ROBUST BIOMOLECULAR COMPLEX DESIGN ACROSS PROTEIN CONFORMATIONAL LANDSCAPES

Anonymous authors
Paper under double-blind review

ABSTRACT

Proteins populate conformational ensembles, yet structure-based biomolecular design typically optimizes candidates against a single target conformation. Consequently, a candidate that fits one state can lose favorable interactions or develop steric clashes when the target adopts another. We introduce FLEXEVO, a model-agnostic evolutionary framework that adapts candidates once at inference time from a single target conformation to improve compatibility with alternative natural conformations unseen during adaptation, without retraining the source model or requiring a conformational ensemble. FLEXEVO casts cross-state adaptation as geometry-constrained bi-objective optimization, balancing preservation of input-state interactions against robustness to plausible conformational perturbations. To limit the search space and reduce invalid structural edits, geometry-derived *FlexBoxes* define protected anchor regions, adaptable regions for local exploration, and forbidden regions for clash avoidance. A unified all-atom representation supports topology-preserving adaptation across diverse binder categories, while Pareto selection preserves nondominated candidates across the two objectives. We evaluate FLEXEVO across multiple generation baselines and nine representative binder categories spanning diverse molecular sizes and structural topologies. FLEXEVO reduces the category-balanced mean relative performance degradation from 47.8% to 4.4%, while adding only 1.4–3.1 minutes of adaptation per sample. These results establish single-state inference-time adaptation as a practical route toward robust biomolecular complex design across protein conformational landscapes.

1 INTRODUCTION

Structure-based prediction and design of biomolecular complexes provide a computational route to developing molecular binders for target proteins. Recent advances in geometric learning and generative modeling have enabled the prediction and generation of three-dimensional candidate complexes conditioned on target protein structures (Guan et al., 2023a;b; Kong et al., 2024; Watson et al., 2023). However, many existing approaches rely on a single-conformation assumption at inference time, optimizing candidates against one supplied or predicted realization of the binding pocket or interaction surface (Zhang et al., 2024; Zhou et al., 2025). Proteins instead populate conformational ensembles, where transitions among apo, holo, and alternative functional states can reshape binding geometries, disrupt favorable contacts, or introduce steric clashes (Boehr et al., 2006; 2009; Lu et al., 2024). Consequently, a candidate optimized for one target state may lose compatibility when the same target adopts another naturally occurring conformation.

Existing approaches can be broadly categorized into two settings. Static structure-conditioned predictors and generators treat the target as a fixed structural input, leaving conformational variation outside the candidate-design objective (Guan et al., 2023a; Kong et al., 2024; Watson et al., 2023). Flexibility-aware methods relax this assumption by incorporating receptor ensembles (Huang & Zou, 2007), modeling conformational motions (Sherman et al., 2006; Lu et al., 2024), or jointly generating receptor–binder complexes (Qiao et al., 2024; Zhang et al., 2024; Zhou et al., 2025), but this does not by itself establish that a single candidate will remain compatible with natural target conformations unseen during optimization. This distinction matters in practice because the target state encountered after design is often unknown: a candidate optimized for one state may lose favorable contacts or develop steric clashes in another. This gap raises a central question: *can a*

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.

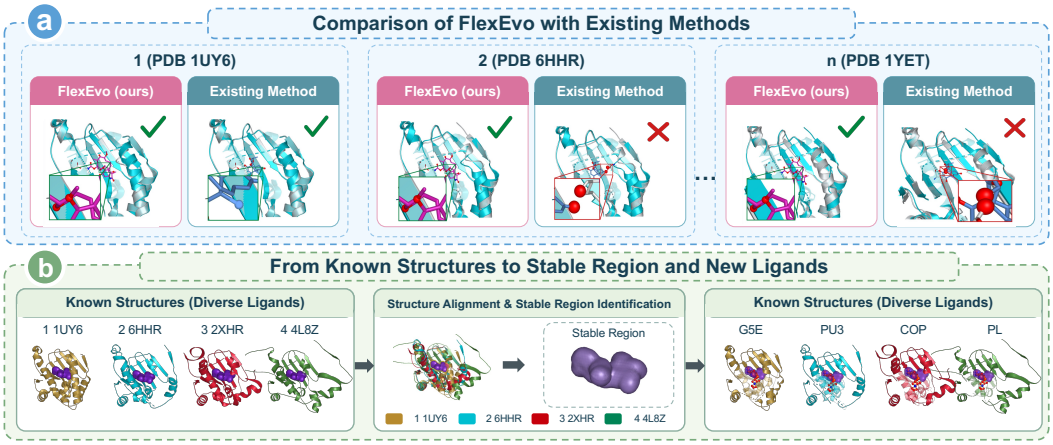

Figure 1: Motivation for cross-conformation ligand design. (a) Comparison of FlexEvo and single-state design methods, illustrating ligand compatibility across alternative target conformations. (b) Alignment of known protein–ligand complexes highlights a shared stable binding region, motivating ligand design that preserves conserved interactions while accommodating conformational variation.

candidate be adapted once from a single observed conformation and still remain compatible with multiple natural conformations of the same target, without redesigning it for each state?

Achieving this goal requires accommodating unobserved target variation while preserving favorable interactions in the observed state. Target conformational changes range from side-chain and loop rearrangements in compact pockets to backbone and interface remodeling across extended surfaces, while binder adaptation may require coordinated changes in pose, torsions, backbone geometry, and side-chain conformations (Lu et al., 2024; Raveh et al., 2011; Marze et al., 2018). Moreover, the number of atomic variables grows with binder size, rapidly expanding the search space. Because bonds, torsions, and steric constraints couple these variables, unrestricted edits can produce invalid geometries (Buttenschoen et al., 2024).

To address these challenges, we introduce FLEXEVO, a model-agnostic evolutionary adaptation framework that refines candidates from existing structure-conditioned predictors and generators without modifying their source models. Given a single observed target conformation and an initial candidate population, FLEXEVO formulates cross-state adaptation as bi-objective optimization over a geometry-constrained search space, balancing retention of input-state interactions against robustness to plausible conformational perturbations. Perturbation robustness is used as a proxy for compatibility with unseen natural target states. To accommodate structural diversity across binder classes, FLEXEVO combines a unified all-atom representation of target–candidate complexes with binder-specific structural variation and repair. To limit the search space and reduce invalid structural edits, FLEXEVO maps local packing and contact geometry of the input complex, together with protein excluded volume, to local *FlexBoxes*: protected anchor regions, adaptable regions for local exploration, and forbidden regions for clash avoidance. FLEXEVO uses structural feedback to update the spatial constraints guiding adaptation, with candidate moves guided by target geometry and available bond topology. Pareto selection retains nondominated solutions across the two objectives. Our contributions can be summarized as follows.

- We formulate inference-time binder adaptation from a single target conformation as a geometry-constrained bi-objective optimization, balancing input-state interaction preservation with robustness to reference-derived geometric perturbations. Adaptation requires neither source-model retraining nor access to alternative natural target conformations.

- We propose FLEXEVO, which couples an outer evolutionary search over residue-level constraint configurations with inner topology-preserving candidate adaptation guided by decoded *FlexBoxes*. Structural feedback updates the configurations, while Pareto selection retains nondominated preservation–robustness trade-offs.

- We conduct extensive evaluations across nine representative binder categories spanning diverse molecular sizes, structural topologies, and interaction geometries, using multiple generation baselines. FLEXEVO reduces the category-balanced mean relative performance degradation from 47.8% to 4.4%, while adding only 1.4–3.1 minutes per sample.

2 RELATED WORK

2.1 SINGLE-STATE BIOMOLECULAR COMPLEX PREDICTION AND DESIGN

Existing methods predict or design complexes around a single target realization. For small molecules, Pocket2Mol uses autoregressive generation, whereas TargetDiff, DiffSBDD, and DecompDiff use pocket-conditioned diffusion (Peng et al., 2022; Guan et al., 2023a; Schneuing et al., 2024; Guan et al., 2023b). PepGLAD extends full-atom generation to peptides, while RFdiffusion and BindCraft design target-specific protein binders (Kong et al., 2024; Watson et al., 2023; Pacesa et al., 2025). General-purpose predictors, including AlphaFold 3, Chai-1, Protenix, and BoltzGen, predict complexes spanning these binder classes (Abramson et al., 2024; Chai Discovery et al., 2024; ByteDance AML AI4Science Team et al., 2025; Stark et al., 2025). Despite their different architectures, these methods optimize or assess candidates against one realization of the target geometry, effectively treating it as rigid. Because target conformational variation is absent from the design objective, a candidate that fits one state may lose favorable contacts or develop steric clashes in another.

