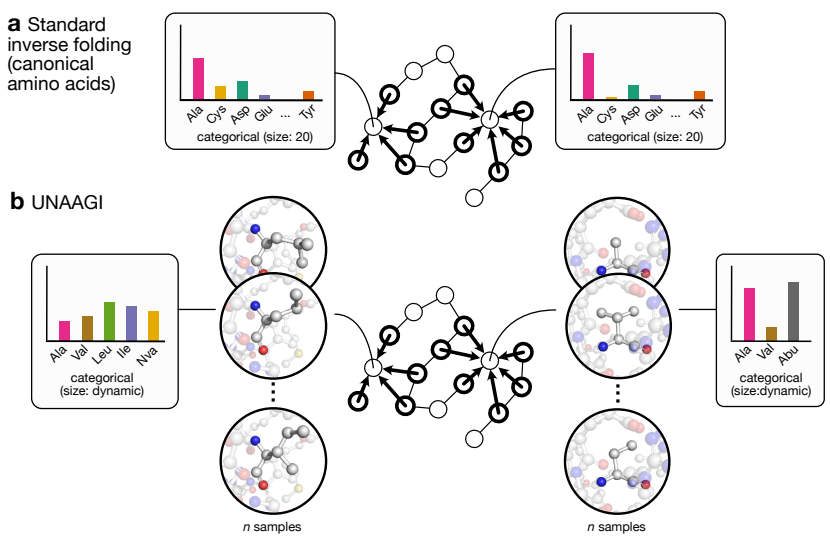



Figure 1: Illustration of UNAAGI compared with a standard inverse-folding model. **(a)** Inverse-folding models parameterize the output distribution as a categorical distribution over a fixed-size vocabulary. **(b)** UNAAGI generates amino acids directly in chemical space. A finite set of samples defines an empirical distribution over the context-dependent region of chemical space, typically including both canonical amino acids (CAAs) and non-canonical amino acids (NCAAs).

In this paper, we test whether atom-level generative modeling can provide a route beyond fixed amino-acid vocabularies. The key idea is that local geometric and chemical constraints learned from canonical amino acids may transfer to chemically proximal NCAAs. This formulation addresses both challenges: it avoids defining a fixed NCAA vocabulary while leveraging abundant canonical amino acid data to support design beyond the canonical set. Although such transfer is unlikely to cover the full diversity of NCAA chemical space, it may enable structure-conditioned assessment of substitutions whose side-chain chemistry remains close to that of canonical residues.

To test this hypothesis, we propose a generative framework for residue-level sequence design based on atom-level side-chain generation through equivariant molecular diffusion. We term our approach **U**ncanonical **N**ovel **A**mino **A**cid **G**enerative **I**nference (**UNAAGI**). By modeling side-chain atomic coordinates directly, UNAAGI implicitly infers residue identity and can propose plausible non-canonical substitutions. Figure 1 illustrates the central idea.

To evaluate the generalizability of UNAAGI, we first trained models while holding out selected canonical amino acids from the training set, testing whether UNAAGI could recover unseen amino-acid chemistries, particularly at structural sites where the held-out residue is experimentally favored. We then evaluated UNAAGI trained on the complete dataset using experimentally measured mutational effects (Rogers et al., 2018; Notin et al., 2023), comparing its generative likelihoods with observed substitution outcomes. UNAAGI substantially improves predictive performance for non-canonical substitutions relative to existing methods, while maintaining consistent performance across canonical and non-canonical residues.

Specifically, our contributions are as follows:

- We introduce UNAAGI, an E(3)-equivariant diffusion framework for atom-level side-chain generation and residue identity inference.
- We evaluate controlled generalization beyond the training vocabulary by holding out selected canonical amino acids and testing whether UNAAGI can recover them as unseen chemistries.
- We demonstrate that UNAAGI can generalize to chemically proximal non-canonical amino acids, enabling structure-conditioned assessment of NCAA substitutions.

- We evaluate UNAAGI in zero-shot variant effect prediction, without mutational-effect supervision or task-specific fine-tuning, and demonstrate improved performance on experimental NCAA substitution benchmarks for the identified chemically proximal subclass.

## 2 Related Work

Our approach connects prior work in molecular diffusion, variant effect prediction, and peptide design.

### 2.1 Molecular Diffusion

Hoogeboom et al. (2022) introduced E(3)-equivariant diffusion frameworks for molecular generation, enabling the direct generation of molecular structures in 3D space through a diffusion process. Subsequent works, including Peng et al. (2023) and Vignac et al. (2023), incorporated additional inductive biases from molecular bonding structures, leading to improved molecular consistency and sample quality. More recently, Le et al. (2024) conducted a systematic comparison of several molecular diffusion frameworks, examining how different architectural and modeling choices affect generative performance and identifying combinations of design choices that yield strong performance across molecular generation tasks.

Beyond unconditional molecular generation, an important application is the conditional generation of compounds that bind to specific protein targets, commonly 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 related models for target-aware ligand sampling. More recently, Schneuing et al. (2025) introduced models that sample ligands with variable atom counts while incorporating protein-pocket side-chain conformations during generation.

### 2.2 Mutational Effect Prediction

Mutational effect prediction aims to estimate how amino acid substitutions affect a protein property of interest, such as stability, affinity, or function. Several branches of machine learning models have demonstrated sensitivity to this problem, including pure sequence-based models, structure-to-sequence (inverse-folding) models, and hybrids of the two

Statistical models of multiple sequence alignments remain strong baselines when many sequences are available for a given protein family. Notable examples are the DeepSequence VAE model and the EVE model (Riesselman et al., 2018; Frazer et al., 2021). Language-based models such as ESM (Meier et al., 2021; Lin et al., 2023) obviate the need for alignments, leveraging conservation and coevolutionary signals learned from large protein sequence datasets. Although sequence-based models can predict mutational effects in many contexts, their performance declines when evolutionary signals are sparse, as in shallow MSAs, antibodies, or de novo designed peptides.

Structure-based approaches replace the sequence context of a mutation with structural context. Early models considered the local environment around one amino acid at a time (Boomsma & Frellsen, 2017; Torng & Altman, 2017); this idea was later generalized to entire sequences in inverse-folding models (Ingraham et al., 2019; Dauparas et al., 2022; Hsu et al., 2022). Although primarily developed for protein design, these models are also useful for mutational effect prediction (Frellsen et al., 2025).

Many existing all-atom protein models still operate in two stages: they first predict a canonical amino acid identity and then generate its corresponding atomic conformation. UNAAGI similarly models one residue at a time, conditioned on its surrounding structural environment. The crucial difference is that UNAAGI generates the side chain directly at the atomic level, without first assigning it to a predefined amino acid class. This formulation allows the model to represent side-chain chemistries beyond the canonical amino-acid alphabet and thereby extend mutational effect prediction to 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 representation enables the model to generate amino acid SMILES

strings at masked positions. The strategy supports generalization to non-canonical amino acids, but the original study does not validate the resulting sample probabilities against experimental measurements. We therefore include PepINVENT as a baseline.

RareFold (Li et al., 2025) and NCFlow (Lee & Kim, 2026) are based on AlphaFold-style architectures. RareFold introduces new tokens for NCAAs and focuses on structure prediction for proteins containing NCAA substitutions and on the design of NCAA-containing peptides. The study reports experimentally validated peptide designs, but the model itself does not produce in silico affinity or mutational-effect scores, making direct comparison on NCAA mutational effect prediction infeasible. NCFlow adopts a flow-matching framework and evaluates mutational effect prediction on the same benchmark as ours, enabling direct comparison (Section 4).

# 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 canonical amino acid space. The model learns patterns in the chemistry of canonical residues and uses them to generalize to chemically proximal non-canonical amino acids. We next outline equivariant graph diffusion and then proceed by describing the model architecture, graph construction, and sampling procedure.

## 3.1 Multimodal Diffusion for Molecular Graphs

We summarise the standard denoising-diffusion formulation, and the equivariance definitions used below, in Appendix A.

