Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Site-Conditioned Edit Chemistry for Cyclic Peptide Permeability Modeling

Minrui Chen
Kyushu University
Guanjie Zou
The University of Tokyo
Minqi Chen
West China Hospital, Sichuan University
Bowen Song
Kyushu University
Hiroto Saigo
saigo@inf.kyushu-u.ac.jp

Kyushu University

Research Article

Keywords: cyclic peptides, permeability prediction, edit chemistry, site-conditioned modeling, chemical descriptors, cheminformatics, mechanistic modeling

Posted Date: May 18th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9541490/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Site-Conditioned Edit Chemistry for Cyclic Peptide Permeability Modeling

Minrui Chen[1], Guanjie Zou[2, 3], Minqi Chen[4,5], Bowen Song[1], Hiroto Saigo[1,2*]

[1]Department of Information Science and Technology, Kyushu University, Motoka, Nishi-ku, 8190395, Fukuoka, Japan.
[2]Center for Advanced Intelligence Project (AIP), RIKEN, Tokyo, 103-0027, Japan.
[3]Department of Computational Biology and Medical Science, The University of Tokyo, Chiba, 277-8563, Japan.
[4]Department of Thoracic Surgery and Institute of Thoracic Oncology, West China Hospital, Sichuan University, Chengdu, Sichuan, 610041, China.
[5]Western China Collaborative Innovation Center for Early Diagnosis and Multidisciplinary Therapy of Lung Cancer, Sichuan University, Chengdu, Sichuan, 610041, China.

*Corresponding author(s). E-mail(s): saigo@inf.kyushu-u.ac.jp;
Contributing authors: chen.minrui.978@s.kyushu-u.ac.jp;
8923236439@edu.k.u-tokyo.ac.jp; 2019141230152@stu.scu.edu.cn;
song.bowem.884@s.kyushu-u.ac.jp;

Abstract

Localized chemical edits such as noncanonical substitutions, post-translational modifications, cyclization, and linker-mediated conjugation can strongly alter cyclic peptide permeability. However, most predictive studies do not explicitly separate the contribution of edit chemistry itself from the contribution of assigning that chemistry to a specific peptide site. Here, we study this question in a deliberately restricted curated known-site regime, in which modification sites can be resolved after rule-based preprocessing, and formulate the resulting task as chemistry–site contribution analysis under explicit site observability. Using the assay-harmonized CycPeptMPDB-PAMPA dataset, we compare sequence-only, chemistry-only, and site-aware descriptor models through controlled feature-block

comparisons. We further interrogate the learned signal using wrong-site and coarse-site controls, graded site perturbation, feature ablation, generalized additive modeling, anchor-sensitive kernel regression, parser auditing, and a harder sequence-cluster split.
Across the standard random split, chemistry-aware descriptor modeling substantially outperforms sequence-only prediction, indicating that edit chemistry provides the dominant predictive signal for passive permeability in this dataset. Adding site-aware features yields a reproducible but quantitatively modest refinement, improving test $\boldsymbol{R^2}$ from 0.433 to 0.472. Perturbation controls show that this refinement is strongly position-sensitive: wrong-site, coarse-site, and graded site-shift settings all degrade performance while holding edit chemistry fixed. Feature ablation further indicates that the strongest individually effective site-aware component in the current descriptor family is site-position statistics, whereas anchor-site residue-composition and residue-property summaries contribute mainly in combination with positional information. However, under a harder sequence-cluster split, the site-aware gain nearly vanishes, with test $\boldsymbol{R^2}$ changing only from 0.268 to 0.270. This result indicates that the current site-conditioned benefit is context-dependent and has limited transfer across stronger scaffold shift.
We therefore position the present study not as a universal peptide-property predictor and not as a solution to uncertain-site inference, but as a mechanistic analysis of chemistry-versus-site contributions in a curated known-site cyclic peptide permeability setting. The main contribution is to quantify the present scale, locality, and boundary conditions of explicit site conditioning within an interpretable descriptor framework. The results support a chemistry-dominant view of permeability prediction, in which site information provides a local within-regime refinement rather than a broadly transferable predictive advantage under scaffold shift.

Keywords: cyclic peptides, permeability prediction, edit chemistry, site-conditioned modeling, chemical descriptors, cheminformatics, mechanistic modeling

1 Introduction

Cyclic peptides are increasingly explored as drug candidates because they can combine high target affinity with substantial opportunities for physicochemical optimization [1, 2]. Their optimization often involves localized chemical interventions, including noncanonical amino acid replacement, post-translational modification, cyclization, stapling, and linker-mediated conjugation [1–3]. These edits can alter permeability, conformational preference, protease resistance, and target engagement. Consequently, cyclic peptide property prediction is rarely determined by sequence alone; the relevant signal may depend jointly on the host peptide scaffold, the chemistry of the installed edit, and the peptide site at which that chemistry is assigned.

Recent predictive studies have combined sequence and chemistry through descriptor fusion, sequence tagging, graph representations, and joint sequence–chemistry

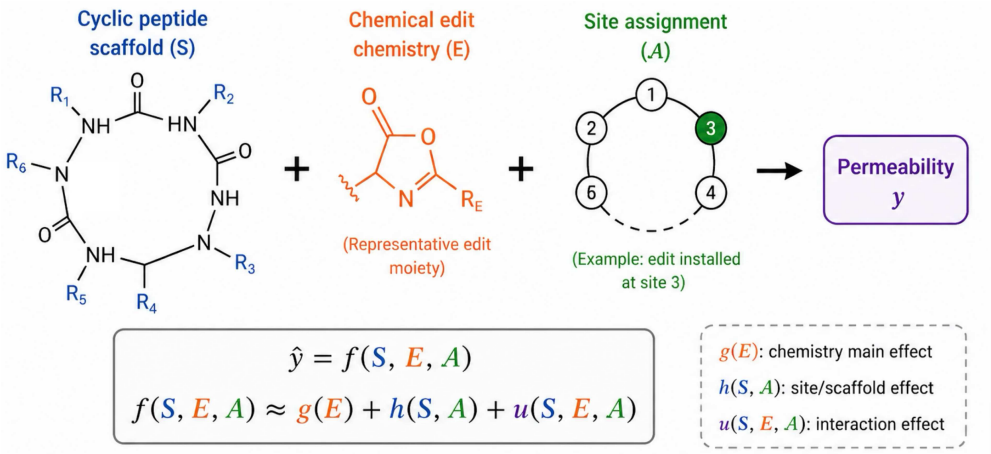


Fig. 1 Problem framing for chemistry–site contribution analysis in cyclic peptide permeability modeling. A cyclic peptide record is represented conceptually by a peptide scaffold S, an edit chemistry component E, and an observable site assignment A, which together define the permeability prediction problem $\hat{y} = f(S, E, A)$. The conceptual decomposition $f(S, E, A) \approx g(E) + h(S, A) + u(S, E, A)$ is used to organize the distinction between chemistry-associated main effects, site-indexed contributions, and chemistry–site interactions. This decomposition is an interpretive and representational framing rather than a claim of strict causal identifiability.

encoders [4–9]. These approaches can improve predictive performance, but an important interpretive ambiguity often remains: when performance improves after adding edit-related information, is the gain driven mainly by the chemistry of the edit itself, by the site-indexed placement of that chemistry in the peptide context, or by increased feature capacity and model flexibility? This distinction is especially relevant for modified peptides, because the same chemical moiety may plausibly have different effects when assigned to different positions on a cyclic peptide scaffold.