2.2 FLEXIBILITY-AWARE BIOMOLECULAR COMPLEX MODELING

To relax this rigid-target assumption, classical methods explicitly search receptor and binder flexibility through induced-fit docking, peptide folding and docking, or protein backbone ensembles (Sherman et al., 2006; Raveh et al., 2011; Marze et al., 2018). Ensemble docking and multistate design further consider multiple supplied structures during optimization (Huang & Zou, 2007; Amaro et al., 2018; Sauer et al., 2020). Learning-based methods such as NeuralPLexer and DynamicBind model coupled receptor–ligand changes, while FlexSBDD, DynamicFlow, and YuelDesign incorporate protein flexibility into ligand generation (Qiao et al., 2024; Lu et al., 2024; Zhang et al., 2024; Zhou et al., 2025; Wang et al., 2025). These methods model target flexibility, but do not ensure compatibility when the target state encountered after design is unknown. State-specific methods construct a favorable complex for one recovered conformation, leaving compatibility with other states unresolved. Ensemble-based methods broaden coverage only over conformations available beforehand; however, available structures often represent only a small fraction of the natural ensemble, and the relevant post-design state may not be known in advance. FlexEvo instead adapts one candidate from one observed conformation for compatibility with unseen natural states, extending this capability across diverse binder categories.

3 METHODOLOGY

3.1 PROBLEM SETUP AND OVERVIEW

Given a single observed target conformation P^0 and an initial pool $\mathcal{X}_0$ of binder structures positioned relative to it, FLEXEVO adapts candidates from an existing predictor or generator once at inference time. The goal is to preserve favorable interactions with P^0 while improving compatibility with unseen natural conformations of the same target. Adaptation uses only P^0 and $\mathcal{X}_0$, without retraining the source model; alternative natural conformations are reserved for evaluation. Let $G \in \mathcal{G}(P^0)$ denote an admissible residue-level constraint configuration (*Box genome*), and let $\mathcal{X}_G = \mathcal{A}(\mathcal{X}_0; P^0, G)$ denote the candidate pool obtained through constrained editing and repair. We formulate the search over G as a bi-objective surrogate problem:

$$\max_{G \in \mathcal{G}(P^0)} \quad \big(F_{\mathrm{pres}}(\mathcal{X}_G; P^0, G), F_{\mathrm{rob}}(\mathcal{X}_G; P^0, G)\big). \tag{1}$$

Here, F_{pres} and F_{rob} are genome-level scores for interaction preservation and perturbation robustness, respectively, with geometric validity included in the latter. Robustness is assessed using local

geometric perturbations derived from P^0 under G, as a proxy for cross-state compatibility. As illustrated in Fig. 2, each genome encodes region labels and box scales that decode into *FlexBoxes*, specifying where and how candidates can be edited. For each genome, the inner procedure adapts the same initial pool $\mathcal{X}_0$; candidate scores are then aggregated for genome evaluation. The outer search uses genome variation and structural feedback to update constraints, while Pareto-based selection retains configurations with different objective trade-offs.

3.2 EVOLVABLE FLEXBOX PRIOR

Static packing and radial position provide structural cues to protein fluctuations (Jamroz et al., 2012; Lin et al., 2008; Shih et al., 2007). Studies of protein–protein recognition further show that preorganized anchors can coexist with local side-chain adjustment (Rajamani et al., 2004). These observations, together with the need to avoid steric clashes (Buttenschoen et al., 2024), motivate distinct editing roles for anchoring, local adaptation, and avoidance. The outer search in FLEXEVO evolves the *spatial prior governing candidate adaptation*, separating the allocation of local editing freedom from the structures produced under that prior. We first screen frozen source-model outputs for basic geometric defects and redundancy to obtain $\mathcal{X}_0$ (Fig. 2(1)). This pool is constructed once, without using either evolutionary objective, and is shared across genome evaluations.

From P^0, we compute fixed residue descriptors, centroids $c_r \in \mathbb{R}^3$, and orthonormal frames $Q_r \in \mathbb{R}^{3\times 3}$ for the R represented target residues. A Box genome specifies the residues covered by the prior, their editing roles, and per-residue scales:

$$\begin{aligned} G &= \big(\mathcal{R}_G, \{\ell_r\}_{r\in\mathcal{R}_G}, Z\big), && Z \in [z_{\min}, z_{\max}]^R, \\ \mathcal{K} &= \{\text{stable}, \text{flexible}, \text{forbidden}\}, && \ell_r \in \mathcal{K}, \quad r \in \mathcal{R}_G, \end{aligned} \tag{2}$$

where $\mathcal{R}_G \subseteq \{1, \dots, R\}$ and $0 < z_{\min} < z_{\max}$. Only residues in $\mathcal{R}_G$ instantiate a *FlexBox*:

$$\mathcal{B}_r(G) = \left\{ y \in \mathbb{R}^3 : \left\| Q_r^\top (y - c_r) \right\|_\infty \le \frac{b_{\ell_r} z_r}{2} \right\}, \qquad r \in \mathcal{R}_G, \tag{3}$$

with role-dependent base extent b_{ℓ_r}. The three roles implement *Bind–Adapt–Avoid* guidance (Fig. 2(2)). Stable FlexBoxes encourage proximity to anchor regions; flexible FlexBoxes support broader local exploration; forbidden FlexBoxes penalize occupancy and guide clash-relieving corrections. These labels specify candidate-editing roles rather than measured residue dynamics; forbidden regions mark occupancy to avoid, not target atoms assumed to be immobile. A geometry-derived seed and geometry-biased variants initialize the outer population. Evolution changes the support, roles, and scales while the residue descriptors, centers, and frames remain fixed. Appendix A.2 specifies their construction and initialization.

3.3 GENOME-CONDITIONED ADAPTATION AND SINGLE-STATE ROBUSTNESS

FlexBox-guided candidate adaptation. A genome is evaluated through the candidate population it produces. For each G, FLEXEDITOR starts from the shared $\mathcal{X}_0$ with a common adaptation budget and reset random seed, keeping the decoded FlexBoxes fixed. The boxes guide the locations and types of candidate edits. FlexEditor supports a broader coordinate-level operator family, whereas the reported molecular experiments use its graph-preserving mode, including rigid translations and rotations together with annotated bond-axis torsions. These candidate-side operations are summarized in the lower row of Fig. 2(4). Let $\mathcal{C}_G^t$ be the inner population and $\mathcal{E}_G^t \subseteq \mathcal{C}_G^t$ its reproduction subset. For each parent x and proposal index $v \in \{1, \dots, n_v\}$, editing $\mathcal{T}_G$ and geometric repair $\mathcal{R}$ produce $\widehat{x}_{x,v} = \mathcal{R}(\mathcal{T}_G(x; \xi_{t,x,v}))$, where ξ denotes proposal randomness. Writing $n(x)$ for the number of represented coordinates and $c(x)$ for local neighbor support, an adaptation round is

$$\begin{aligned} \mathcal{V}_G^t &= \left\{ \widehat{x}_{x,v} : x \in \mathcal{E}_G^t,\ 1 \le v \le n_v,\ n(\widehat{x}_{x,v}) > 0,\ c(\widehat{x}_{x,v}) \ge \tau_c \right\}, \\ \mathcal{C}_G^{t+1} &= \mathcal{S}_x\big(\mathcal{C}_G^t \cup \mathcal{V}_G^t\big), \qquad \mathcal{X}_G = \mathcal{C}_G^T. \end{aligned} \tag{4}$$

Here T is the number of inner rounds, and $\mathcal{S}_x$ combines nondominated sorting, objective-space crowding, and candidate-space niching. Repair and screening precede survivor selection, so proposals failing the specified geometric checks cannot compensate through favorable scores. The edit primitives and geometric screening criteria are specified in Appendix A.3.