Molecular data are inherently *multimodal*: atomic coordinates are continuous, whereas atom types, bond types, and other chemical features are discrete. Diffusion over discrete variables can be defined through categorical transition kernels, which are discrete analogues of Gaussian noise in the forward process (Hoogeboom et al., 2021; Austin et al., 2021). Recent molecular diffusion models combine continuous and categorical processes to jointly model these heterogeneous modalities (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

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

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., 2021), 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 is the standard diffusion ELBO, stated in full in Appendix A. It 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., 2021). We adopt the $\hat{\mathbf{x}}_0$-parameterization to predict the clean data directly. The loss combines mean-squared error for atomic coordinates with cross-entropy for 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. The relevant symmetry group is $\mathrm{E}(3)$ (rotations, reflections and translations); the formal equivariance conditions are given in Appendix A.

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 and charge) and vector features $\mathbf{v}_i \in \mathbb{R}^3$, while edges encode bond types and interatomic distances. Further architectural and training details are provided in Appendix B.

### 3.3 Graph Topology and Virtual Node Padding

Unlike traditional structure-based drug design (SBDD), in which generative models sample ligands conditioned on a protein pocket, our task focuses on reconstructing amino acid side chains. Each generated side chain is covalently bonded to a fixed protein backbone, imposing local structural constraints around the masked region. We therefore fix the positions and types of the backbone atoms for the residue under reconstruction, anchoring generation to the known structural context.

A further challenge is unified sampling across canonical and non-canonical amino acids, whose side chains contain different numbers of atoms. In standard SBDD models, the atom count must be fixed before diffusion because the graph topology is predetermined. We address this challenge using a virtual-node strategy inspired by DrugFlow (Schneuing et al., 2025). Unlike empirical size-sampling schemes (Hoogeboom et al., 2022; Schneuing et al., 2024; Guan et al., 2023) or separate size-estimation networks (Igashov et al., 2024), this approach is fully end to end and allows the model to learn a distribution over variable side-chain 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 $n_{\text{virt}} \sim U(0, N_{\max}^{\text{train}})$ are added to the side chain. Following Schneuing et al. (2025), these nodes are initialized at the side-chain center of mass, allowing the model to learn which nodes correspond to real atoms. At sampling time, the model denoises a graph initialized with $N_{\max}^{\text{sample}}$ 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 construct a curated training dataset by combining data from multiple sources:

- **Canonical protein structures.** We use a filtered set of protein structures from the Protein Data Bank (PDB), with full dataset construction details provided in Appendix B.3. For each structure, we extract every residue together with its local structural environment, defined as all surrounding residues whose centers of mass lie within 10 Å of the target residue.
- **Non-canonical amino acids.** To expose the model to chemical structures beyond the canonical amino-acid alphabet, we incorporate non-canonical amino acids (NCAAs) from Ilardo et al. (2019) and SwissSidechain (Gfeller et al., 2013). These datasets contain isolated amino-acid structures without associated protein environments. During training, the four backbone atoms (N, $C_\alpha$, C, and O) are kept fixed, while noise is applied only to the side-chain atoms.
- **Protein–ligand complexes.** We include protein–ligand complexes from PDBBind (Wang et al., 2005) to expose the model to a broader range of local chemical interactions and ligand-like molecular structures. Following the structure-based drug design setup of Le et al. (2024), we keep the protein-pocket atoms fixed and apply noise to all ligand atoms.
- **PDB structures containing NCAAs.** We additionally collect protein structures from the PDB that contain non-canonical amino acids, providing examples in which NCAA side chains occur in explicit protein environments. Details of the collection and filtering procedure are provided in Appendix B.3.

| | Rebuilds native | Designs at favorable sites | | | Prefers favorable over harmful sites | | | |
|---|---|---|---|---|---|---|---|---|
| Excluded residue | without it (rate) | without it (rate) | full model (rate) | % kept vs. full | favorable sites (rate) | harmful sites (rate) | fold change | $p$ |
| VAL | 0.990 | 0.921 | 0.972 | 95% | 0.928 | 0.843 | 1.10× | 3.2e−6 |
| SER | 1.000 | 0.896 | 0.977 | 92% | 0.908 | 0.720 | 1.26× | 1.2e−18 |
| PHE | 0.929 | 0.674 | 0.896 | 75% | 0.687 | 0.589 | 1.16× | 4.2e−4 |
| ILE | 0.885 | 0.357 | 0.863 | 41% | 0.407 | 0.295 | 1.38× | 1.1e−4 |

Table 1: **The model generates held-out residues with a systematic difficulty gradient.** Each value is the fraction of sites at which a model trained *without* the excluded residue nonetheless generates it ("without it"), compared with the model trained on all 20 canonical amino acids ("full model"). **Rebuilds native** is reconstruction ($\mathrm{wt} = X$); **designs at favorable sites** is generalization ($\mathrm{wt} \neq X$, with mutation to $X$ stabilizing according to deep mutational scanning); **prefers favorable over harmful sites** is selectivity, and "fold change" is the ratio of those two rates. Per-residue site counts, Fisher exact-test $p$-values and pooled Wilson intervals are given in Appendix D.1.

To assess the contribution of each supplementary data source, we perform ablation experiments in which individual components are removed from the training set and evaluate the resulting changes in model performance. Further details are provided in Section 4.6.

### 4.2 Accelerated Sampling and Residue Identification

Given a fixed local protein environment, UNAAGI generates variable-size side-chain molecular graphs by jointly predicting atomic coordinates, atom features, and bond types. After virtual nodes are removed, generated side chains are assigned to known canonical or non-canonical amino acids by matching their labeled molecular graphs to reference structures using graph isomorphism (Weisfeiler & Lehman, 1968). We use the NetworkX 3.5 implementation of `is_isomorphic`, which is based on the VF2 algorithm (Cordella et al., 2001).

Downstream evaluation requires many independent samples at each residue position to estimate the model-induced distribution over amino acid identities (Section 4.4). Standard ancestral DDPM sampling requires 500 denoising steps per sample and is therefore computationally expensive. At inference time, we instead apply DPM-Solver++ (Lu et al., 2025) to the continuous coordinates and continuous-time Markov chain updates (Campbell et al., 2022) to the categorical atom and bond features. Together, these methods reduce the number of denoising steps from 500 to 20 per sample.

### 4.3 Rediscovering Canonical Amino Acids from Held-out Models

Large-scale and diverse experimental ground truth for NCAA mutational effects is unavailable, preventing direct evaluation of whether a generative model designs them appropriately. We therefore construct a surrogate task by removing one canonical amino acid entirely from training. The model must then treat the held-out residue as unseen chemistry and reconstruct it atom by atom, analogous to the setting encountered for a true NCAA. Using ProteinGym measurements, we test whether the model can (i) generate the held-out residue and (ii) place it at positions where it is favored. We train four separate held-out models, each excluding one amino acid both as a reconstruction target and from the surrounding structural context to prevent data leakage.

We compare each held-out model with a model trained on all 20 canonical amino acids. For evaluation, we randomly select 25 assays from ProteinGym (Notin et al., 2023) and separately hold out Ser, Ile, Phe, and Val. As shown in Table 1, the held-out models recover residues that never appear in their training data, both at native sites and at positions where the substitution is stabilizing. The models also exhibit selectivity, generating the held-out residue more frequently at favorable positions (native and stabilizing sites combined) than at harmful positions.

### 4.4 Evaluation on Deep Mutational Scanning (DMS)

Given a fixed local protein environment, UNAAGI generates diverse side-chain identities and conformations. We interpret the resulting frequencies over residue identities as an empirical estimate of the model-induced distribution at each position.

To evaluate predictive performance, we compare this learned distribution with experimental deep mutational scanning (DMS) data.[1] For each site in a DMS benchmark, we perform 1,000 sampling iterations and record the frequency of each sampled amino acid.[2] Although larger sample sizes would provide smoother frequency estimates, we use 1,000 iterations as a trade-off between statistical stability and computational cost. We repeat each experiment five times and report error bars to quantify sampling variability.

From the sampling frequencies, we estimate the probabilities assigned to the mutant and wild-type residues. We define the mutation score as the difference between their negative log-likelihoods:

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

where $P(\text{mutant})$ and $P(\text{wild-type})$ denote the estimated probabilities of the mutant and wild-type residues, respectively. Larger values therefore indicate that the mutant is less probable than the wild type under the model. We correlate this score with experimentally measured mutational effects.