In this study, we address this question in a deliberately restricted setting. We do not attempt to infer uncertain modification sites from raw records, and we do not propose a general-purpose framework for all peptide-property prediction tasks. Instead, we focus on a curated known-site regime in which localized edits can be mapped to explicit peptide sites after rule-based preprocessing. Within this regime, we ask a focused representational question: given a peptide scaffold S, edit chemistry E, and observable site assignment A, how much additional predictive information is associated with explicit site conditioning once edit chemistry has already been represented? This problem framing is illustrated in Fig. 1.

This framing leads to three concrete questions. First, does chemistry-aware modeling substantially improve over sequence-only modeling? Second, when edit chemistry is held fixed, does site-indexed information retain measurable predictive value compared with degraded positional alternatives? Third, which components of a site-aware representation are responsible for any observed gain?

To answer these questions, we study the assay-harmonized CycPeptMPDB-PAMPA dataset and compare sequence-only, chemistry-only, and site-conditioned

descriptor models through controlled feature-block comparisons. We then interrogate the fitted signal using wrong-site and coarse-site controls, graded site perturbation, feature ablation, parser auditing, auxiliary interpretable models, and a harder sequence-cluster split. This design is intended to separate three issues that are often conflated in peptide-property modeling: the predictive contribution of edit chemistry, the incremental contribution of explicit site assignment, and the robustness of that contribution under stronger scaffold shift.

The central finding is intentionally specific rather than universal. In the present dataset, edit chemistry provides the dominant predictive signal for passive permeability. Explicit site conditioning provides a smaller, position-sensitive refinement under standard random-split evaluation, but this refinement is strongly attenuated under a harder sequence-cluster split. Thus, the observed site-conditioned effect is best interpreted as a within-regime, sequence-indexed descriptor refinement rather than as evidence that the current representation captures general scaffold-transferable or three-dimensional structural mechanisms.

This distinction is important for both interpretation and use. The present study should be viewed as a mechanistic and representational analysis in a curated known-site regime, not as a claim that the current descriptor family solves cyclic peptide permeability prediction in general. In particular, the present encoding is based on one-dimensional sequence position and two-dimensional chemistry descriptors; it does not directly encode conformational ensembles, chameleonic behavior, intramolecular hydrogen-bond organization, or solvent-exposure patterns, all of which are known to influence cyclic peptide permeability [10–14]. Accordingly, any position-sensitive effect reported here should be interpreted within the present sequence-indexed representation regime rather than as a direct proxy for three-dimensional spatial proximity or conformation-aware local environment.

2 Computational Methods

2.1 Study Design

We investigated cyclic peptide permeability prediction in a curated known-site setting, where localized chemical edits could be associated with explicit peptide sites after rule-based preprocessing. The study was organized around one assay-harmonized dataset, CycPeptMPDB-PAMPA, and three primary predictor families: sequence-only models, chemistry-aware but site-agnostic descriptor models, and site-conditioned descriptor models.

The main goal was not to propose a universal cyclic peptide permeability predictor. Instead, we asked a narrower representational question: after localized edit chemistry has been encoded, how much additional predictive information is contributed by explicit site assignment? To answer this question, we compared matched descriptor families and then interrogated the fitted representations using site perturbation controls, feature ablation, interpretable auxiliary models, parser auditing, and a harder sequence-cluster split.

Throughout the study, site-aware effects are interpreted as sequence-indexed descriptor effects within the present representation regime. They should not be

interpreted as direct evidence of three-dimensional spatial proximity, conformational organization, intramolecular hydrogen-bond topology, or causal biochemical site effects.

2.2 Dataset Curation

The primary dataset was CycPeptMPDB-PAMPA [15], selected because it provides assay-comparable cyclic peptide permeability measurements together with structured edit records that can often be resolved to peptide sites. PAMPA is used here as a proxy assay for passive permeability characterization [16, 17].

The raw release contained 7298 records. All 7298 records were schema-parseable, and 7224 samples (99.0%) were successfully resolved to site-level anchor annotations by the rule-based parser described below. The final usable dataset therefore contained 7224 samples derived from 1325 unique peptide sequences, corresponding to an average of 5.5 derivatives per unique sequence. The 74 excluded records were those for which site-level resolution failed.

The default train/validation/test split was 70/15/15, yielding 5056, 1084, and 1084 samples, respectively. Because the dataset contains multiple derivatives from shared or closely related peptide scaffolds, this random split is interpreted as a within-regime evaluation rather than as a strict test of scaffold-level transfer.

To evaluate robustness under stronger sequence distribution shift, we additionally constructed a 70% sequence-identity cluster split. This split produced 4801 training samples, 1306 validation samples, and 1117 test samples from 352, 287, and 125 clusters, respectively. The sequence-cluster split was used as a stress test for whether chemistry-aware and site-conditioned gains persisted beyond scaffold-nearby derivatives.

Table 1 Primary dataset census for CycPeptMPDB-PAMPA.

Quantity	Value	Notes
Raw records	7298	downloaded release
Schema-parseable records	7298	100% schema parse success
Site-resolved records	7224	99.0% anchor mappable
Final usable samples	7224	retained after filtering
Unique sequences	1325	average 5.5 derivatives per sequence
Train/validation/test (random)	5056/1084/1084	70/15/15 split
Train/validation/test (cluster)	4801/1306/1117	70% sequence-identity cluster split

2.3 Unified Sample Representation

Each retained sample was normalized into

$$x = (\texttt{sample_id}, S, E, A, Y, M),$$

where S denotes the peptide sequence, $E = [e_1, \ldots, e_K]$ is the list of localized chemical edits, A denotes the resolved anchor annotations, Y is the permeability label, and M

contains assay and provenance metadata. Each edit was represented as

$$e_k = (\texttt{edit_type}_k,\ \texttt{anchor}_k,\ \texttt{chem_rep}_k,\ \texttt{rule_id}_k).$$

This representation separates peptide scaffold, edit chemistry, and observable site assignment. It therefore allows sequence-only, chemistry-only, and site-conditioned models to be compared through controlled feature-block comparisons. We do not assume that the feature blocks contain equal information. Instead, the comparisons quantify how much predictive refinement is obtained when explicit site-indexed and attachment-aware descriptors are added to chemistry descriptors under matched training and evaluation protocols.