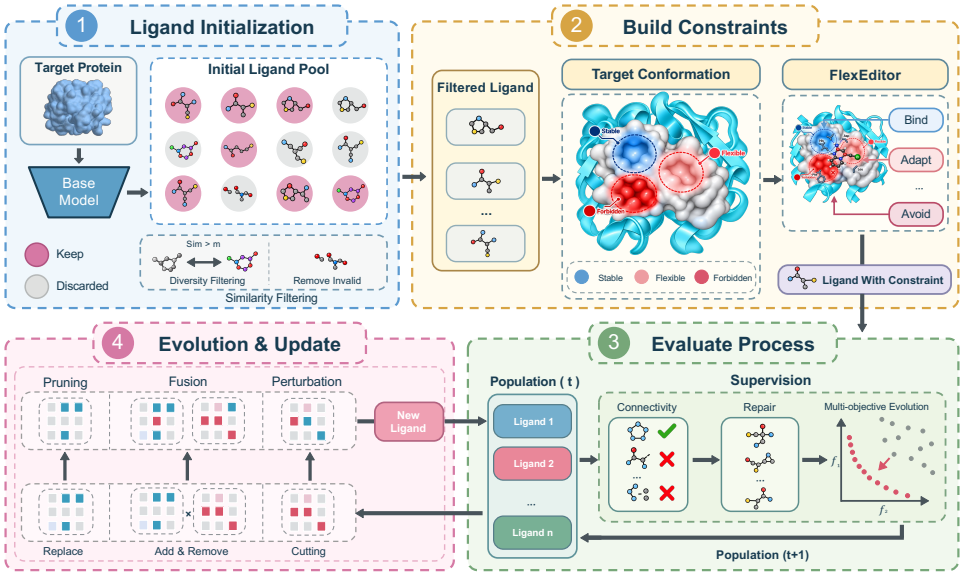

Figure 2: Overview of the FlexEvo framework. Starting from a single target conformation, FlexEvo filters an initial candidate pool and constructs geometry-derived FlexBoxes defining protected, adaptable, and forbidden regions. Iterative structural editing, repair, and Pareto-based selection balance input-state interaction preservation with robustness to conformational perturbations.

Single-state perturbation probes. Good contact with P^0 alone does not establish tolerance to target motion. We therefore construct a genome-conditioned probe collection $\widetilde{\mathcal{P}}_G = \{\widetilde{P}_G^0, \ldots, \widetilde{P}_G^M\}$, with $\widetilde{P}_G^0 = P^0$, by applying local radial and tangential perturbations to target regions selected by the prior. Candidate coordinates remain fixed across probes: evaluation measures the response of the *same* candidate, not its performance after separate re-optimization for each state. The probes are synthetic geometric stress tests constructed solely from P^0; alternative natural conformations are reserved for evaluation. For candidate x, let q_m and χ_m denote contact quality and clash fraction under $\widetilde{P}_G^m$. Over $m = 0, \ldots, M$, define

$$
\begin{aligned}
q_{\mathrm{rob}} &= \rho \min_m q_m + (1-\rho)\bar{q}, \qquad \bar{q} = \frac{1}{M+1}\sum_m q_m, \\
\Delta_q &= \max_m q_m - \min_m q_m, \qquad \chi_{\max} = \max_m \chi_m, \qquad \kappa_q = \frac{1}{M+1}\sum_m \mathbf{1}_{\{q_m \geq \tau q_0\}}.
\end{aligned}
\tag{5}
$$

Here $\rho \in [0,1]$ controls worst-case emphasis, $\tau \in (0,1)$ specifies the retained-contact ratio, and $\kappa_q = 0$ when $q_0 \leq \varepsilon_q$ for $\varepsilon_q > 0$. Together, q_{rob}, Δ_q, and κ_q characterize retained contact, perturbation sensitivity, and contact consistency. App. D derives finite-probe score bounds and conditional transfer bounds under explicit target-proximity assumptions.

Bi-objective evaluation. We balance interaction preservation against geometric quality and perturbation tolerance using two maximized candidate objectives:

$$
\begin{aligned}
f_{\mathrm{pres}}(x;G) &= \alpha_a a + \alpha_0 q_0 + \alpha_r q_{\mathrm{rob}} - \alpha_\chi \chi_{\max}, \\
f_{\mathrm{rob}}(x;G) &= \beta_g g + \beta_c c + \beta_k \kappa_q - \beta_d \Delta_q - \Pi(x;G).
\end{aligned}
\tag{6}
$$

We suppress component arguments for readability. The anchor-proximity score a and reference contact q_0 encourage compatibility with P^0, while q_{rob} rewards contact maintained under perturbation. The internal-geometry score g, neighbor support c, and consistency κ_q favor geometrically supported candidates with stable contact quality. The penalty Π combines flexible-region exposure, forbidden occupancy, probe clashes, and accumulated edit cost. All coefficients are fixed and nonnegative; Appendix A.4 defines the components and their values. The search uses geometric scores

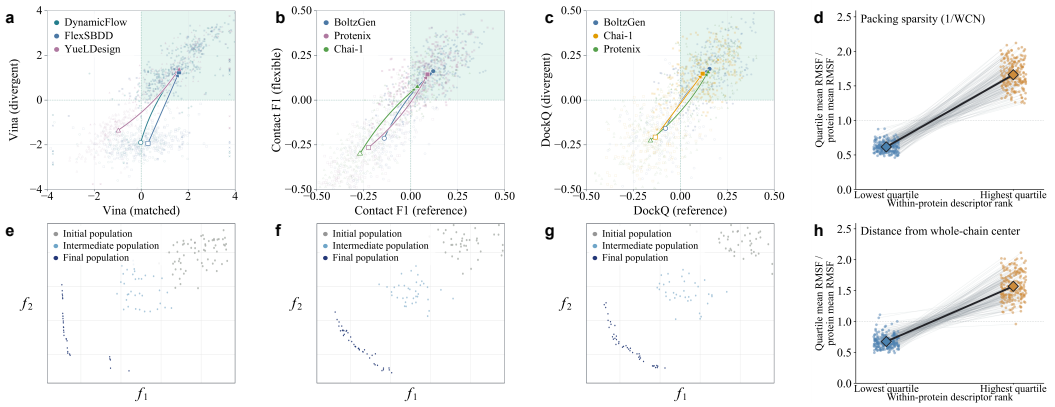

Figure 3: Cross-conformation evaluation and analysis of FlexEvo. (a–c) Distributions of scores derived from Vina, contact F1, and DockQ before and after adaptation for small molecules, peptides, and protein binders, respectively, under the paired target conditions shown. (e–g) Initial, intermediate, and final populations in the corresponding two-objective spaces. (d,h) Mean residue RMSF in the lowest and highest within-protein quartiles of packing sparsity ($1/\mathrm{WCN}$) and distance from the whole-chain C_α geometric center, respectively, for 200 ATLAS proteins. Quartile means are normalized by each protein's whole-chain mean RMSF. In (d,h), points denote proteins, faint lines pair quartiles, and diamonds with thick lines indicate equal-weight means across proteins.

rather than external docking or affinity estimators. The cost schematic in Fig. 2(3) corresponds to $f_1 = -f_{\mathrm{pres}}$ and $f_2 = -f_{\mathrm{rob}}$. To assess the prior rather than only its best candidate, we aggregate both leading-candidate and pool-average quality. For nonempty $\mathcal{X}_G$, the genome objectives in Eq. equation 1 are

$$F_j(\mathcal{X}_G; P^0, G) = \lambda \max_{x \in \mathcal{X}_G} f_j(x; G) + \frac{1-\lambda}{|\mathcal{X}_G|} \sum_{x \in \mathcal{X}_G} f_j(x; G) - \eta_j \Omega(G),$$
$$\Omega(G) = \frac{|\mathcal{R}_G|}{R}, \qquad j \in \{\mathrm{pres}, \mathrm{rob}\}, \quad \eta_{\mathrm{pres}} = 0, \quad \eta_{\mathrm{rob}} = \eta. \tag{7}$$

Here $\lambda \in [0, 1]$ balances the maximum and mean, and $\eta \geq 0$ penalizes extensive constraint assignment. The two objective-wise maxima need not correspond to the same candidate.