For non-canonical amino acids, we evaluate UNAAGI on the DMS benchmark introduced by Rogers et al. (2018). The study measured mutational effects for two peptide–protein complexes: PDB ID 5LY1 (JMJD2A/KDM4A bound to the macrocyclic peptide CP2) and PDB ID 2ROC (Mcl-1 bound to PUMA). The dataset reports $\Delta\Delta G$ measurements for substitutions spanning 20 canonical and 20 non-canonical amino acids. To assess performance on canonical amino acids, we use the same 25 ProteinGym assays as in Section 4.3 and evaluate the canonical amino-acid substitutions generated by UNAAGI.[3]

## 4.5 Results

### 4.5.1 Mutational Effect Prediction for Non-Canonical Amino Acids

We evaluate predictive performance using the Spearman correlation between mutation scores and experimental $\Delta\Delta G$ values. We first report results on the DMS benchmark containing non-canonical amino acid substitutions. We include PepINVENT, NCFlow, and AlphaFold 3 as baselines. NCFlow was evaluated on the same benchmark in its original study, whereas PepINVENT supports the generation of non-canonical amino acids but was not originally evaluated on DMS data. For PepINVENT, we follow the repository guidelines by masking one residue at a time and generating candidate substitutions at every position in each assay. For peptide–protein complexes, we concatenate the peptide and protein sequences as input to the language model. Sampled SMILES strings corresponding to amino acids in the benchmark are mapped back to residue identities to estimate frequencies. AlphaFold 3 accepts NCAAs with Chemical Component Dictionary (CCD) codes, including those covered by this benchmark; we use the predicted interface TM-score (iPTM) as its mutation score. To ensure consistent directionality across methods, we reverse the signs of the AlphaFold 3 and NCFlow scores. RareFold, by contrast, restricts NCAA inputs to a predefined residue dictionary. Since none of the NCAAs in our benchmark are included in this dictionary, RareFold cannot be evaluated on this benchmark.

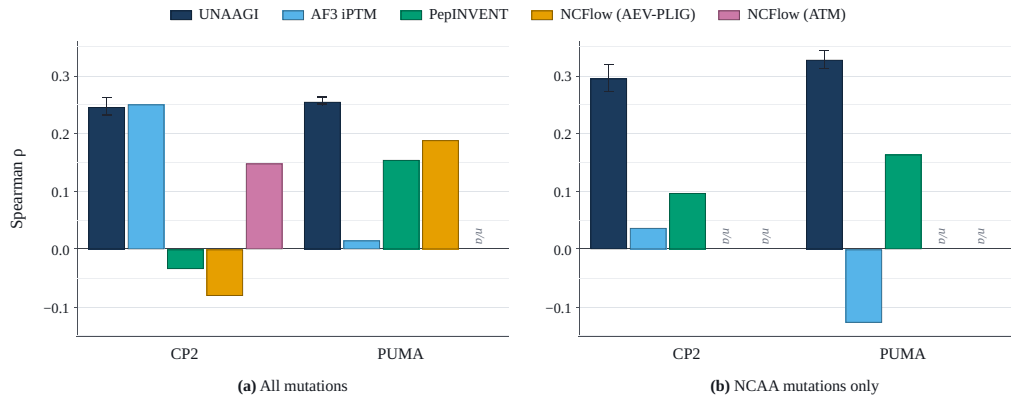
As shown in Figure 2, UNAAGI performs consistently across the all-substitution and NCAA-only evaluations. Its correlations on NCAA substitutions are comparable to those obtained when canonical substitutions are also included, suggesting that UNAAGI can generalize beyond 20 canonical amino acids. In contrast, PepINVENT shows limited signal on NCAA substitutions, and NCFlow yields

[1] ProteinGym is also used in Section 4.3, but there only to identify favorable and harmful substitution sites for the held-out-residue analysis. The present section directly evaluates mutational-effect prediction using DMS benchmarks containing both canonical and non-canonical amino-acid substitutions.

[2] We assign unsampled amino acids a pseudocount of 1 (Laplace add-one smoothing), equivalent to the posterior mean of the sampling distribution under a uniform Dirichlet(1) prior. This bounds the log-likelihood ratio when a substitution is never sampled. In Appendix D.2, we assess robustness by excluding unsampled amino acids from the evaluation.

[3] The computational cost of UNAAGI scales approximately linearly with protein size because it operates on local structural environments. Scaling to larger proteins is therefore not fundamentally prohibitive; however, due to limited computational resources, we evaluate a randomly selected subset rather than the full ProteinGym benchmark.

weak or negative correlations regardless of the affinity-prediction module used[4]. AlphaFold 3 performs comparably to UNAAGI on CP2 when all substitutions are included, but its performance decreases on NCAA-only substitutions and on PUMA. The PUMA receptor contains less evolutionary information than the CP2 receptor, which may partly explain this performance gap (Appendix E.4).




**(a)** All mutations **(b)** NCAA mutations only

Figure 2: Mutational effect prediction on the CP2 and PUMA DMS benchmarks, evaluated using Spearman correlation. **(a)** Performance across all measured substitutions, compared with AF3 iPTM, PepINVENT, and NCFlow variants. **(b)** Performance restricted to non-canonical amino acid substitutions, compared with AF3 iPTM and PepINVENT.

#### 4.5.2 Sanity Check on Canonical Mutational Effect Prediction

As a sanity check, we evaluate UNAAGI on ProteinGym by computing the Spearman correlation between predicted mutation scores and experimental fitness measurements, consistent with the standard metric reported by ProteinGym. Across most assays, UNAAGI exhibits meaningful correlation with experimental fitness (Figure C1, Appendix C), indicating that its atom-level generative distribution captures biologically relevant variation among canonical amino-acid substitutions. Although UNAAGI is not designed specifically for canonical sequence-based mutational-effect prediction and does not achieve state-of-the-art performance on individual assays, these results provide an independent validation that the structural preferences learned by the model transfer to mutational-effect prediction.

Among the baselines, PepINVENT is methodologically closest to UNAAGI because it also represents and samples residues at the atomic level. UNAAGI consistently outperforms PepINVENT across all evaluated assays: PepINVENT shows little or no correlation on most assays, whereas UNAAGI retains substantial predictive signal.[5]

### 4.6 Ablation Studies

To assess the contribution of each supplementary training source, we train separate models after removing (i) NCAA-containing PDB structures, (ii) PDBBind complexes, (iii) the independent NCAA dataset, or (iv) all three sources simultaneously. These four ablated models are summarized in Table 2. UNAAGI (Ablation: NCAA in PDB), UNAAGI (Ablation: PDBBind), and UNAAGI (Ablation: NCAA) denote models trained without the corresponding source; UNAAGI (Ablation: All) is trained only on the canonical PDB dataset. We evaluate the effect of each ablation by measuring the average Spearman correlation across ProteinGym assays.

Removing any supplementary data source reduces performance, indicating that each contributes useful information. The model trained only on the canonical PDB dataset performs worst overall.

[4]NCFlow results are taken from the original paper. Since NCFlow is not open-sourced, we only report results for all mutations. The original paper reports that the ATM variant fails to produce results for PUMA.

[5]PepINVENT cannot generate samples when the evaluated sequence exceeds its maximum input length.

Table 2: Ablation study of the supplementary training-data sources.

| **Model** | **Spearman $\rho$** |
|---|---|
| UNAAGI | **0.310** |
| UNAAGI (Ablation: NCAA in PDB) | 0.267 |
| UNAAGI (Ablation: PDBBind) | 0.219 |
| UNAAGI (Ablation: NCAA) | 0.205 |
| UNAAGI (Ablation: All) | 0.198 |

### 4.7 Capabilities and Limitations

To further investigate the behavior of UNAAGI, we manually inspected a set of sampled amino acids, as shown in Figure E1. We observed two main phenomena:

1. **Chemical diversity.** UNAAGI generates amino acids spanning a broad range of chemistries, including hydrophobic, uncharged polar, positively charged, and negatively charged residues.
2. **Sampled NCAAs are structurally proximal to canonical amino acids.** UNAAGI is designed to explore non-canonical amino acids in the vicinity of known amino-acid chemistry rather than to generate arbitrarily distant molecular structures. Consistent with this objective, the generated NCAAs often resemble canonical residues or interpolate between them, reflecting both the predominance of canonical amino acids in the training data and the model's focus on chemically proximal substitutions.