2.4 Rule-Based Site Mapping and Parser Auditing

Anchor annotations were assigned using a versioned rule-based parser that maps local edits to peptide sequence positions whenever a site-level assignment can be resolved from the curated record representation. Single-site edits were encoded as a single anchor index, whereas pairwise attachment events such as cyclization were represented through multi-anchor metadata and anchor-pair statistics. Ambiguous or invalid cases that could not be resolved to a site-level representation were excluded during preprocessing.

Because parser validity is a structural assumption of the analysis, we treated the parser as part of the analysis substrate rather than as a minor preprocessing step. Automated consistency checks were used to identify anchor out-of-bounds cases, duplicate anchors, multi-anchor inconsistencies, edit-type incompatibilities, and unresolved mapping cases. We additionally constructed a stratified manual audit set of 150 records spanning single-anchor events, multi-anchor events, major edit families, and structural edge cases.

In the full dataset, automated checks identified a per-edit issue rate of 9.55%. Most flagged cases corresponded to expected global-anchor or structurally special cases rather than obvious catastrophic mapping failures. A conservative high-confidence subset retained 68.1% of usable samples. The parser audit is therefore used to define the confidence boundary of the site-conditioned analysis. It is not used to claim perfect site recovery or parser independence.

2.5 Descriptor Design

The fixed-length descriptor vector was organized into three chemically interpretable blocks: Group A, global edit-chemistry descriptors; Group B, explicit site-indexed descriptors; and Group C, attachment-aware and multi-edit context descriptors. In the interpretation below, Group A represents chemistry-only information, Group B represents stricter site-indexed information, and Group C represents broader attachment and multi-edit context. Consequently, improvements from the full A+B+C model are interpreted as full-context site-conditioned gains rather than as pure site-assignment gains.

Table 2 Parser-audit flags for site-mapping uncertainty.

Flag	Definition	Analysis treatment
Anchor out of bounds	Anchor index outside the peptide length.	Excluded as invalid.
Duplicate anchor	Repeated anchor within one edit event.	Flagged; retained only if chemically interpretable.
Multi-anchor inconsistency	Anchor set incompatible with a pairwise or multi-site event.	Flagged as structurally ambiguous.
Global-anchor case	Edit represented without a unique local site.	Treated as a special case; excluded from strictly local analyses when unresolved.
Edit-type mismatch	Edit type inconsistent with the resolved anchor structure.	Flagged for rule-level audit.
Unresolved mapping	Site-level assignment cannot be recovered.	Excluded from the site-resolved set.

Group A: global edit-chemistry descriptors.

Group A consisted of 10 descriptor types computed from edit-level chemical representations using RDKit: molecular weight, octanol/water partition coefficient (LogP), topological polar surface area, hydrogen-bond donor count, hydrogen-bond acceptor count, rotatable bond count, aromatic ring count, aliphatic ring count, fraction of sp^3 carbons, and heteroatom count. These descriptor types are motivated by prior work linking lipophilicity, polarity, hydrogen-bonding capacity, and conformational restriction to cyclic peptide permeability [10, 14, 18, 19]. For peptides with multiple edits, descriptor values were aggregated using order-invariant summary operations.

Group B: explicit site-indexed descriptors.

Group B comprised three subblocks: B1, site-position statistics; B2, anchor-site residue-composition summaries; and B3, anchor-site residue-property summaries.

Let the peptide length be L and the resolved anchor set be

$$A = \{a_1, \ldots, a_m\},$$

where $a_j \in \{1, \ldots, L\}$. The normalized anchor position was defined as

$$p_j = \frac{a_j - 1}{L - 1}.$$

For peptides with $L = 1$, the normalized position was set to zero by convention. B1 encoded order-invariant summaries over $\{p_j\}_{j=1}^{m}$, including anchor count, anchor density, minimum position, maximum position, mean position, and standard deviation of normalized anchor position.

For B2, let $\phi_{\mathrm{aa}}(r)$ denote a 20-dimensional one-hot encoding of the residue type at site r. The anchor-site residue-composition summary was defined as

$$\mathbf{b}_{\mathrm{comp}} = \frac{1}{m} \sum_{j=1}^{m} \phi_{\mathrm{aa}}(a_j).$$

This yields a 20-dimensional frequency vector over amino-acid identities at anchor sites.

For B3, let $\phi_{\mathrm{prop}}(r)$ denote a residue-property encoding at site r, summarizing coarse physicochemical classes such as hydrophobic, aromatic, polar, positively charged, negatively charged, sulfur-containing, conformationally restricted, hydrogen-bond-capable, or small residues. The anchor-site residue-property summary was defined as

$$\mathbf{b}_{\mathrm{prop}} = \frac{1}{m} \sum_{j=1}^{m} \phi_{\mathrm{prop}}(a_j).$$

The current implementation summarizes residue identity and residue class at the anchor positions themselves. It does not encode a sliding local sequence window around each anchor and does not encode three-dimensional neighborhood structure. Group B should therefore be interpreted as a sequence-indexed contextual summary rather than as a conformation-aware local environment model.

Group C: attachment-aware and multi-edit context descriptors.

Group C summarized edit family, edit-type diversity, edit-type counts, and multi-edit structural aggregation. These features included anchor-pair count, cyclization ring size when available, edit-anchor overlap, sequence length, a binary cyclicity indicator, and modification rate normalized by peptide length.

Group C provides attachment and multi-edit context rather than a pure site-assignment signal. For this reason, the ablation analysis reports both A+B, the stricter chemistry-plus-site comparison, and A+B+C, the full-context site-conditioned comparison. For peptides containing multiple edits, the descriptor was constructed using order-invariant aggregation so that the same multi-edit state yielded the same feature representation regardless of input ordering. This improves comparability across records but does not preserve full anchor topology or edit ordering beyond the explicitly defined pairwise summary features.

2.6 Mathematical Framing

We write the prediction problem as

$$\hat{y} = f(S, E, A),$$

where S is the peptide sequence, E is edit chemistry, and A is observable site assignment. Conceptually, we decompose this predictor as

$$f(S, E, A) = g(E) + h(S, A) + u(S, E, A),$$

where $g(E)$ denotes the chemistry-associated main effect, $h(S, A)$ denotes the site-indexed contribution, and $u(S, E, A)$ denotes chemistry–site interaction.