3.4 Feedback-Driven Genome Evolution

Spatial genome variation. Each offspring applies one of the three operators in Fig. 2(4): *Pruning*, *Fusion*, or *Perturbation*. Pruning retires weakly used stable or flexible constraints, reducing the support of the prior without removing target atoms. Fusion inherits a spatial configuration from a donor G_b. For a sampled residue s and neighborhood $\mathcal{U} = \{r : \|c_r - c_s\|_2 \leq d_{\mathrm{cross}}\}$, it updates

$$\mathcal{R}_{\widetilde{G}} = (\mathcal{R}_{G_a} \setminus \mathcal{U}) \cup (\mathcal{R}_{G_b} \cap \mathcal{U}),$$
$$(\widetilde{\ell}_r, \widetilde{z}_r) = \begin{cases} (\ell_r^b, z_r^b), & r \in \mathcal{R}_{G_b} \cap \mathcal{U}, \\ (\ell_r^a, z_r^a), & r \in \mathcal{R}_{G_a} \setminus \mathcal{U}. \end{cases} \tag{8}$$

Transferring roles and scales together retains the donor's configuration within the selected spatial block. Perturbation changes support or roles within a local neighborhood and resizes a separately sampled residue subset $\mathcal{M}_z$:

$$\widetilde{z}_r = \mathrm{clip}\big(z_r e^{\epsilon_r}, z_{\min}, z_{\max}\big), \qquad \epsilon_r \sim \mathcal{N}(0, \sigma_z^2), \quad r \in \mathcal{M}_z. \tag{9}$$

The role changes are spatially local; scale updates need not be. Sampling rules are given in Appendix A.5.

Structural feedback. Initialization alone does not reveal which constraints will be exercised during adaptation. We therefore update an offspring proposal using evidence from its primary parent's

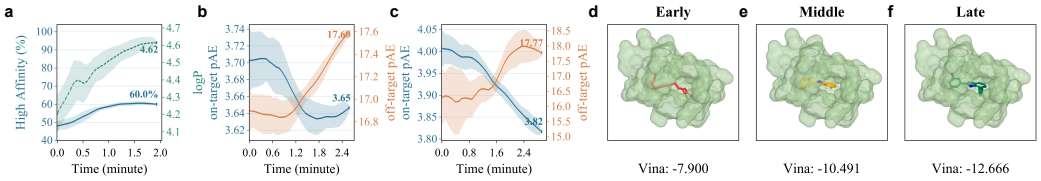

Figure 4: Metric trajectories and structural progression. (a) High-affinity fraction (blue, left axis) and $\log P$ (green, right axis) over adaptation time. (b,c) On-target pAE (blue, left axis) and off-target pAE (orange, right axis). Time is shown in minutes. (d–f) Early, intermediate, and late binding poses for one ligand example, with the corresponding Vina scores shown below.

adapted pool $\mathcal{X}_{G_a}$. Let u_r count uses of residue r's FlexBox in surviving edit histories. Usage and proximity define the feedback signal

$$D_r = \min_{x \in \mathcal{X}_{G_a}, i} \|p_i(x) - c_r\|_2, \qquad s_r = \max\left\{ \frac{u_r}{\max(1, \sum_k u_k)}, \left[1 - \frac{D_r}{d_c}\right]_+ \right\}, \tag{10}$$

where $[z]_+ = \max(z, 0)$ and $d_c > 0$ is the proximity radius. Usage records constraints exercised during editing, while proximity can identify relevant regions with no active box or use history. High-signal uncovered residues are activated as flexible constraints; low-signal stable or flexible assignments are retired under a bounded budget. For the corresponding sets $\mathcal{A}_+$ and $\mathcal{A}_-$,

$$\mathcal{R}_{G^+} = (\mathcal{R}_{\widetilde{G}} \cup \mathcal{A}_+) \setminus \mathcal{A}_-, \qquad \ell_r^+ = \text{flexible for } r \in \mathcal{A}_+. \tag{11}$$

Pruning and feedback retirement exclude forbidden constraints. After genome repair, the offspring is evaluated again from $\mathcal{X}_0$. Thus parent candidates inform *where to search*, without passing their adapted coordinates to the offspring evaluation.

Diversity-aware Pareto survival. We apply nondominated sorting and objective-space crowding (Deb et al., 2002) to the parent–offspring union. To discourage redundant spatial priors, genomes close to an already selected configuration in residue-role space are deferred, then reconsidered if slots remain. This is a survivor-diversity criterion. For parent population $\mathcal{H}^g$, evaluated offspring $\mathcal{Q}^g$, and survivor population size N_G,

$$\mathcal{H}^{g+1} = \mathcal{S}_G\big(\mathcal{H}^g \cup \mathcal{Q}^g; N_G\big), \qquad \mathcal{Y} = \{(G, \mathcal{X}_G) : G \in \mathcal{H}^{g_{\max}}\}. \tag{12}$$

The output retains each surviving prior with its adapted candidate pool. Algorithm 1 summarizes the search procedure; Apps. A.5 and A.6 detail the diversity rule and execution settings.

4 Experiments

4.1 Experimental Setup

We construct a cross-conformation benchmark covering nine binder categories: small molecules, nonpeptidic macrocycles, linear peptides, cyclic peptides, non-antibody proteins, full-length antibodies, antibody fragments and nanobodies, oligonucleotides, and long nucleic acids including mRNA. Data sources, category sizes, and construction procedures are described in App. B. Initial candidates come from DynamicFlow, FlexSBDD, and YuelDesign for small molecules and nonpeptidic macrocycles, and from BoltzGen, Chai-1, and Protenix for the remaining seven binder categories. We compare each source model's outputs before and after FlexEvo adaptation. Evaluation uses three natural target conformations at the same binding site: *Reference*, *Matched*, and *Divergent*. In this benchmark, Matched preserves the Reference regulatory state, whereas Divergent involves a propeptide-associated state change with local structural rearrangements. Grouping follows biological evidence rather than RMSD thresholds. Adaptation uses only Reference and the initial candidate pool, with source models frozen. The same selected candidate is evaluated across all three states without further adaptation. We report Vina scores and high-affinity fractions for ligand generation, and DockQ, ligand RMSD (LRMSD), and the fraction with LRMSD below 2 Å for complexes with native reference structures. Contact F1, interaction scores, molecular properties, and model confidence provide complementary assessments (App. B.2). These metrics are distinct from the geometric surrogate objectives used during search.

Table 1: Cross-conformation performance across nine binder categories. Reference is the adaptation conformation; Matched and Divergent are alternative natural conformations held out from adaptation. Black entries report post-FLEXEVO scores. Superscripts give changes from the unadapted baseline (Δ = adapted − unadapted); green and red indicate improvement and deterioration according to each metric's preferred direction. Changes in percentage-valued metrics are expressed in percentage points. Vina, DockQ, and QED use means; LRMSD uses medians. Values and changes are rounded independently from the source values. QED is reported once per method in the Reference columns and applies to all three conformations; dashes denote repeated values omitted from display.