In the benchmark described in Section 4.5.1, UNAAGI consistently recovers a subset of the evaluated NCAAs: **Nle**, **Nva**, **Abu**, **Aoc**, **tBu**, **hSM**, **Cpa**, and **TMe**. However, it does not reliably sample other NCAA classes, including residues with *N*-methyl substitutions or more complex ring systems. Appendix E.2 provides a more detailed analysis of NCAA coverage. We attribute this limitation in large part to the scarcity of relevant NCAA data. Although consistent with its intended scope, UNAAGI primarily explores NCAAs that are chemically proximal to canonical residues.

## 5 Conclusion

We introduced UNAAGI, a molecular diffusion model that samples both canonical and non-canonical amino acids. Across deep mutational scanning benchmarks, UNAAGI achieves meaningful correlations with experimental variant effects without using mutational-effect supervision. Crucially, this predictive signal extends to NCAA substitutions, making UNAAGI, to our knowledge, the first atom-level protein design model to demonstrate positive predictive signal on experimental NCAA mutational-effect benchmarks, while achieving state-of-the-art performance among the evaluated methods.

Methodologically, UNAAGI follows the molecular diffusion paradigm developed for structure-based drug design (SBDD), learning a distribution over chemically plausible side chains in protein environments. This formulation highlights a promising connection between SBDD and variant effect prediction, as both depend on modeling local non-covalent interactions. 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.

Looking forward, several directions could extend UNAAGI. Scaling model capacity and training data may improve coverage, while conditional diffusion could enable classifier-free guidance toward specific NCAAs or desired properties, such as synthetic accessibility.

## AI use statement

We used generative AI tools as general-purpose writing and coding assistants in the preparation of this work. Specifically, AI assistance was used for manuscript preparation—copy-editing, LaTeX formatting, and tightening the exposition of text drafted by the authors—and for writing and debugging

portions of the analysis, evaluation and plotting code supporting the experiments reported in Section 4. We did not use generative AI tools to generate research ideas, to formulate or design the method, to produce or interpret experimental results, or to retrieve and summarise related work; the scientific claims, modelling choices and reported numbers in this paper originate with the authors. All AI-assisted text was reviewed and revised by the authors, and all AI-assisted code was checked by the authors against the results it produced. We take responsibility for the final content of this work, including text, claims and artifacts produced with the aid of generative AI.

## ETHICS STATEMENT

This work uses only publicly available structural and mutational-effect data— the Protein Data Bank, PDBBind, and ProteinGym—and involves no human subjects, no personally identifiable information, and no newly collected experimental data. UNAAGI extends structure-based protein design beyond the canonical amino acid vocabulary, which broadens the accessible chemical space for protein engineering, therapeutic development and synthetic biology. As with other generative models in structural biology, this capability carries a dual-use risk: the same modelling that supports beneficial protein engineering could in principle be directed toward harmful designs. We regard that risk as limited in the present setting, since UNAAGI proposes and scores single-residue substitutions in a given structural context rather than designing complete functional molecules, and any proposed substitution requires substantial downstream experimental work before it has a biological effect. We nonetheless encourage that applications of this and related methods, particularly de novo design, be pursued with appropriate experimental validation and in adherence to established biosafety and research-ethics guidelines. The authors declare no conflicts of interest.

## REPRODUCIBILITY STATEMENT

The information required to reproduce our results is distributed across the main text and the appendix. The model, its equivariant diffusion formulation and the training objective are specified in Section 3, with the full architecture and hyperparameter settings in Appendix B. Dataset composition and pre-processing are described in Section 4.1, and the construction of the NCAA-containing portion of the training data in Appendix B.3. The sampling and residue-identification procedure used at inference is given in Section 4.2; the held-out amino acid protocol in Section 4.3; and the mutational-effect evaluation, including the DMS benchmarks, in Section 4.4 and Section 4.5.1. Appendix F states precisely what our zero-shot claim does and does not cover, including the relationship between benchmark proteins and the structural training corpus. Two sensitivities of the reported numbers are analysed explicitly: robustness to the pseudocount assigned to unsampled amino acids in Appendix D.2, and the geometric quality control applied to generated residues in Appendix E. All datasets used are publicly available. We will release all code upon acceptance.

## 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. In *Advances in Neural Information Processing Systems*, volume 34, pp. 17981–17993, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/958c530554f78bcd8e97125b70e6973d-Abstract.html.

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

Andrew Campbell, Joe Benton, Valentin De Bortoli, Tom Rainforth, George Deligiannidis, and Arnaud Doucet. A continuous time framework for discrete denoising models. In *Advances in Neural Information Processing Systems*, volume 35, pp. 28266–28279, 2022. doi: 10.52202/068431-2049. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/b5b528767aa35f5b1a60fe0aaeca0563-Abstract.html.

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.

Luigi P. Cordella, Pasquale Foggia, Carlo Sansone, and Mario Vento. An improved algorithm for matching large graphs. In *Proceedings of the IAPR-TC15 Workshop on Graph-based Representations*, 2001.

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.

Jonathan Frazer, Pascal Notin, Mafalda Dias, Aidan Gomez, Joseph K Min, Kelly P. Brock, Yarin Gal, and Debora S. Marks. Disease variant prediction with deep generative models of evolutionary data. *Nature*, 2021.

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. In *Advances in Neural Information Processing Systems*, volume 38, pp. 94328–94354, 2025. doi: 10.52202/085713-2836. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/7a209a1c557e5087272014e7dd37035f-Abstract-Conference.html.

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. Swisssidechain: a molecular and structural database of non-natural sidechains. *Nucleic Acids Research*, 41(Database issue):D327–D332, 2013. doi: 10.1093/nar/gks991.

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. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=kJqXEPXMsE0.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/4c5bcfec8584af0d967f1ab10179ca4b-Abstract.html.

Emiel Hoogeboom, Didrik Nielsen, Priyank Jaini, Patrick Forré, and Max Welling. Argmax flows and multinomial diffusion: Learning categorical distributions. In *Advances in Neural Information Processing Systems*, volume 34, pp. 12454–12465, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/67d96d458abdef21792e6d8e590244e7-Abstract.html.

Emiel Hoogeboom, Víctor Garcia Satorras, Clément Vignac, and Max Welling. Equivariant diffusion for molecule generation in 3D. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 8867–8887. PMLR, 2022. URL https://proceedings.mlr.press/v162/hoogeboom22a.html.

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 noncanonical amino acids using flow matching. *Bioinformatics*, 42(9):btag654, 2026. doi: 10.1093/bioinformatics/btag654. URL https://doi.org/10.1093/bioinformatics/btag654.

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.

Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver++: Fast solver for guided sampling of diffusion probabilistic models. *Machine Intelligence Research*, 22 (4):730–751, 2025.

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. In *Advances in Neural Information Processing Systems*, volume 34, pp. 29287–29303, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/f51338d736f95dd42427296047067694-Abstract.html.

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. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 27611–27629. PMLR, 2023. URL https://proceedings.mlr.press/v202/peng23b.html.

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.

Yu Rong, Yatao Bian, Tingyang Xu, Weiyang Xie, Ying Wei, Wenbing Huang, and Junzhou Huang. Self-supervised graph transformer on large-scale molecular data. In *Advances in Neural Information Processing Systems*, volume 33, pp. 12559–12571, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/94aef38441efa3380a3bed3faf1f9d5d-Abstract.html.

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 Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 2256–2265. PMLR, 2015. URL https://proceedings.mlr.press/v37/sohl-dickstein15.html.

Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. In *Advances in Neural Information Processing Systems*, volume 32, pp. 11895–11907, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/3001ef257407d5a371a96dcd947c7d93-Abstract.html.

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=AkIttWFnxS9.

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021b. URL https://openreview.net/forum?id=PxTIG12RRHS.

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

Clément Vignac, Nagham Osman, Laura Toni, and Pascal Frossard. MiDi: Mixed graph and 3D denoising diffusion for molecule generation. In *Machine Learning and Knowledge Discovery in Databases: Research Track*, volume 14170 of *Lecture Notes in Computer Science*, pp. 560–576. Springer Nature Switzerland, 2023. doi: 10.1007/978-3-031-43415-0_33. URL https://doi.org/10.1007/978-3-031-43415-0_33.

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 Diffusion and Equivariance Preliminaries

This appendix collects the standard formulations referenced in Section 3. None of it is specific to UNAAGI.