This decomposition is used as an organizing principle for descriptor construction, feature-block comparison, and interpretation. It is not a claim of strict causal identifiability, because edit chemistry, scaffold identity, and site placement are not experimentally randomized in the dataset.

2.7 Predictor Families

Sequence-only baseline.

The sequence-only baseline used amino-acid composition features derived exclusively from peptide sequence and was trained with XGBoost regression. This model was used as a minimal sequence reference for estimating how much signal could be obtained without explicit edit chemistry.

Chemistry-only baseline.

The chemistry-only baseline used Group A descriptors alone and was trained with XGBoost regression. This model tests how much predictive signal is carried by edit chemistry without explicit site-indexed descriptors.

Site-block descriptor model.

The stricter site-block descriptor model combined Group A and Group B descriptors and was trained with XGBoost regression. This model augments edit chemistry with explicit site-position statistics, anchor-site residue-composition summaries, and anchor-site residue-property summaries. It is used primarily in the ablation analysis to separate the contribution of explicit site-indexed descriptors from broader attachment-aware context.

Full-context site-conditioned descriptor model.

The full-context site-conditioned descriptor model combined Group A, Group B, and Group C descriptors and was trained with XGBoost regression. Unless otherwise specified, the term "site-aware" in the main predictive comparison refers to this full-context site-conditioned model. Because Group C contains sequence length, cyclicity, modification rate, ring-size, and edit-aggregation features, performance gains from this model are interpreted as full-context site-conditioned gains rather than pure site-assignment gains.

Position-only control.

A position-only XGBoost control was trained using anchor count and site-location summary statistics without chemistry descriptors. This control was included to test whether site-conditioned gains could be explained by positional shortcuts alone.

Auxiliary interpretable models.

To support mechanistic interpretation, we additionally trained an anchor-sensitive generalized additive model (GAM) and an anchor-sensitive kernel ridge model. These

models were not intended to be the primary predictive models. They were used to examine whether alternative inductive biases supported the same qualitative hierarchy observed in the main XGBoost descriptor models.

For the kernel model, the total kernel was defined as

$$K_{\text{total}}\big((e_i, a_i), (e_j, a_j)\big) = K_{\text{chem}}(e_i, e_j) \cdot K_{\text{anchor}}(a_i, a_j),$$

where K_{chem} measures chemistry similarity and K_{anchor} decays with site discrepancy. In the present implementation, the anchor kernel takes the radial form

$$K_{\text{anchor}}(a_i, a_j) = \exp\left(-\gamma_a d(a_i, a_j)^2\right),$$

where $d(a_i, a_j)$ is a normalized anchor-distance function. For single-anchor events,

$$d(a_i, a_j) = \left| \frac{a_i}{L_i} - \frac{a_j}{L_j} \right|,$$

and for multi-anchor events, d is defined by the average minimum matching distance over normalized anchor sets. The product kernel is used as an interpretable inductive-bias model that explicitly couples chemistry similarity and site discrepancy.

2.8 Site Perturbation Controls

To test whether explicit site information mattered beyond chemistry alone, we performed descriptor-level perturbation controls while holding edit chemistry fixed. These controls are interpreted as sensitivity analyses of the learned representation. Because test-time site corruption can create feature combinations that are less common in the training distribution, these experiments are not treated as standalone proof of causal site effects.

Binary site perturbation.

In the wrong-site control, each edit retained its chemistry but was reassigned to an incorrect site using a deterministic perturbation rule. In the coarse-position control, exact site identities were replaced by a coarse regional representation along peptide length. These controls test whether exact site-indexed descriptors carry information beyond degraded positional alternatives.

Graded anchor-shift perturbation.

To test whether the site-conditioned effect was distance-sensitive, we shifted anchors by ± 1, ± 2, ± 3, and ± 5 residues while holding edit chemistry fixed. Performance degradation under graded perturbation was interpreted as evidence of sensitivity to sequence-indexed site placement within the current descriptor family.

2.9 Feature Ablation

Systematic ablation was performed across chemistry-only, partial site-indexed, attachment-context, and full-context descriptor variants. The evaluated feature sets

were A, A+B1, A+B2, A+B3, A+B, A+C, and A+B+C. In this notation, A denotes global edit chemistry, B1 denotes site-position statistics, B2 denotes anchor-site residue composition, B3 denotes anchor-site residue properties, and C denotes attachment-aware and multi-edit context descriptors.

A+B is interpreted as the stricter chemistry-plus-site comparison, whereas A+B+C is interpreted as the full-context site-conditioned comparison. This distinction is important because the full model contains both explicit site-indexed features and broader attachment-aware context features.

2.10 Model Training and Evaluation

All primary sequence-only and descriptor-based models were trained as XGBoost regressors under the corresponding train/validation/test definitions. Model selection was performed on the validation split, and final performance was reported on the held-out test split. The primary random split was used to evaluate within-regime predictive performance, whereas the 70% sequence-cluster split was used to evaluate robustness under stronger scaffold shift.

Model performance was evaluated using root mean squared error (RMSE), mean absolute error (MAE), coefficient of determination (R^2), and Spearman rank correlation. RMSE and MAE quantify absolute error, R^2 quantifies variance explained, and Spearman correlation quantifies ranking agreement. Where multiple random seeds were used, results were summarized as mean performance with standard deviation across seeds.

The main random-split comparison reported three primary model families: sequence-only, chemistry-only, and full-context site-conditioned models. The stricter A+B site-block model was reported in the ablation analysis to clarify how much of the full-context improvement was attributable to explicit site-indexed descriptors alone.

2.11 Statistical Analysis

Statistical testing was used to support interpretation of effect direction and repeat-level stability, not as a substitute for effect-size reporting. The primary evidence for model comparison was the combined pattern of R^2, RMSE, MAE, Spearman correlation, perturbation controls, ablation results, parser audit, and sequence-cluster split performance.

We used two complementary levels of statistical analysis. First, seed-level tests were used to assess whether performance differences were stable across repeated model training runs. For matched-seed comparisons, we used a paired t-test on seed-level performance values. When the number of recorded seeds differed between two model families, we used Welch's t-test and interpreted the result as an auxiliary repeat-level stability check rather than as definitive evidence of generalization.

Second, observation-level paired tests were used to examine whether one model produced systematically smaller errors than another model on the same held-out samples. For two models a and b, we computed the paired absolute-error difference

$$d_i = |y_i - \hat{y}_i^{(a)}| - |y_i - \hat{y}_i^{(b)}|$$

for each held-out test sample i. A two-sided Wilcoxon signed-rank test was then applied to the resulting paired differences.