A De novo ligand generation

Base model	Vina↓			High affinity (%)↑			Vinardo↑			AutoDock4↑			RTMScore↑			QED↑		
	Reference	Matched	Divergent	Reference	Matched	Divergent	Reference	Matched	Divergent	Reference	Matched	Divergent	Reference	Matched	Divergent	Reference	Matched	Divergent
Small molecule																		
DynamicFlow	-9.18$^{-2.57}$	-8.36$^{-1.89}$	-7.80$^{-2.95}$	58.2$^{+6.9}$	56.7$^{+16.8}$	61.7$^{+35.4}$	.698$^{+.282}$	.537$^{+.131}$	.587$^{+.232}$	.597$^{+.019}$	.570$^{+.141}$	.672$^{+.298}$	.619$^{+.248}$	.594$^{+.135}$	.571$^{+.129}$	.594$^{+.060}$	–	–
FlexSBDD	-9.22$^{-2.32}$	-8.27$^{-1.58}$	-7.74$^{-2.89}$	55.4$^{+8.3}$	57.1$^{+17.8}$	56.5$^{+17.6}$	.552$^{+.142}$	.607$^{+.218}$	.603$^{+.242}$	.585$^{+.244}$	.643$^{+.201}$	.588$^{+.182}$	.532$^{+.204}$	.572$^{+.110}$	.576$^{+.211}$	.589$^{+.215}$	–	–
YuelDesign	-9.18$^{-3.46}$	-8.44$^{-2.86}$	-7.67$^{-3.72}$	62.5$^{+13.0}$	61.3$^{+24.5}$	59.5$^{+28.3}$	.584$^{+.192}$	.643$^{+.269}$	.578$^{+.185}$	.652$^{+.220}$	.616$^{+.177}$	.619$^{+.268}$	.656$^{+.188}$	.584$^{+.171}$	.587$^{+.198}$	.612$^{+.191}$	–	–
Nonpeptidic macrocycle																		
DynamicFlow	-8.32$^{-2.40}$	-7.37$^{-1.63}$	-7.05$^{-2.70}$	51.6$^{+5.1}$	49.9$^{+14.6}$	55.0$^{+31.8}$	.633$^{+.264}$	.486$^{+.125}$	.529$^{+.207}$	.539$^{+.020}$	.508$^{+.121}$	.601$^{+.265}$	.559$^{+.230}$	.538$^{+.132}$	.516$^{+.117}$	.616$^{+.133}$	–	–
FlexSBDD	-8.24$^{-2.00}$	-7.50$^{-1.55}$	-6.83$^{-2.42}$	50.0$^{+8.1}$	50.3$^{+15.7}$	51.1$^{+16.1}$	.487$^{+.120}$	.534$^{+.182}$	.540$^{+.223}$	.531$^{+.229}$	.584$^{+.182}$	.523$^{+.164}$	.475$^{+.179}$	.504$^{+.090}$	.513$^{+.191}$	.535$^{+.204}$	–	–
YuelDesign	-8.23$^{-3.07}$	-7.46$^{-2.47}$	-6.90$^{-3.36}$	56.3$^{+12.2}$	54.9$^{+22.3}$	53.6$^{+25.7}$	.520$^{+.173}$	.582$^{+.248}$	.524$^{+.176}$	.590$^{+.199}$	.549$^{+.158}$	.562$^{+.252}$	.581$^{+.161}$	.528$^{+.153}$	.517$^{+.173}$	.538$^{+.167}$	–	–

B Biomolecular complex design

Base model	DockQ↑			< 2Å (%)↑			LRMSD (Å)↓			MM/GBSA↓			HADDOCK↓			On-target pAE↓		
	Reference	Matched	Divergent	Reference	Matched	Divergent	Reference	Matched	Divergent	Reference	Matched	Divergent	Reference	Matched	Divergent	Reference	Matched	Divergent
Linear peptide																		
BoltzGen	.874$^{+.217}$	.840$^{+.451}$	.850$^{+.566}$	40.2$^{-2.0}$	40.2$^{+13.2}$	36.9$^{+18.5}$	5.32$^{-.16}$	4.82$^{-2.77}$	5.56$^{-5.51}$	-44.6$^{-4.2}$	-44.7$^{-11.3}$	-45.7$^{-26.6}$	-126.6$^{+.5}$	-130.8$^{-33.5}$	-122.3$^{-53.2}$	3.56$^{-.12}$	3.49$^{-2.25}$	3.11$^{-3.71}$
Chai-1	.772$^{+.128}$	.832$^{+.429}$	.822$^{+.533}$	39.7$^{+.0}$	38.6$^{+11.0}$	35.8$^{+17.1}$	4.99$^{-.86}$	4.60$^{-4.09}$	5.99$^{-5.59}$	-43.9$^{-3.3}$	-42.4$^{-9.4}$	-40.2$^{-19.3}$	-124.9$^{-2.0}$	-122.7$^{-11.4}$	-122.0$^{-44.1}$	4.35$^{+.22}$	3.71$^{-2.82}$	4.03$^{-2.26}$
Protenix	.820$^{+.104}$	.863$^{+.394}$	.836$^{+.484}$	39.7$^{-.3}$	38.8$^{+11.0}$	36.4$^{+17.4}$	4.59$^{+.31}$	4.73$^{-3.80}$	4.25$^{-8.03}$	-46.3$^{-1.3}$	-43.2$^{-8.7}$	-40.8$^{-22.3}$	-123.3$^{-.1}$	-123.4$^{-13.4}$	-128.3$^{-51.8}$	3.86$^{+.64}$	3.22$^{-1.01}$	3.08$^{-2.15}$
Cyclic peptide																		
BoltzGen	.902$^{+.219}$	.880$^{+.488}$	.873$^{+.578}$	42.2$^{-1.0}$	41.2$^{+13.6}$	39.1$^{+19.4}$	5.64$^{-.04}$	4.94$^{-2.82}$	5.84$^{-5.27}$	-46.4$^{-3.6}$	-45.1$^{-10.2}$	-47.0$^{-26.8}$	-132.6$^{+.0}$	-133.3$^{-32.5}$	-128.2$^{-55.4}$	3.79$^{-.06}$	3.51$^{-2.59}$	3.29$^{-3.68}$
Chai-1	.861$^{+.005}$	.867$^{+.411}$	.845$^{+.588}$	41.0$^{-.3}$	40.0$^{+5.1}$	37.0$^{+19.3}$	4.74$^{-.90}$	5.05$^{-3.82}$	6.10$^{-6.34}$	-44.7$^{+1.6}$	-44.9$^{-7.2}$	-41.0$^{-18.2}$	-126.4$^{+3.5}$	-122.9$^{-14.8}$	-128.0$^{-29.2}$	4.19$^{+.34}$	4.05$^{-3.08}$	3.98$^{-2.71}$
Protenix	.862$^{+.105}$	.877$^{+.403}$	.854$^{+.500}$	42.1$^{+1.0}$	40.1$^{+11.5}$	37.9$^{+17.8}$	4.71$^{+.26}$	4.92$^{-4.08}$	4.40$^{-8.03}$	-48.5$^{-1.1}$	-46.0$^{-9.3}$	-41.3$^{-21.5}$	-131.2$^{-2.1}$	-130.5$^{-16.4}$	-130.1$^{-49.6}$	4.06$^{+.64}$	3.41$^{-1.06}$	3.27$^{-2.16}$
Non-antibody protein																		
BoltzGen	.862$^{+.219}$	.850$^{+.457}$	.833$^{+.536}$	42.7$^{+4.2}$	39.7$^{+13.9}$	39.0$^{+21.4}$	4.22$^{-1.20}$	4.86$^{-3.32}$	5.90$^{-5.53}$	-40.7$^{-.7}$	-42.7$^{-13.6}$	-38.5$^{-21.5}$	-119.6$^{-1.8}$	-112.1$^{-26.7}$	-117.5$^{-56.1}$	3.87$^{-.19}$	3.17$^{-2.15}$	2.95$^{-4.42}$
Chai-1	.833$^{+.102}$	.807$^{+.335}$	.815$^{+.463}$	44.8$^{+3.5}$	40.8$^{+18.4}$	39.8$^{+24.9}$	3.86$^{-.82}$	5.11$^{-4.11}$	4.97$^{-7.87}$	-39.4$^{-1.0}$	-43.5$^{-12.7}$	-40.3$^{-21.7}$	-111.1$^{+8.3}$	-117.2$^{-19.0}$	-122.0$^{-44.4}$	4.24$^{+.68}$	3.55$^{-.38}$	3.38$^{-1.47}$
Protenix	.818$^{+.152}$	.836$^{+.455}$	.817$^{+.477}$	45.3$^{+1.1}$	41.8$^{+19.3}$	40.2$^{+23.3}$	3.79$^{-.03}$	5.02$^{-4.45}$	4.59$^{-7.92}$	-41.3$^{-.8}$	-41.2$^{-6.2}$	-39.5$^{-22.8}$	-113.9$^{+5.3}$	-111.7$^{-16.7}$	-114.8$^{-45.3}$	3.76$^{-.43}$	3.61$^{-1.11}$	3.83$^{-3.05}$
Full-length antibody																		
BoltzGen	.800$^{+.190}$	.800$^{+.432}$	.772$^{+.496}$	39.9$^{+3.6}$	37.7$^{+13.5}$	37.0$^{+20.6}$	3.98$^{-1.17}$	4.58$^{-3.18}$	5.54$^{-5.07}$	-37.7$^{-.3}$	-39.9$^{-12.3}$	-35.7$^{-19.7}$	-113.6$^{-4.0}$	-104.9$^{-25.0}$	-109.9$^{-52.4}$	3.58$^{-.20}$	3.00$^{-2.02}$	2.78$^{-4.03}$
Chai-1	.791$^{+.110}$	.755$^{+.311}$	.764$^{+.433}$	42.1$^{+2.9}$	38.6$^{+17.6}$	37.5$^{+23.5}$	3.60$^{-.73}$	4.82$^{-3.89}$	4.71$^{-7.44}$	-37.0$^{-1.6}$	-40.5$^{-11.5}$	-38.1$^{-20.7}$	-103.7$^{+8.2}$	-108.5$^{-15.3}$	-115.1$^{-43.3}$	3.97$^{+.62}$	3.34$^{-.34}$	3.18$^{-1.34}$
Protenix	.762$^{+.144}$	.792$^{+.435}$	.770$^{+.456}$	42.9$^{+2.0}$	38.8$^{+17.6}$	37.6$^{+21.7}$	3.53$^{-.01}$	4.75$^{-4.06}$	4.35$^{-7.49}$	-39.1$^{-1.5}$	-38.3$^{-5.3}$	-37.3$^{-21.8}$	-106.2$^{+4.1}$	-104.4$^{-16.2}$	-108.3$^{-42.9}$	3.52$^{-.38}$	3.37$^{-1.11}$	3.59$^{-2.86}$
Antibody fragment and nanobody																		
BoltzGen	.824$^{+.207}$	.822$^{+.442}$	.810$^{+.525}$	41.5$^{+4.3}$	38.0$^{+13.1}$	37.1$^{+20.4}$	4.05$^{-1.19}$	4.69$^{-3.26}$	5.65$^{-5.42}$	-39.2$^{-1.0}$	-41.0$^{-12.8}$	-37.4$^{-21.1}$	-114.8$^{-2.8}$	-107.0$^{-25.5}$	-111.8$^{-52.9}$	3.71$^{-.22}$	3.01$^{-2.09}$	2.85$^{-4.24}$
Chai-1	.802$^{+.105}$	.770$^{+.321}$	.794$^{+.460}$	43.1$^{+3.3}$	39.3$^{+17.6}$	37.9$^{+23.8}$	3.72$^{-.76}$	4.93$^{-3.98}$	4.77$^{-7.64}$	-37.6$^{-.7}$	-41.8$^{-12.1}$	-38.9$^{-21.2}$	-105.7$^{+7.7}$	-112.9$^{-19.0}$	-117.2$^{-42.5}$	4.06$^{+.64}$	3.42$^{-.32}$	3.24$^{-1.39}$
Protenix	.783$^{+.149}$	.814$^{+.450}$	.796$^{+.469}$	43.6$^{+1.3}$	39.9$^{+18.1}$	38.5$^{+22.0}$	3.60$^{-.07}$	4.88$^{-4.13}$	4.46$^{-7.51}$	-40.2$^{-1.3}$	-39.4$^{-5.5}$	-38.4$^{-22.4}$	-108.4$^{+7.2}$	-108.4$^{-17.0}$	-110.9$^{-44.4}$	3.63$^{-.45}$	3.46$^{-1.12}$	3.72$^{-2.90}$
Oligonucleotide																		
BoltzGen	.792$^{+.211}$	.772$^{+.414}$	.768$^{+.497}$	38.9$^{+3.4}$	36.0$^{+12.4}$	35.5$^{+19.7}$	3.87$^{-1.02}$	4.47$^{-2.93}$	5.34$^{-4.95}$	-37.2$^{-.3}$	-38.7$^{-12.0}$	-35.4$^{-19.6}$	-108.9$^{-.7}$	-100.9$^{-22.3}$	-108.3$^{-51.9}$	3.50$^{-.16}$	2.92$^{-1.90}$	2.68$^{-4.03}$
Chai-1	.757$^{+.083}$	.745$^{+.316}$	.737$^{+.413}$	40.5$^{+3.1}$	37.7$^{+17.4}$	36.7$^{+23.0}$	3.51$^{-.80}$	4.61$^{-3.85}$	4.47$^{-7.24}$	-36.2$^{-1.2}$	-40.0$^{-11.9}$	-36.7$^{-19.5}$	-100.5$^{+7.4}$	-106.5$^{-17.9}$	-110.0$^{-38.8}$	3.88$^{+.68}$	3.26$^{-.31}$	3.07$^{-1.32}$
Protenix	.742$^{+.128}$	.759$^{+.407}$	.755$^{+.449}$	40.8$^{+.9}$	38.5$^{+18.2}$	36.3$^{+20.9}$	3.45$^{-.02}$	4.56$^{-4.10}$	4.15$^{-7.20}$	-38.0$^{-.8}$	-38.0$^{-6.0}$	-36.0$^{-21.0}$	-105.3$^{+3.1}$	-101.6$^{-14.3}$	-103.9$^{-41.2}$	3.46$^{-.32}$	3.31$^{-1.02}$	3.48$^{-2.86}$
mRNA / long nucleic acid																		
BoltzGen	.862$^{+.218}$	.849$^{+.457}$	.833$^{+.536}$	45.3$^{+4.0}$	42.1$^{+14.8}$	41.3$^{+22.5}$	4.44$^{-1.35}$	5.19$^{-3.52}$	6.22$^{-5.77}$	-43.7$^{-1.2}$	-45.0$^{-14.0}$	-41.1$^{-22.9}$	-127.0$^{-.9}$	-118.4$^{-27.5}$	-125.9$^{-60.3}$	4.14$^{-.16}$	3.33$^{-2.31}$	3.14$^{-4.61}$
Chai-1	.833$^{+.102}$	.808$^{+.336}$	.814$^{+.463}$	47.9$^{+3.8}$	43.3$^{+19.7}$	42.3$^{+26.6}$	4.09$^{-.88}$	5.38$^{-4.38}$	5.24$^{-8.45}$	-41.9$^{-1.6}$	-45.7$^{-12.7}$	-43.1$^{-23.3}$	-117.2$^{+8.8}$	-125.8$^{-20.3}$	-129.9$^{-48.4}$	4.54$^{+.75}$	3.74$^{-.45}$	3.61$^{-1.48}$
Protenix	.818$^{+.153}$	.836$^{+.455}$	.817$^{+.477}$	47.7$^{+.3}$	44.5$^{+20.5}$	42.9$^{+25.0}$	4.05$^{-.02}$	5.38$^{-4.59}$	4.83$^{-8.33}$	-44.0$^{-.6}$	-44.2$^{-7.4}$	-42.0$^{-24.3}$	-119.7$^{+7.6}$	-117.5$^{-17.7}$	-123.0$^{-49.4}$	3.97$^{-.50}$	3.80$^{-1.25}$	4.08$^{-3.23}$