## A.1 Denoising Diffusion

Diffusion models have emerged as a powerful class of generative models (Sohl-Dickstein et al., 2015; Ho et al., 2020; Song & Ermon, 2019; 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

$$q(\mathbf{x}_{1:T} \mid \mathbf{x}_0) = \prod_{t=1}^{T} q(\mathbf{x}_t \mid \mathbf{x}_{t-1}), \tag{2}$$

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

$$p_\theta(\mathbf{x}_{0:T}) = p(\mathbf{x}_T) \prod_{t=1}^{T} p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t), \tag{3}$$

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

## A.2 Training Objective

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.

## A.3 Equivariance

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 rotations, reflections, and translations in 3D. 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 $\mathbf{t} \in \mathbb{R}^3$.

## A.4 Message Passing

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 by MLPs over concatenated scalar and edge features and then partitioned into components that update scalar, vector, and positional embeddings equivariantly. This design preserves E(3)-equivariance throughout the network and enables joint modeling of geometry, connectivity, and chemistry during denoising.

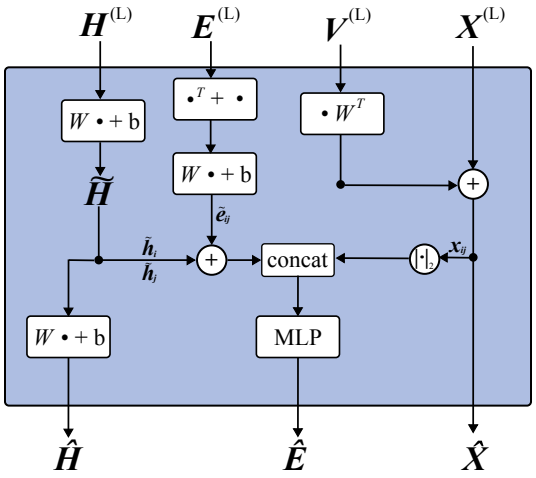



Figure A1: Backbone architecture of UNAAGI, closely following EQGAT-diff (Le et al., 2024).

# B Training Details

## B.1 Architecture Details

Our model architecture and hyperparameter choices closely follow EQGAT-diff (Le et al., 2024). The message-passing function at layer $l$ is defined as

$$\mathbf{m}_{ji}^{(l)} = \mathrm{MLP}\Big(\Big[\mathbf{h}_j^{(l)} \,;\, \mathbf{h}_i^{(l)} \,;\, \mathbf{W}_{\mathrm{e}_0}^{(l)}\mathbf{e}_{ji}^{(l)} \,;\, d_{ji}^{(l)} \,;\, d_j^{(l)} \,;\, d_i^{(l)} \,;\, \mathbf{p}_j^{(l)} \cdot \mathbf{p}_i^{(l)}\Big]\Big)$$

where ";" denotes concatenation and the MLP is a two-layer perceptron. The resulting message embedding is partitioned into subembeddings:

$$\mathbf{m}_{ji}^{(l)} = \big(\mathbf{a}_{ji}^{(l)}, \mathbf{b}_{ji}^{(l)}, \mathbf{c}_{ji}^{(l)}, \mathbf{d}_{ji}^{(l)}, \mathbf{s}_{ji}^{(l)}\big)^\top \in \mathbb{R}^K$$

The node, edge, vector, and coordinate updates are then computed as

$$\mathbf{h}_i^{(l+1)} = \mathbf{h}_i^{(l)} + \sum_j \frac{\exp(a_{ji}^{(l)})}{\sum_{j'} \exp(a_{j'i}^{(l)})} \mathbf{W}_h^{(l)}\mathbf{h}_j^{(l)} \qquad \text{and} \qquad \mathbf{e}_{ji}^{(l+1)} = \mathbf{W}_{e_1}^{(l)}\sigma(\mathbf{e}_{ji}^{(l)} + \mathbf{d}_{ji}^{(l)}),$$

$$\mathbf{v}_i^{(l+1)} = \mathbf{v}_i^{(l)} + \frac{1}{N}\sum_j \mathbf{x}_{ji,n} \otimes \mathbf{b}_{ji}^{(l)} + \big(\mathbf{1} \otimes \mathbf{c}_{ji}^{(l)}\big) \odot \mathbf{v}_j^{(l+1)}\,\mathbf{W}_v^{(l)},$$

$$\mathbf{x}_i^{(l+1)} = \mathbf{x}_i^{(l)} + \frac{1}{N}\sum_j s_{ji}^{(l)}\,\mathbf{x}_{ji,n}.$$

where $\mathbf{1} = (1,1,1)^\top$ and $\sigma$ is the SiLU activation function. After $L$ layers of message passing, the embeddings are transformed as shown in Figure A1 to obtain the final representations. Categorical atomic features, including hybridization states, formal charges, atom types, aromaticity, ring membership, and atomic degree, are concatenated to form $\hat{H}$. We use the final-layer prediction $\hat{x}_0 = (\hat{\mathbf{X}}, \hat{\mathbf{H}}, \hat{\mathbf{E}})$ for denoising.

## B.2 Hyperparameters

In all experiments, the forward diffusion process uses 500 timesteps during training. At inference time, we use the 20-step accelerated sampler described in Section 4.2. Scalar feature dimensions are set to 256, vector feature dimensions to 64, and edge feature dimensions to 32. We use 7 layers of message passing, resulting in a total of approximately 3.6M trainable parameters.

For optimization, we use Adam with a learning rate of $5 \cdot 10^{-4}$ and weight decay of $1 \cdot 10^{-4}$. Gradients are clipped to a maximum norm of 10. We maintain an exponential moving average (EMA) of the model weights with a decay of 0.9999. The training batch size is 8.

For virtual-node padding, we set $N_{\max}^{\text{train}} = 5$ and sample $n_{\text{virt}} \sim U(0, 5)$ for each residue during training. Here, $N_{\max}^{\text{train}}$ denotes the maximum number of virtual atoms added to a ground-truth side chain. During sampling, $N_{\max}^{\text{sample}} = 15$ denotes the number of nodes in the initial graph. This value covers all amino acids in the benchmark: the largest, 2NP, contains 11 side-chain atoms. During training, side-chain graph sizes range from 0 nodes (Gly with no virtual nodes) to 15 nodes (Trp with five virtual nodes), exposing the model to the graph size used at inference.

### B.3 Details on Training Data Containing NCAAs

We construct the training dataset by filtering PDB entries deposited before January 26, 2026. We retain structures with resolution finer than 2 Å, determined by X-ray crystallography or electron microscopy, and containing full-atom coordinates. We remove hydrogen atoms and isolated water molecules, yielding 102,048 PDB structures for training.

As noted in the dataset composition section, we augment the training set with independent NCAA data from Ilardo et al. (2019) and SwissSidechain (Gfeller et al., 2013). The dataset of Ilardo et al. (2019) contains 1,913 distinct NCAAs, and SwissSidechain contains 230 NCAAs. These amino acids are used for training as individual molecules without associated structural context.

For the selected PDB structures containing NCAAs, we assemble a dedicated corpus of experimentally determined structures containing NCAAs or stabilizing post-translational modifications (PTMs) *in native structural context*. This dataset, denoted NCAA_PDB, comprises 14,937 high-resolution PDB structures spanning 61,937 modified-residue sites.

## C Full ProteinGym Comparison

Figure C1 reports UNAAGI's Spearman correlation against experimental DMS measurements across all 25 ProteinGym assays, alongside the baselines discussed in Section 4.4. This is the canonical-substitution sanity check described in Section 4.5.2; UNAAGI is not designed for canonical sequence-based prediction, and the comparison is included for completeness rather than as a headline result.