Because the dataset contains multiple related derivatives from shared or closely related peptide scaffolds, observation-level samples are not fully independent. Therefore, observation-level p-values were interpreted as complementary descriptive evidence about paired sample-wise error shifts rather than as the primary inferential basis. This distinction is especially important for the site-aware refinement, where effect sizes are modest and cluster-split transfer is limited.

Formal p-values were reported only for the main baseline comparisons. Perturbation controls, feature ablation, parser audit, GAM analysis, kernel analysis, and cluster-split evaluation were treated primarily as mechanistic and robustness analyses rather than as independent hypothesis-testing exercises.

3 Results and Discussion

3.1 Edit Chemistry Is the Dominant Predictive Axis in the Standard Random-Split Regime

We first tested whether access to edit chemistry materially improved prediction relative to a sequence-only baseline. On the standard random split, the sequence-only XGBoost model reached test $R^2 = 0.2635 \pm 0.0019$, whereas the chemistry-only descriptor model reached $R^2 = 0.4327 \pm 0.0042$ (Table 3). This corresponds to an absolute increase of 0.1692 in R^2, together with lower RMSE and MAE and higher Spearman correlation.

This improvement was supported by both levels of statistical analysis. At the seed level, the improvement was significant under the paired t-test ($p = 2.9 \times 10^{-8}$). At the paired-sample level, the Wilcoxon signed-rank test over the 1084 held-out samples also indicated a strong reduction in absolute error ($p = 2.1 \times 10^{-19}$). Because related derivatives are present in the dataset, the observation-level test is interpreted as complementary evidence rather than as a fully independent-sample inference. Taken together, the effect size and repeat-level stability support the conclusion that edit chemistry carries the dominant predictive signal for cyclic peptide permeability in the present curated known-site dataset.

For passive permeability, this hierarchy is chemically plausible. Quantities such as lipophilicity, polarity, hydrogen-bonding capacity, ring content, and conformational restriction are expected to influence transport behavior [10, 14, 18, 19]. The observed chemistry-driven gain over sequence-only modeling is therefore both statistically stable and chemically interpretable, while still remaining a dataset-specific predictive finding rather than a causal claim.

3.2 Explicit Site Conditioning Adds a Reproducible but Modest Refinement

We next asked whether explicit site-aware representation improved upon chemistry-only aggregation. The site-aware descriptor model improved test R^2 from 0.4327 to 0.4717 (Table 3), corresponding to an absolute gain of 0.0390. This gain is much smaller than the gain obtained by adding chemistry to the sequence-only baseline,

Table 3 Main model performance on the standard random split. Values are means across seeds; numbers in parentheses denote SD in the last digits. Seed-level p-values assess repeat-level stability, whereas observation-level p-values are complementary paired-error tests on the same held-out samples.

Model	RMSE	MAE	R^2	Sp.	vs prev. seed p	vs prev. obs. p
Seq-only	0.9805(13)	0.6491(17)	0.2635(19)	0.5213(34)	N/A	N/A
Chem-only	0.8605(32)	0.5242(23)	0.4327(42)	0.7183(32)	2.9×10^{-8}	2.1×10^{-19}
Site-aware	**0.8304(15)**	**0.5056(2)**	**0.4717(19)**	**0.7331(12)**	2.6×10^{-6}	0.054

Note. p-values are computed relative to the previous row. For Seq-only versus Chem-only, the seed-level test is a paired t-test across matched repeated runs. For Chem-only versus Site-aware, Welch's t-test is used because the recorded number of seeds differs between the two model families; this value is interpreted as a repeat-level stability check. Observation-level p-values are two-sided Wilcoxon signed-rank tests on paired absolute-error differences across the 1084 held-out samples. Because the dataset is derivative-heavy, observation-level p-values are complementary descriptive evidence rather than the primary inferential basis.

indicating that site assignment acts as a secondary refinement rather than as the dominant predictive axis.

At the seed level, the site-aware model outperformed the chemistry-only model under Welch's t-test ($p = 2.6 \times 10^{-6}$). However, the paired observation-level Wilcoxon signed-rank test over the 1084 held-out samples yielded $p = 0.054$, which did not cross the nominal $\alpha = 0.05$ threshold. We therefore avoid framing the site-aware model as strongly or universally superior. A more defensible interpretation is that, in this curated known-site setting, explicit site assignment provides a reproducible but quantitatively modest refinement over chemistry-only descriptors.

This interpretation is important for practical use. In early triage across chemically diverse edits, chemistry-aware modeling may already capture most of the actionable signal. In later-stage scaffold-focused optimization, where related edits are compared on a narrower family of peptide backbones, a smaller site-aware refinement may still be useful for prioritization.

3.3 Position Alone Is Not Sufficient, but It Is the Strongest Single Site-Aware Signal

A major concern is that the apparent site-aware gain might arise from positional shortcuts rather than chemistry–site interplay. To test this, we trained a position-only XGBoost control based solely on anchor count and site-location summary statistics. This model achieved $R^2 = 0.175 \pm 0.002$, far below both the chemistry-only baseline ($R^2 = 0.433 \pm 0.004$) and the full site-aware model ($R^2 = 0.472 \pm 0.002$). Thus, anchor position alone is weakly predictive and cannot explain the main performance gain by itself.

At the same time, the ablation analysis below shows that site-position statistics are the strongest individually effective site-aware component in the current descriptor family. The appropriate conclusion is therefore not that position is merely an artifact, but that the present site-aware gain is best understood as chemistry plus position-sensitive refinement, with richer contextual summaries contributing mainly in combination.

3.4 Wrong-Site and Coarse-Site Controls Indicate That Site Fidelity Matters Beyond Chemistry Alone

To isolate the contribution of site assignment, we perturbed the anchor representation while keeping edit chemistry fixed. Wrong-site reassignment reduced test R^2 from 0.4720 to 0.3828, and coarse regional encoding reduced it to 0.4105 (Table 4). These correspond to relative drops of 18.9% and 13.0%, respectively. This result indicates that, once chemistry is supplied, the model remains sensitive to site fidelity under the present representation.

Table 4 Binary site-perturbation controls for the site-aware descriptor model (seed 42).

Setting	RMSE	MAE	R^2	Spearman
Correct site	0.8302	0.5056	0.4720	0.7326
Wrong site	0.9231	0.6372	0.3828	0.6464
Coarse position	0.8770	0.5703	0.4105	0.6820

These controls are useful because they hold edit chemistry fixed while degrading positional fidelity. They therefore support the interpretation that the model is not simply reacting to the presence of an edit, but also to where that edit is represented as being placed. However, these perturbations are artificial corruptions of the descriptor space. The resulting performance drops should therefore be interpreted as sensitivity evidence under controlled descriptor perturbation, not as direct proof of a causal biochemical site effect.