4.2 MAIN RESULTS

Cross-conformation performance. Table 1 shows improvements in Vina for ligand generation and DockQ for complex design under both Matched and Divergent conditions across the evaluated source–category combinations. For DynamicFlow small molecules, mean Vina under Divergent improves from -4.85 to -7.80 kcal/mol, while the high-affinity fraction increases from 26.4% to 61.7%. For Protenix linear peptides, mean DockQ increases from 0.3519 to 0.8363 under Divergent, compared with 0.7154 to 0.8198 under Reference. Across all 27 source–category combinations, the primary-metric gain is larger under Divergent than under Reference, indicating that the benefit is most pronounced on the biologically divergent held-out state. Importantly, FLEXEVO optimizes only Reference-derived geometric surrogates rather than Vina, DockQ, or other external evaluators; these held-out improvements therefore provide out-of-objective evidence for cross-conformation transfer. Figure 3(a–c) further shows the joint score distributions under paired target conditions, complementing the condition-wise summaries in Table 1.

Input-state quality and metric trade-offs. Reference-state Vina and DockQ also improve in Table 1, so gains on alternative conformations coexist with improved input-state scores. The trend is less uniform across complementary metrics. For example, BoltzGen linear peptides have higher Reference-state DockQ, but the fraction with LRMSD below 2 Å decreases from 42.2% to 40.2%.

Thus, the dominant cross-state gains do not imply uniform improvement on every secondary metric, highlighting the complementary information captured by the reported structural measures.

Geometric basis of initialization. We examine the geometric cues used for FlexBox initialization using 200 ATLAS proteins (Vander Meersche et al., 2024), averaging residue RMSF over three MD simulations per protein. Figure 3(d,h) compares the lowest and highest within-protein quartiles of packing sparsity (inverse C_α weighted contact number) and distance from the whole-chain C_α geometric center. Each quartile's mean RMSF is normalized by the corresponding protein's whole-chain mean. For both descriptors, the highest quartile has greater normalized RMSF on average. These associations motivate the initial allocation of editing freedom; the residue roles are subsequently refined by search.