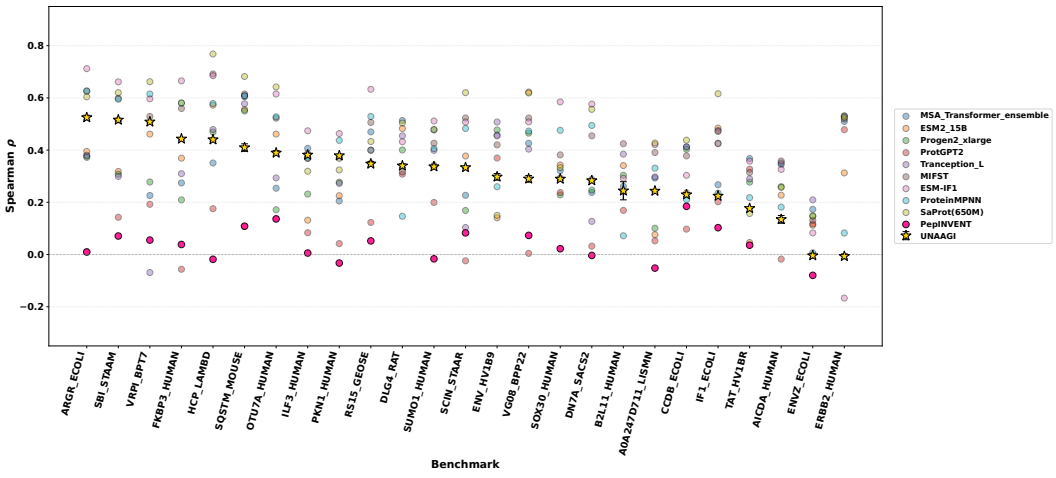



Figure C1: Results on 25 ProteinGym assays, evaluated using the Spearman correlation with experimental DMS measurements. UNAAGI outperforms its closest methodological baseline, PepINVENT, across the evaluated assays.

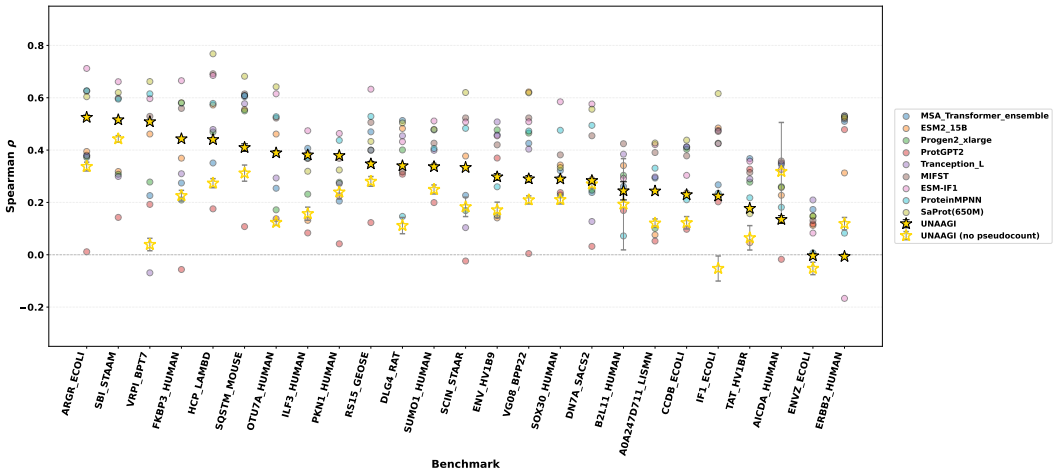



Figure D1: Robustness of ProteinGym performance to the treatment of unsampled amino acids. Spearman correlations are shown for UNAAGI using the add-one pseudocount employed in the main evaluation (black stars) and after excluding unsampled substitutions (yellow outlined stars), alongside sequence- and structure-based baselines.

# D Additional Robustness Analyses

## D.1 Held-out Residue Statistics

Sample sizes for Table 1. Reconstruction (wt $= X$): $n = 316$, with 96/91/42/87 sites for VAL/SER/PHE/ILE. Design at favorable sites (wt $\neq X$): $n = 3{,}086$, with 785/704/778/819 sites. Selectivity compares the favorable-site rate ($n = 3{,}402$) with the harmful-site rate ($n = 1{,}833$, with 434/510/492/397 sites). "% kept" is the ratio between the held-out and full-model rates; the full model achieves $100\%$ reconstruction throughout.

All differences between the held-out and full models at favorable sites are significant (Fisher exact test, $p$: VAL $8.6\text{e}{-6}$, SER $2.1\text{e}{-10}$, PHE $3.0\text{e}{-27}$, ILE $9.2\text{e}{-104}$). Pooled Wilson $95\%$ confidence intervals are $[0.927, 0.973]$ for reconstruction, $[0.687, 0.719]$ for design, $[0.711, 0.741]$ for favorable sites and $[0.599, 0.644]$ for harmful sites.

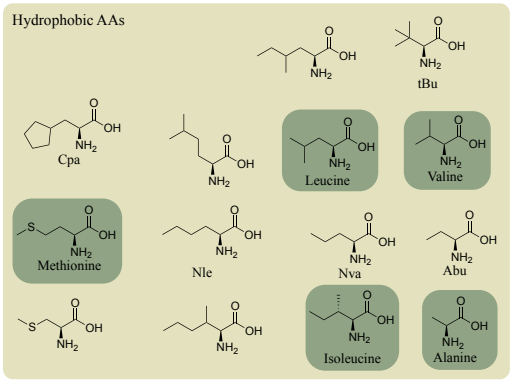
## D.2 Robustness to the Pseudocount for Unsampled Amino Acids

We assess whether the add-one pseudocount assigned to unsampled amino acids drives UNAAGI's ProteinGym performance by recomputing correlations after excluding substitutions that are never sampled. Figure D1 compares the original evaluation with this alternative treatment across all 25 ProteinGym assays.

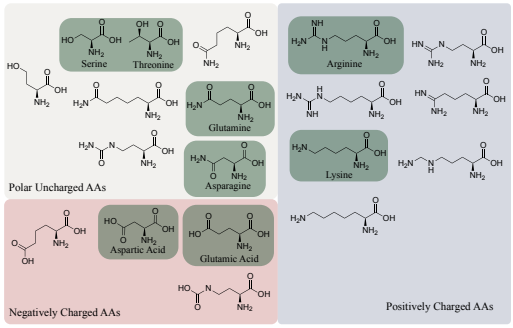
# E Additional Qualitative Results

## E.1 Analysis of Wild-Type Coverage

Because both UNAAGI and PepINVENT sample residues atom by atom, they must be able to reconstruct wild-type residues reliably. We quantify this ability using the *wild-type coverage rate*, defined as the fraction of positions in an assay for which the wild-type residue is sampled at least once. Table E1 reports the average across all benchmarks, and Figure E2 provides assay-level results. UNAAGI achieves substantially higher coverage, whereas PepINVENT rarely recovers the wild-type residue.




(a) Hydrophobic amino acids sampled by UNAAGI, including both canonical and non-canonical examples (not an exhaustive set).




(b) Polar and charged amino acids sampled by UNAAGI, including both canonical and non-canonical examples (not an exhaustive set).

Figure E1: Qualitative inspection of sampled non-canonical amino acids. Canonical amino acids are highlighted with a green background.

Table E1: Comparison of wild-type coverage rates for UNAAGI and PepINVENT, averaged across all benchmarks, including those introduced in Section 4.5.1.

| Model | Wild-Type Coverage Rate |
|---|---|
| UNAAGI (ours) | **0.9368** |
| PepINVENT | 0.2365 |

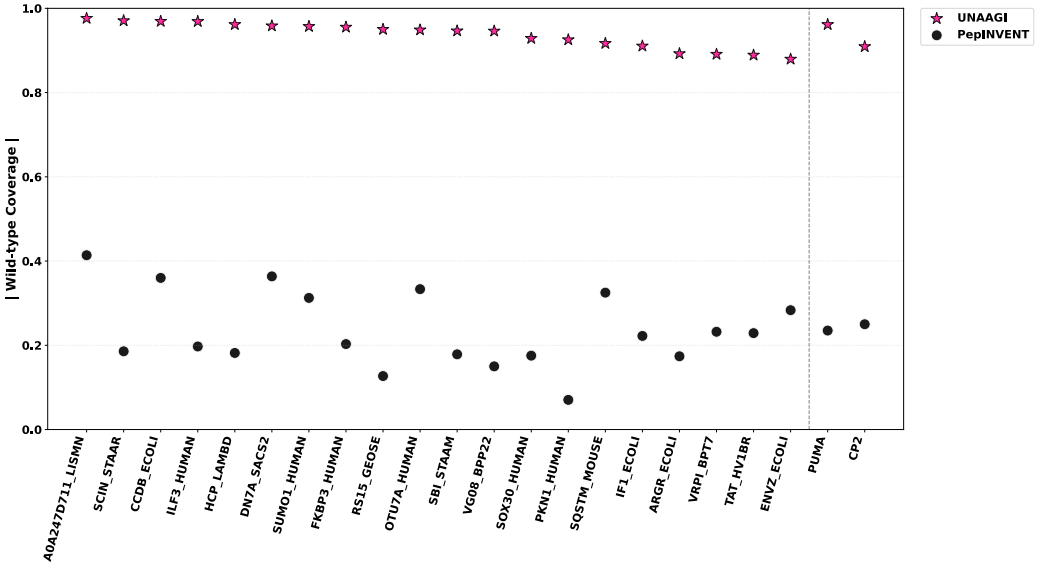



Figure E2: Wild-type coverage rate of UNAAGI and PepINVENT across benchmark assays.