3.5 The Site Effect Is Sharply Local in the Present 1D Positional Encoding

We next examined whether the site effect was binary or graded. Shifting anchors by just one residue reduced test R^2 from 0.4720 to 0.2881 or 0.2020, depending on direction (Table 5). At two residues, performance fell below the mean-predictor baseline, producing negative R^2 values. This indicates that the site-aware component captured by the present descriptor family is not a broad regional context signal but a highly localized positional signal.

This result substantially narrows the interpretation of "site awareness" in the current work. The model is not learning a broad topology-aware positional prior; rather, it appears to rely on sharp local site identity encoded through normalized sequence position and anchor-linked summaries. Because these shifted-anchor settings can also generate descriptor combinations that are less represented in training, the magnitude of the performance collapse should be interpreted cautiously. The key conclusion is not that one-residue physical movement necessarily causes such a large biochemical change, but that the current representation is highly sensitive to precise sequence-indexed site assignment.

Given that cyclic peptide permeability depends on conformational organization rather than only on 1D sequence index, the present site effect should be interpreted as

Table 5 Graded anchor-shift perturbation analysis for the site-aware descriptor model (seed 42).

Shift setting	RMSE	MAE	R^2	Spearman
Correct site	0.8302	0.5056	0.4720	0.7326
−1 residue	0.9624	0.6715	0.2881	0.5618
+1 residue	1.0187	0.7422	0.2020	0.4689
−2 residues	1.2024	0.9678	−0.1121	0.2212
+2 residues	1.2154	0.9815	−0.1363	0.1964
−3 residues	1.2879	1.0732	−0.2761	0.1083
+3 residues	1.3014	1.0901	−0.3028	0.0846
−5 residues	1.3361	1.1327	−0.3732	0.0418
+5 residues	1.3448	1.1435	−0.3911	0.0275

a sequence-indexed local refinement rather than as a full structural notion of anchor context.

3.6 Feature Ablation Shows that Position Statistics Carry the Strongest Standalone Site Signal

To determine which site-aware features actually drive the gain, we evaluated systematic ablations over the descriptor blocks (Table 6). Starting from chemistry alone (A), adding only site-position statistics (A+B1) increased test R^2 from 0.4327 to 0.4443. By contrast, adding only anchor-site residue composition (A+B2) or only residue-property summaries (A+B3) yielded weaker performance, with $R^2 = 0.4236$ and 0.4185, respectively. Combining all site-aware features (A+B) raised performance to 0.4625, and the full model including attachment-aware aggregation (A+B+C) reached 0.4720.

Table 6 Feature-group ablation on the standard random split (seed 42). A: chemistry descriptors; B1: site-position statistics; B2: anchor-site residue composition; B3: anchor-site residue properties; C: attachment-aware and multi-edit aggregation.

Feature set	RMSE	MAE	R^2	Spearman
A (chemistry only)	0.8606	0.5242	0.4327	0.7183
A+B1 (chemistry + position)	0.8517	0.5174	0.4443	0.7194
A+B2 (chemistry + residue composition)	0.8674	0.5363	0.4236	0.7194
A+B3 (chemistry + residue properties)	0.8712	0.5429	0.4185	0.7074
A+B (chemistry + all site-aware)	0.8376	0.5102	0.4625	0.7375
A+C (chemistry + attachment-aware)	0.8634	0.5244	0.4289	0.7148
A+B+C (full model)	**0.8302**	**0.5056**	**0.4720**	**0.7326**

This narrows the claim of the paper in an important way. The present descriptor family should not be interpreted as fully capturing local anchor-context chemistry semantics. Rather, the current evidence supports a more modest but still useful statement: the observed gain is best understood as chemistry plus position-sensitive refinement, with residue-aware summaries helping mainly in combination rather than independently.

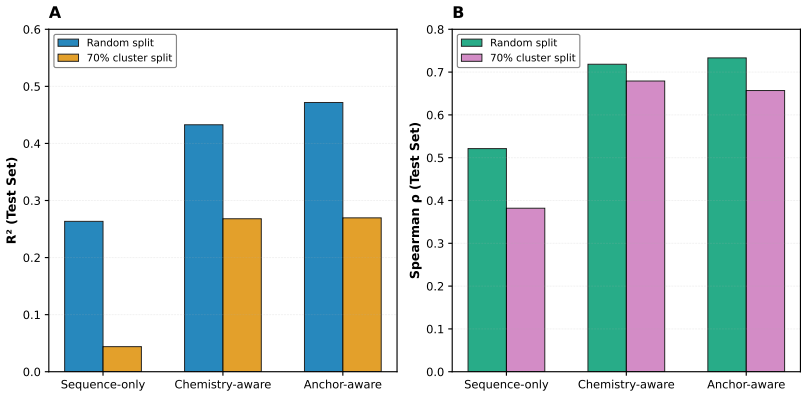


Fig. 2 Predictive performance under standard random-split and sequence-cluster evaluation. (A) Test-set R^2 for sequence-only, chemistry-aware, and anchor-aware descriptor models under the standard random split and the 70% sequence-identity cluster split. (B) Test-set Spearman rank correlation for the same model families and split settings. Chemistry-aware descriptors substantially improve over the sequence-only baseline in both evaluation regimes, indicating that edit chemistry carries the dominant predictive signal. Anchor-aware descriptors provide an additional gain under the random split, but this gain is strongly attenuated under the sequence-cluster split, consistent with a site-conditioned refinement that is mainly effective within scaffold-nearby derivatives rather than broadly transferable across stronger scaffold shift.

3.7 Auxiliary Interpretable Models Support the Same Hierarchy

The GAM and kernel branches were used to test whether different inductive biases led to the same mechanistic interpretation. The GAM reached test $R^2 = 0.174$ and Spearman $= 0.512$ on the standard split but enabled explicit decomposition of predictive contribution. On the fitted additive scale used by the GAM, the mean absolute contribution of the chemistry block was 5.905, compared with 0.217 for the anchor block and 0.037 for the chemistry–anchor interaction block. These values are used for within-model comparison rather than as absolute physical effect sizes. They support a chemistry-dominant regime with a smaller site-refinement component.

The anchor-sensitive kernel ridge model reached test $R^2 = 0.410$ and Spearman $= 0.712$ on the standard split, showing that an explicit chemistry-similarity-times-site-discrepancy formulation can recover much of the in-distribution signal. However, it generalized poorly under stronger distribution shift ($R^2 = -0.056$, Spearman $=$ 0.448), reinforcing the conclusion that the current site-aware benefit is not yet broadly robust.