Multi-objective search behavior. Figure 3(e–g) shows initial, intermediate, and final populations in the plotted objective spaces. Later populations shift toward lower plotted values of f_1 and f_2, while retaining candidates with different objective trade-offs. This spread is consistent with Pareto-based survival preserving alternative preservation–robustness trade-offs rather than collapsing the search to a single operating point. These snapshots characterize internal search behavior under genome-dependent perturbation probes; compatibility with held-out natural conformations is assessed separately in Table 1.

Adaptation dynamics and runtime. Figure 4(a–c) plots evaluation metrics against elapsed adaptation time, linking the attained scores to the computation time of the displayed runs. In panel (a), the high-affinity fraction reaches 60.0% by approximately two minutes and changes little near the end of the run, while $\log P$ continues to increase. In panels (b,c), on-target pAE decreases overall and off-target pAE increases overall over the displayed intervals of approximately 2–3 minutes, with local reversals in some traces.

Structural progression. Figure 4(d–f) provides a structural view of adaptation in one ligand-design example. The early, intermediate, and late candidates have Vina scores of -7.900, -10.491, and -12.666 kcal/mol, respectively. The snapshots show changes in candidate geometry within the binding pocket alongside progressively more favorable docking scores.

4.3 Further Analyses

Among six optimization strategies, FlexEvo ranks first on the reported HV-AUC and B-pocket top-10 scores (App. C.4). Across nine categories, the full method achieves better Divergent-state Vina or DockQ than each of the three component-removal variants (Table 11). Runtime comparisons report an additional 1.4–3.1 minutes of adaptation per sample (App. C.5). Detailed results cover all three target states across 27 source–category combinations (App. C.2). Joint-state plots distinguish simultaneous threshold attainment from attainment in only one target state (App. C.3). Reference-library comparisons quantify candidate similarity to sampled public molecules across nine categories (App. C.6). On/off-target diagnostics report the fraction meeting joint pAE criteria within the evaluated off-target panel (App. C.9). Multimetric profiles compare docking, structural, and molecular-property scores across methods and states (App. C.8). Structure-pool analyses characterize structural composition and biological annotation coverage (App. C.1). Public reference structures and property profiles illustrate binder architectures (App. C.7).

5 Conclusion

We introduced FlexEvo, a model-agnostic framework that adapts binder candidates once at inference time using a single observed target conformation, without retraining the source model. An outer evolutionary search optimizes residue-level constraint configurations, whose decoded *FlexBoxes* guide structural editing and repair. The bi-objective search balances input-state interaction preservation with robustness to local geometric perturbations, using perturbation robustness as a proxy for cross-state compatibility. Experiments across nine binder categories and multiple candidate sources show improved Vina and DockQ scores on natural target conformations unseen during adaptation, while maintaining input-state performance on these metrics. These findings support constrained adaptation of existing model outputs to improve compatibility across target conformations.

AI USE STATEMENT

We used generative AI tools for manuscript writing, language editing, organization, and feedback on the manuscript.

ETHICS STATEMENT

This computational study does not establish biological safety or clinical efficacy; biomedical use requires further validation.

REPRODUCIBILITY STATEMENT

Implementation details, default settings, and evaluation procedures are documented in Apps. A and B.

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.

Rommie E. Amaro, Jerome Baudry, John Chodera, Özlem Demir, J. Andrew McCammon, Yinglong Miao, and Jeremy C. Smith. Ensemble docking in drug discovery. *Biophysical Journal*, 114(10): 2271–2278, 2018. doi: 10.1016/j.bpj.2018.02.038.

David D. Boehr, Dan McElheny, H. Jane Dyson, and Peter E. Wright. The dynamic energy landscape of dihydrofolate reductase catalysis. *Science*, 313(5793):1638–1642, 2006. doi: 10.1126/science.1130258.

David D. Boehr, Ruth Nussinov, and Peter E. Wright. The role of dynamic conformational ensembles in biomolecular recognition. *Nature Chemical Biology*, 5(11):789–796, 2009. doi: 10.1038/nchembio.232.

Andrew A. Bogan and Kurt S. Thorn. Anatomy of hot spots in protein interfaces. *Journal of Molecular Biology*, 280(1):1–9, 1998. doi: 10.1006/jmbi.1998.1843.

Martin Buttenschoen, Garrett M. Morris, and Charlotte M. Deane. PoseBusters: AI-based docking methods fail to generate physically valid poses or generalise to novel sequences. *Chemical Science*, 15(9):3130–3139, 2024. doi: 10.1039/D3SC04185A.

ByteDance AML AI4Science Team, Xinshi Chen, Yuxuan Zhang, Chan Lu, Wenzhi Ma, Jiaqi Guan, Chengyue Gong, Jincai Yang, Hanyu Zhang, Ke Zhang, Shenghao Wu, Kuangqi Zhou, Yanping Yang, Zhenyu Liu, Lan Wang, Bo Shi, Shaochen Shi, and Wenzhi Xiao. Protenix: Advancing structure prediction through a comprehensive AlphaFold3 reproduction. *bioRxiv*, 2025. doi: 10.1101/2025.01.08.631967.

Chai Discovery, Jacques Boitreaud, Jack Dent, Matthew McPartlon, Joshua Meier, Vinicius Reis, Alex Rogozhnikov, and Kevin Wu. Chai-1: Decoding the molecular interactions of life. *bioRxiv*, 2024. doi: 10.1101/2024.10.10.615955.

K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan. A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE Transactions on Evolutionary Computation*, 6(2):182–197, 2002. doi: 10.1109/4235.996017.

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

Jiaqi Guan, Xiangxin Zhou, Yuwei Yang, Yu Bao, Jian Peng, Jianzhu Ma, Qiang Liu, Liang Wang, and Quanquan Gu. DecompDiff: Diffusion models with decomposed priors for structure-based drug design. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 11827–11846. PMLR, 2023b. URL https://proceedings.mlr.press/v202/guan23a.html.

Sheng-You Huang and Xiaoqin Zou. Ensemble docking of multiple protein structures: Considering protein structural variations in molecular docking. *Proteins: Structure, Function, and Bioinformatics*, 66(2):399–421, 2007. doi: 10.1002/prot.21214.

Michal Jamroz, Andrzej Kolinski, and Daisuke Kihara. Structural features that predict real-value fluctuations of globular proteins. *Proteins: Structure, Function, and Bioinformatics*, 80(5):1425–1435, 2012. doi: 10.1002/prot.24040.

Xiangzhe Kong, Yinjun Jia, Wenbing Huang, and Yang Liu. Full-atom peptide design with geometric latent diffusion. In *Advances in Neural Information Processing Systems*, volume 37, pp. 74808–74839. Curran Associates, Inc., 2024. doi: 10.52202/079017-2379.

Chih-Peng Lin, Shao-Wei Huang, Yan-Long Lai, Shih-Chung Yen, Chien-Hua Shih, Chih-Hao Lu, Cuen-Chao Huang, and Jenn-Kang Hwang. Deriving protein dynamical properties from weighted protein contact number. *Proteins*, 72(3):929–935, 2008. doi: 10.1002/prot.21983.

Wei Lu, Jixian Zhang, Weifeng Huang, Ziqiao Zhang, Xiangyu Jia, Zhenyu Wang, Leilei Shi, Chengtao Li, Peter G. Wolynes, and Shuangjia Zheng. DynamicBind: Predicting ligand-specific protein–ligand complex structure with a deep equivariant generative model. *Nature Communications*, 15(1):1071, 2024. doi: 10.1038/s41467-024-45461-2.

Nicholas A. Marze, Shourya S. Roy Burman, William Sheffler, and Jeffrey J. Gray. Efficient flexible backbone protein–protein docking for challenging targets. *Bioinformatics*, 34(20):3461–3469, 2018. doi: 10.1093/bioinformatics/bty355.