### E.2 Comprehensive Analysis of NCAA Coverage

We examine how UNAAGI's coverage of the NCAAs in the benchmark of Rogers et al. (2018) relates to their chemical similarity to canonical amino acids. Figure E3 summarizes these relationships as a molecular graph-edit network. The sampled NCAAs tend to be structurally similar to canonical residues, consistent with the model's ability to extend beyond the canonical vocabulary to chemically

proximal substitutions. Coverage remains limited for other classes, including *N*-methylated residues and residues with more complex ring systems.

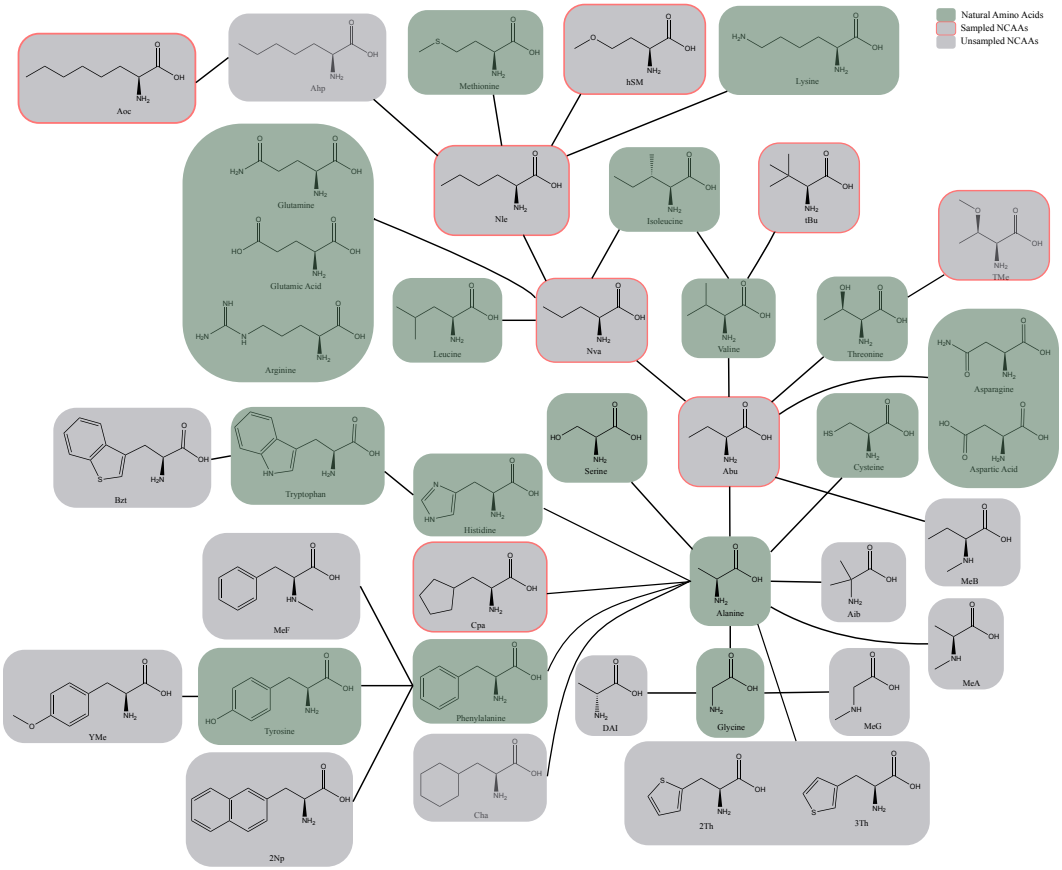



Figure E3: Relationships between canonical amino acids and the NCAAs in the benchmark of Rogers et al. (2018). Canonical amino acids are shown with a green background, NCAAs with a gray background, and NCAAs sampled by UNAAGI are highlighted with a red outline. Edges connect amino acids that differ by one graph edit.

To complement this graph-based analysis, we obtain molecular embeddings for the amino acids in the benchmark using the encoder of Rong et al. (2020) and project them into two dimensions using UMAP (Figure E4). The projection reveals chemically related groups, with sampled NCAAs concentrated among residues with aliphatic side chains and related modifications. In contrast, NCAAs containing aromatic rings or *N*-methyl modifications are largely unsampled. This qualitative pattern is consistent with selective coverage of NCAA chemical space, rather than uniform generalization across all chemical classes; the projection itself does not quantify mutational-effect prediction accuracy.

### E.3 Quality Control of Generated Amino Acids with PoseBusters

We evaluate the chemical and geometric validity of 13,472 sampled amino acids using PoseBusters. A sample is considered strictly valid if it passes every included criterion. Figure E5 reports the strict validity rate for each benchmark, yielding an overall pass rate of 77.2%. Figure E6 disaggregates these results by criterion. The generated molecules almost universally pass sanitization, ring-planarity, and molecular-loadability checks, while most failures arise from deviations in bond lengths and bond angles. These results indicate that UNAAGI generally produces chemically coherent molecular graphs, with local geometric accuracy representing the primary remaining source of invalidity.

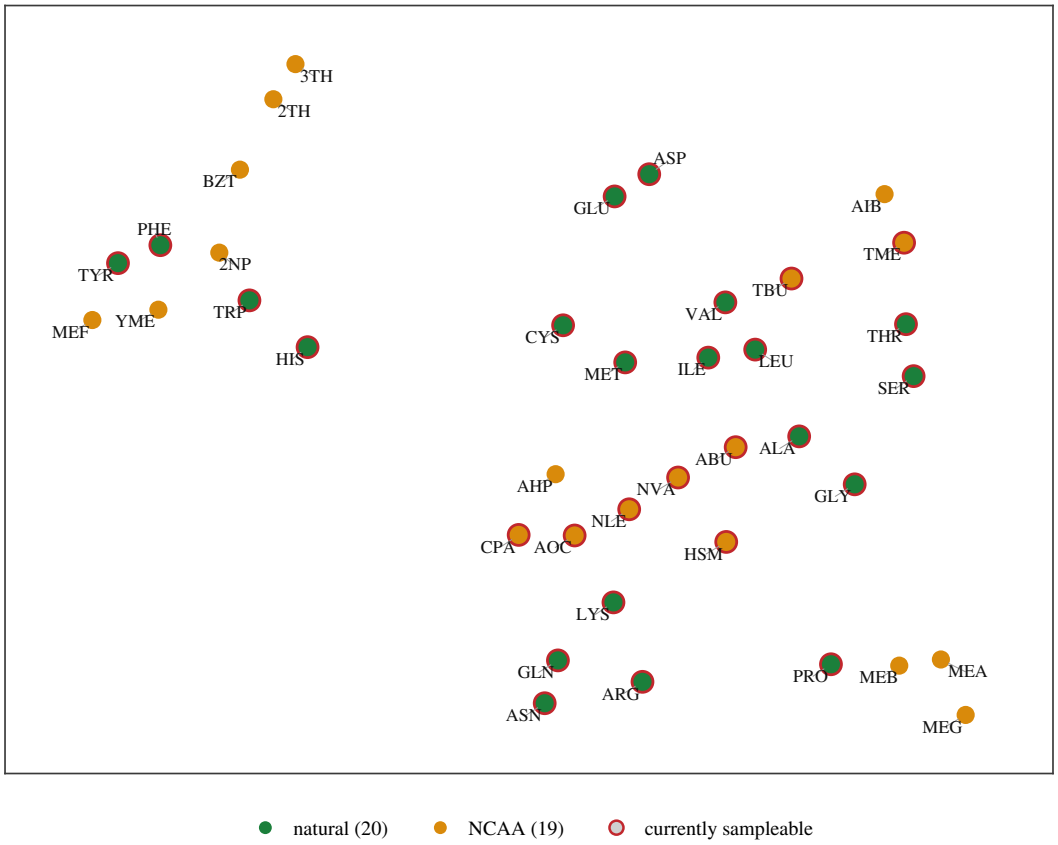



Figure E4: Chemical-space visualization of amino-acid coverage by UNAAGI. Molecular embeddings from the encoder of Rong et al. (2020) are projected into two dimensions using UMAP. Each point represents an amino acid in the benchmark: green denotes the 20 canonical amino acids, orange denotes 19 NCAAs, and red outlines indicate amino acids sampled by UNAAGI. Sampled NCAAs are concentrated among residues with aliphatic side chains and related modifications, whereas aromatic and *N*-methylated NCAAs are largely unsampled.