Together, these auxiliary analyses support the same hierarchy observed in the main descriptor models: edit chemistry provides the dominant signal, while site-aware information contributes a smaller and less transferable refinement. The corresponding random-split and sequence-cluster comparison is summarized in Fig. 2, which should appear near this discussion.

3.8 Parser Audit Supports Bounded Robustness but Also Defines an Interpretation Boundary

Because the framework depends on rule-based site resolution, parser artifacts could in principle manufacture apparent site sensitivity. Automated consistency checks identified a per-edit issue rate of 9.55%, with most flagged cases corresponding to expected global-anchor or structurally special cases rather than obvious catastrophic failures. The stratified audit comprised 150 manually inspected records, and a conservative high-confidence subset retained 68.1% of the usable dataset.

These results suggest that parser uncertainty is bounded rather than overwhelming, and that the chemistry-dominant/site-refinement hierarchy is not obviously an artifact of gross parsing failure. At the same time, the current audit is not large enough to justify strong claims of parser independence. The audit should therefore be read as a robustness boundary rather than as proof that parser error has no effect on downstream site-sensitive conclusions.

3.9 Stronger Lightweight Baselines Improve Fairness but Do Not Change the Main Interpretation

Additional k-mer sequence baselines and k-mer-plus-chemistry fusion baselines improve upon simple amino-acid composition, tightening the fairness of the sequence comparison. However, these stronger lightweight baselines do not overturn the main interpretation of the study: chemistry remains the dominant predictive signal, and explicit site conditioning provides a smaller within-regime refinement.

This point matters because the current manuscript is best read as a mechanistic descriptor study rather than as a claim that hand-crafted descriptors plus XGBoost are the final modeling solution for modified peptides. Stronger protein language model or graph-based models remain important comparators for future work, but their absence does not negate the specific mechanistic conclusions supported by the present perturbation and ablation results.

3.10 The Site-Aware Gain Collapses under Stronger Sequence Distribution Shift

The harshest stress test is the sequence-cluster split at 70% sequence identity. Under this split, the sequence-only model nearly collapses ($R^2 = 0.044$), the chemistry-only model remains substantially more robust ($R^2 = 0.268$), and the site-aware descriptor model improves only marginally beyond chemistry alone ($R^2 = 0.270$). Thus, the site-aware gain visible under the derivative-rich random split does not transfer strongly across harder scaffold shift in the current descriptor regime.

This result should be treated as a main finding rather than as a footnote. It does not invalidate the perturbation evidence that site fidelity matters when chemistry is held fixed; rather, it shows that the present encoding of site information captures a mostly local, within-regime refinement that is not yet robust under stronger out-of-distribution sequence shift. In this sense, the current representation is more

appropriate for scaffold-nearby prioritization than for broad transfer across unseen peptide families.

3.11 Practical Implications for Cyclic Peptide Design

The practical implication of the present results is not that site-aware descriptors should replace chemistry-first thinking in cyclic peptide design. Rather, they suggest a sequential use case. In early triage across chemically diverse edits, chemistry-aware modeling likely captures most of the actionable signal. In later-stage scaffold-focused optimization, where chemists compare related edits on a narrower family of peptide backbones, explicit site assignment may provide an additional prioritization cue. This use case is also consistent with broader medicinal-chemistry interest in rational optimization of cyclic peptide permeability through localized chemical modification [20, 21].

Accordingly, the present framework is most naturally interpreted as a *refinement tool for within-regime library ranking* rather than a general de novo design engine. A stronger future version of this claim should be supported by explicit scaffold-level case studies, prospective ranking exercises, or independent medicinal chemistry validation.

3.12 Limitations and Outlook

The study has several important limitations.

First, it is based on a single phenotype, cyclic peptide permeability, in which chemistry-related quantities such as lipophilicity and polar surface area are expected to dominate [10, 18]. The same hierarchy may not transfer unchanged to peptide properties such as binding affinity, protease stability, clearance, cell-specific uptake, or activity [1, 2].

Second, the dataset is derivative-heavy, with multiple related edits per sequence. This is useful for isolating chemistry and site effects, but it may also increase the extent to which modest site-aware gains reflect within-family regularities rather than broadly transferable rules. This issue is also why observation-level p-values are treated as complementary rather than primary inferential evidence.

Third, the framework assumes explicit site observability after rule-based parsing. Uncertain-site records, jointly inferred-site settings, and de novo design scenarios remain outside the demonstrated scope.

Fourth, the current descriptor family is intentionally limited to 1D positional summaries and 2D chemistry descriptors. It does not explicitly encode conformational ensembles, IMHB topology, or chameleonic behavior, all of which are likely to be important for cyclic peptide permeability [10, 12–14]. Future work should therefore test whether conformation-aware features change the apparent balance between chemistry and site assignment.

Fifth, although the present manuscript includes stronger lightweight sequence baselines, it does not yet provide a full comparison against stronger modern end-to-end baselines such as peptide language model fusion or chemistry-aware graph architectures [4–8, 22, 23]. These comparisons remain important for establishing how much of the current mechanistic picture is specific to interpretable descriptor regimes and how much persists in richer learned representations.

These limitations do not negate the mechanistic findings of the present study, but they define their boundary conditions.

Code Availability

To support transparency and reproducibility, the implementation used in this study is available in the public GitHub repository: https://github.com/leochenminrui/CASPer. The repository contains the core source code, preprocessing utilities, descriptor construction scripts, model-training and evaluation scripts, and documentation for reproducing the main analyses reported in this paper. When external datasets are not redistributed directly, the repository provides instructions for obtaining the required input data and regenerating the processed analysis tables. Versioned code and configuration files are provided to make the reported random-split, perturbation-control, ablation, parser-audit, and sequence-cluster experiments traceable.

4 Conclusions

We studied cyclic peptide permeability prediction as a chemistry-versus-site disentanglement problem in a curated known-site regime. Rather than presenting a general-purpose cyclic peptide permeability predictor, we asked a narrower mechanistic question: after localized edit chemistry is represented, how much additional predictive information is contributed by explicit site assignment?

Across controlled descriptor comparisons, site-perturbation analyses, feature ablation, auxiliary interpretable models, parser-robustness checks, and a harder sequence-cluster split, a consistent hierarchy emerged. Edit chemistry provided the dominant predictive signal in the present dataset. Explicit site conditioning added a smaller refinement under standard random-split evaluation, but this refinement was modest in magnitude and strongly attenuated under stronger sequence distribution shift. Perturbation controls further showed that the fitted representation was sensitive to site fidelity and local anchor shifts, although these effects should be interpreted as sequence-indexed descriptor sensitivity rather than direct evidence of three-dimensional structural causality.

The main contribution of this work is therefore conceptual and analytical. We provide an interpretable framework for separating edit-chemistry effects from site-conditioned effects in a known-site cyclic peptide setting, quantify the scale of the current site-aware gain, and identify the conditions under which this gain appears reliable. The results suggest that chemistry-aware descriptors are likely to be most useful for broad early triage, whereas site-conditioned descriptors may serve as a secondary refinement for scaffold-nearby library prioritization. Future work should test whether the same hierarchy persists when site information is combined with conformation-aware representations, stronger sequence or graph encoders, and prospective medicinal-chemistry validation.

5 Acknowledgements

This work was supported by JSPS KAKENHI Grant Number JP23H03356.

References

[1] Vinogradov, A.A., Yin, Y., Suga, H.: Macrocyclic peptides as drug candidates: recent progress and remaining challenges. Journal of the American Chemical Society **141**(10), 4167–4181 (2019)

[2] Muttenthaler, M., King, G.F., Adams, D.J., Alewood, P.F.: Trends in peptide drug discovery. Nature reviews Drug discovery **20**(4), 309–325 (2021)

[3] O'Connor, J.J.: De novo design of passively permeable cyclic peptides and incorporation of noncanonical amino acids to improve predicted binding affinity. PhD thesis, University of Washington (2021)

[4] Li, J., Yanagisawa, K., Akiyama, Y.: Cycpeptmp: enhancing membrane permeability prediction of cyclic peptides with multi-level molecular features and data augmentation. Briefings in Bioinformatics **25**(5), 417 (2024)

[5] Cao, L., Xu, Z., Shang, T., Zhang, C., Wu, X., Wu, Y., Zhai, S., Zhan, Z., Duan, H.: Multi_cycgt: a deep learning-based multimodal model for predicting the membrane permeability of cyclic peptides. Journal of medicinal chemistry **67**(3), 1888–1899 (2024)

[6] Zhang, Y., Chen, Z., Wan, Z., Jiang, Q., Lu, X., Yan, B., Qin, J., Liu, Y., Wang, J.: Msf-cpmp: A novel multi-source feature fusion model for prediction of cyclic peptide membrane permeability. Computational and Structural Biotechnology Journal (2025)

[7] Wang, Z., Chen, Y., Shang, Y., Yang, X., Pan, W., Ye, X., Sakurai, T., Zeng, X.: Multicycpermea: accurate and interpretable prediction of cyclic peptide permeability using a multimodal image-sequence model. BMC biology **23**(1), 63 (2025)

[8] Jiang, D., Chen, Z., Du, H.: Cyclic peptide membrane permeability prediction using deep learning model based on molecular attention transformer. Frontiers in Bioinformatics **5**, 1566174 (2025)

[9] Shendre, A., Gahlot, P.S., Raghava, G.P.: Pcppred: Prediction of chemically modified peptide permeability across multiple assays for oral delivery. bioRxiv, 2026–01 (2026)

[10] Rezai, T., Yu, B., Millhauser, G.L., Jacobson, M.P., Lokey, R.S.: Testing the conformational hypothesis of passive membrane permeability using synthetic cyclic peptide diastereomers. Journal of the American Chemical Society **128**(8), 2510–2511 (2006)

[11] Lewis, I., Schaefer, M., Wagner, T., Oberer, L., Sager, E., Wipfli, P., Vorherr, T.: A detailed investigation on conformation, permeability and pk properties of

two related cyclohexapeptides. International Journal of Peptide Research and Therapeutics **21**(2), 205–221 (2015)

[12] Riniker, S.: Toward the elucidation of the mechanism for passive membrane permeability of cyclic peptides. Future medicinal chemistry **11**(7), 637–639 (2019)

[13] Ramelot, T.A., Palmer, J., Montelione, G.T., Bhardwaj, G.: Cell-permeable chameleonic peptides: Exploiting conformational dynamics in de novo cyclic peptide design. Current opinion in structural biology **80**, 102603 (2023)

[14] Sugita, M., Sugiyama, S., Fujie, T., Yoshikawa, Y., Yanagisawa, K., Ohue, M., Akiyama, Y.: Large-scale membrane permeability prediction of cyclic peptides crossing a lipid bilayer based on enhanced sampling molecular dynamics simulations. Journal of chemical information and modeling **61**(7), 3681–3695 (2021)

[15] Li, J., Yanagisawa, K., Sugita, M., Fujie, T., Ohue, M., Akiyama, Y.: Cycpeptmpdb: a comprehensive database of membrane permeability of cyclic peptides. Journal of chemical information and modeling **63**(7), 2240–2250 (2023)

[16] Kansy, M., Senner, F., Gubernator, K.: Physicochemical high throughput screening: parallel artificial membrane permeation assay in the description of passive absorption processes. Journal of medicinal chemistry **41**(7), 1007–1010 (1998)

[17] Azadvari, N., Choi, J., Hosseinzadeh, P.: Peptide permeability measurements in pampa: Sources of variability and implications for data reproducibility (2026)

[18] Li, J., Kannan, S., Aronica, P., Brown, C.J., Partridge, A.W., Verma, C.S.: Molecular descriptors suggest stapling as a strategy for optimizing membrane permeability of cyclic peptides. The Journal of Chemical Physics **156**(6) (2022)

[19] Sugita, M., Noso, Y., Li, J., Fujie, T., Yanagisawa, K., Akiyama, Y.: Protocol for membrane permeability prediction of cyclic peptides using descriptors obtained from extended ensemble molecular dynamics simulations and chemical structures. bioRxiv, 2025–06 (2025)

[20] Aerts, R., Tavernier, J., Kerstjens, A., Ahmad, M., Gómez-Tamayo, J.C., Tresadern, G., De Winter, H.: C2po: an ml-powered optimizer of the membrane permeability of cyclic peptides through chemical modification. Journal of Cheminformatics **17**(1), 1–13 (2025)

[21] Nielsen, A.L., Bartling, C.R., Zarda, A., De Sadeleer, N., Neeser, R.M., Schwaller, P., Strømgaard, K., Heinis, C.: Bulk measurement of membrane permeability for random cyclic peptides in living cells to guide drug development. Angewandte Chemie International Edition **64**(27), 202500493 (2025)

[22] Liu, W., Li, J., Verma, C.S., Lee, H.K.: Systematic benchmarking of 13 ai methods

for predicting cyclic peptide membrane permeability. Journal of Cheminformatics **17**(1), 129 (2025)

[23] Cabezón, A., Otović, E., Kalafatovic, D., Piñeiro, Á., García-Fandiño, R., Mauša, G.: Unravelling cyclic peptide membrane permeability prediction: a study on data augmentation, architecture choices, and representation schemes. Digital Discovery **4**(5), 1259–1275 (2025)