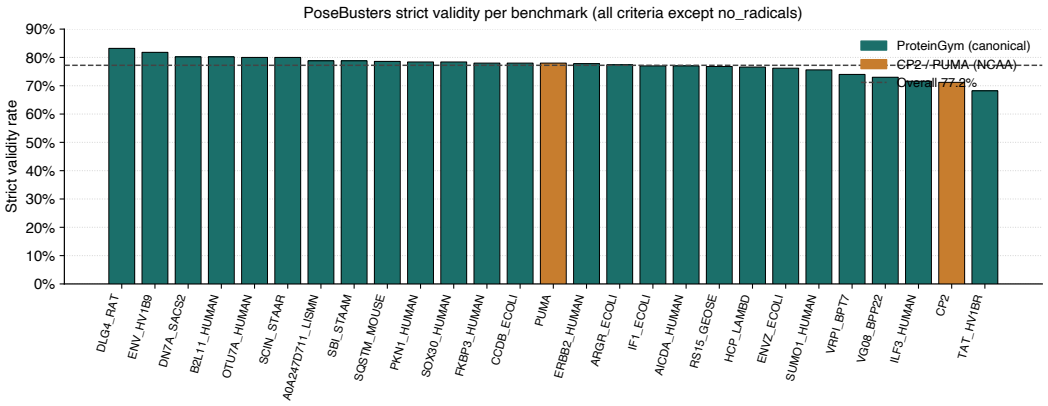



Figure E5: Strict PoseBusters validity rate for each benchmark assay. A generated amino acid is counted as valid only if it passes all included PoseBusters criteria. ProteinGym assays are shown in teal, and the CP2 and PUMA NCAA benchmarks are shown in orange. The dashed line indicates the overall validity rate of 77.2% across all 13,472 samples.

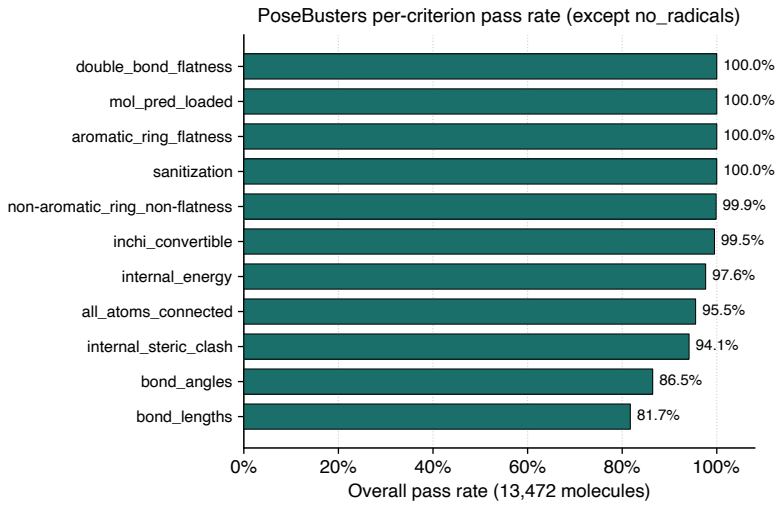



Figure E6: PoseBusters pass rate for each included validity criterion, aggregated across all 13,472 generated amino acids. The `no_radicals` criterion is excluded. Most samples pass chemical-validity, connectivity, energy, and steric-clash checks, whereas bond-length and bond-angle criteria account for the largest fractions of failures.

### E.4 Evolutionary Depth of the CP2 and PUMA Receptors

To characterize the evolutionary information available for the two NCAA mutational-effect benchmarks, we quantify the depth of the multiple-sequence alignments (MSAs) for the CP2 and PUMA receptors. Figure E7 reports both the total number of retrieved sequences and the effective depth after sequence reweighting at an 80% identity threshold. Both receptors have substantial evolutionary coverage, although the CP2 alignment is considerably deeper than the PUMA alignment. This disparity may partly explain the difference in AlphaFold3's mutational-effect prediction performance between the two receptors, as AlphaFold3 uses evolutionary information from MSAs.

## F Clarification on Evaluation

UNAAGI is evaluated zero-shot with respect to mutational-effect prediction. Training uses molecular structures and residue identities but does not include DMS measurements, $\Delta\Delta G$ values, fitness

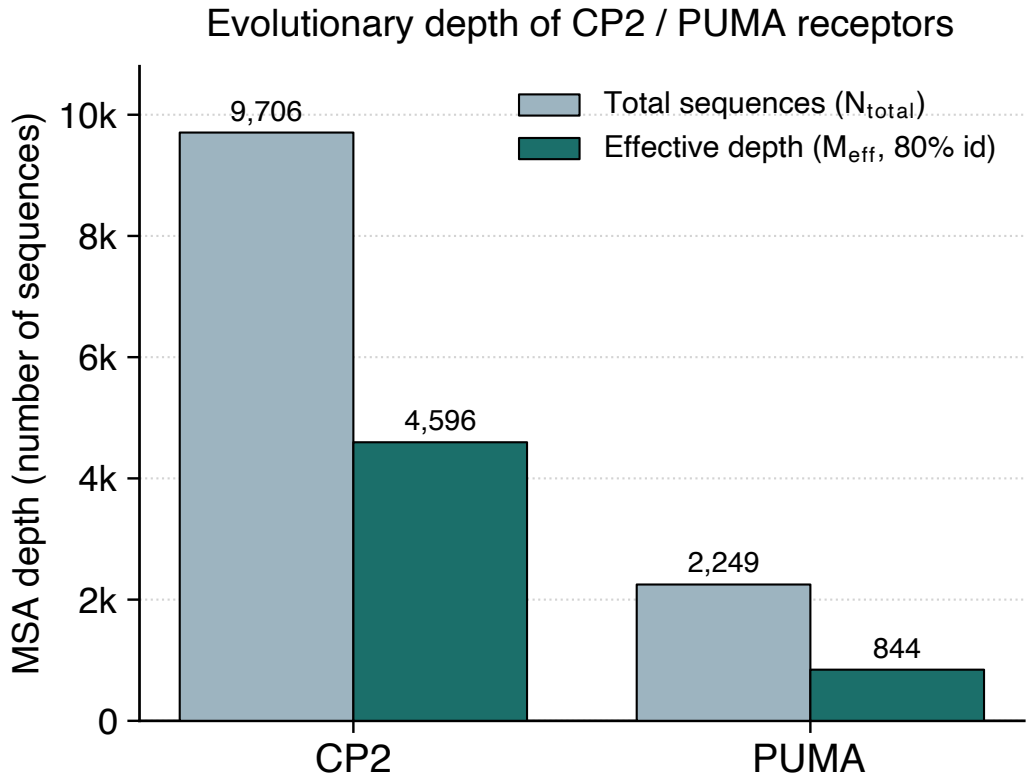



Figure E7: Evolutionary depth of the CP2 and PUMA receptor MSAs. Bars show the total number of sequences ($N_{\text{total}}$) and the effective sequence depth ($M_{\text{eff}}$) after reweighting at an 80% sequence-identity threshold.

scores, or task-specific fine-tuning. As with other pretrained zero-shot predictors, benchmark proteins may occur in the structural training corpus. Our evaluation therefore tests whether knowledge learned from structural data without mutational-effect labels transfers to mutational-effect prediction, rather than whether the model generalizes to entirely unseen protein structures.