Martin Pacesa, Lennart Nickel, Christian Schellhaas, Joseph Schmidt, Ekaterina Pyatova, Lucas Kissling, Patrick Barendse, Jagrity Choudhury, Srajan Kapoor, Ana Alcaraz-Serna, Yehlin Cho, Kourosh H. Ghamary, Laura Vinué, Brahm J. Yachnin, Andrew M. Wollacott, Stephen Buckley, Adrie H. Westphal, Simon Lindhoud, Sandrine Georgeon, Casper A. Goverde, Georgios N. Hatzopoulos, Pierre Gönczy, Yannick D. Muller, Gerald Schwank, Daan C. Swarts, Alex J. Vecchio, Bernard L. Schneider, Sergey Ovchinnikov, and Bruno E. Correia. One-shot design of functional protein binders with BindCraft. *Nature*, 646(8084):483–492, 2025. doi: 10.1038/s41586-025-09429-6.

Xingang Peng, Shitong Luo, Jiaqi Guan, Qi Xie, Jian Peng, and Jianzhu Ma. Pocket2Mol: Efficient molecular sampling based on 3d protein pockets. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 17644–17655. PMLR, 2022. URL https://proceedings.mlr.press/v162/peng22b.html.

Zhuoran Qiao, Weili Nie, Arash Vahdat, Thomas F. Miller, and Animashree Anandkumar. State-specific protein–ligand complex structure prediction with a multiscale deep generative model. *Nature Machine Intelligence*, 6(2):195–208, 2024. doi: 10.1038/s42256-024-00792-z.

Deepa Rajamani, Spencer Thiel, Sandor Vajda, and Carlos J. Camacho. Anchor residues in protein-protein interactions. *Proceedings of the National Academy of Sciences*, 101(31):11287–11292, 2004. doi: 10.1073/pnas.0401942101.

Barak Raveh, Nir London, Lior Zimmerman, and Ora Schueler-Furman. Rosetta FlexPepDock ab-initio: Simultaneous folding, docking and refinement of peptides onto their receptors. *PLOS ONE*, 6(4):e18934, 2011. doi: 10.1371/journal.pone.0018934.

Marion F. Sauer, Alexander M. Sevy, James E. Crowe, and Jens Meiler. Multi-state design of flexible proteins predicts sequences optimal for conformational change. *PLOS Computational Biology*, 16(2):e1007339, 2020. doi: 10.1371/journal.pcbi.1007339.

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

Woody Sherman, Tyler Day, Matthew P. Jacobson, Richard A. Friesner, and Ramy Farid. Novel procedure for modeling ligand/receptor induced fit effects. *Journal of Medicinal Chemistry*, 49 (2):534–553, 2006. doi: 10.1021/jm050540c.

Chien-Hua Shih, Shao-Wei Huang, Shih-Chung Yen, Yan-Long Lai, Sung-Huan Yu, and Jenn-Kang Hwang. A simple way to compute protein dynamics without a mechanical model. *Proteins: Structure, Function, and Bioinformatics*, 68(1):34–38, 2007. doi: 10.1002/prot.21430.

Hannes Stark, Felix Faltings, MinGyu Choi, Yuxin Xie, Eunsu Hur, Timothy O'Donnell, Anton Bushuiev, Talip Uçar, Saro Passaro, Weian Mao, Mateo Reveiz, Roman Bushuiev, Tally Portnoi, Tomáš Pluskal, Josef Sivic, Karsten Kreis, Arash Vahdat, Shamayeeta Ray, Jonathan T. Goldstein, Andrew Savinov, Jacob A. Hambalek, Anshika Gupta, Diego A. Taquiri-Diaz, Yaotian Zhang, Samuel J. Snyder, A. Katherine Hatstat, Angelika Arada, Nam Hyeong Kim, Haoyu Fan, Ethel Tackie-Yarboi, Dylan Boselli, Lee Schnaider, Chang C. Liu, Gene-Wei Li, Denes Hnisz, David M. Sabatini, William F. DeGrado, Jeremy Wohlwend, Gabriele Corso, Regina Barzilay, and Tommi Jaakkola. BoltzGen: Toward universal binder design. *bioRxiv*, 2025. doi: 10.1101/2025.11.20.689494.

Yann Vander Meersche, Gabriel Cretin, Aria Gheeraert, Jean-Christophe Gelly, and Tatiana Galochkina. ATLAS: Protein flexibility description from atomistic molecular dynamics simulations. *Nucleic Acids Research*, 52(D1):D384–D392, 2024. doi: 10.1093/nar/gkad1084.

Jian Wang, Dong Yan Zhang, Shreshty Budakoti, and Nikolay V. Dokholyan. A diffusion-based framework for designing molecules in flexible protein pockets. *bioRxiv*, 2025. doi: 10.1101/2025.05.27.656443.

Joseph L. Watson, David Juergens, Nathaniel R. Bennett, Brian L. Trippe, Jason Yim, Helen E. Eisenach, Woody Ahern, Andrew J. Borst, Robert J. Ragotte, Lukas F. Milles, Basile I. M. Wicky, Nikita Hanikel, Samuel J. Pellock, Alexis Courbet, William Sheffler, Jue Wang, Preetham Venkatesh, Isaac Sappington, Susana Vázquez Torres, Anna Lauko, Valentin De Bortoli, Emile Mathieu, Sergey Ovchinnikov, Regina Barzilay, Tommi S. Jaakkola, Frank DiMaio, Minkyung Baek, and David Baker. De novo design of protein structure and function with RFdiffusion. *Nature*, 620(7976):1089–1100, 2023. doi: 10.1038/s41586-023-06415-8.

Zaixi Zhang, Mengdi Wang, and Qi Liu. FlexSBDD: Structure-based drug design with flexible protein modeling. In *Advances in Neural Information Processing Systems*, volume 37, pp. 53918–53944. Curran Associates, Inc., 2024. doi: 10.52202/079017-1707.

Xiangxin Zhou, Yi Xiao, Haowei Lin, Xinheng He, Jiaqi Guan, Yang Wang, Qiang Liu, Feng Zhou, Liang Wang, and Jianzhu Ma. Integrating protein dynamics into structure-based drug design via full-atom stochastic flows. In *The Thirteenth International Conference on Learning Representations*. OpenReview.net, 2025. URL https://openreview.net/forum?id=9qS3HzSDNv.

APPENDIX CONTENTS

A Additional Details of FlexEvo **14**

A.1 Objects, information flow, and scope 14

A.2 Target Encoding, Screening, and Initialization 16

A.3 Candidate Editing and Structural Safeguards 18

A.4 Single-State Perturbation Scoring 23

A.5 Genome Variation, Feedback, and Survival 26

A.6 Reference Configuration and Execution Schedule 27

B Experimental Protocol and Reproducibility **28**

B.1 Evaluation design 28

B.2 Evaluation metrics 30

B.3 Reproducibility 32

C Supplementary Results and Analyses **34**

C.1 Structure-pool coverage and biological context 34

C.2 Detailed results across nine binder categories 34

C.3 Joint cross-conformation comparisons 37

C.4 Alternative optimizers and component ablations 41

C.5 Runtime and state-wise summary profiles 42

C.6 Similarity to a sampled reference library 44

C.7 Structural examples and physicochemical context 46

C.8 Multimetric profiles 50

C.9 On-target and off-target score diagnostics 54

C.10 Optimization trajectories across binder categories 57

D Deterministic score bounds and conditional robustness of FlexEvo **61**

D.1 Geometry of the spatial prior 61

D.2 Fixed-pose scores and finite-probe bounds 62

D.3 Computable bounds under target perturbations 63

D.4 Transfer from probe neighborhoods 65

D.5 Objective ranges 67

D.6 Selection: what is and is not preserved 68

D.7 Established properties and their role in adaptation 70

APPENDIX OVERVIEW

This supplement separates implementation detail, experimental protocol, supplementary evidence, and mathematical analysis so that each claim can be traced to the information that supports it. Appendix A expands the coordinate-level FLEXEVO core, including target encoding, candidate screening, constrained editing, perturbation scoring, genome evolution, and the reference configuration. Appendix B defines the evaluation design, metric conventions, comparison rules, and information required to reproduce the experiments. Appendix C reports the supplementary results and analyses, including modality-specific tables, cross-conformation comparisons, ablations, contextual data summaries, similarity analyses, and computational cost. Appendix D gives deterministic properties of the implemented geometric scores and states the assumptions needed to transfer finite-probe bounds to an unseen target structure.

The empirical and mathematical parts address different questions. FLEXEVO adapts candidates using a single Reference conformation and evaluates the frozen representatives on held-out Matched and Divergent natural conformations. Reference-derived geometric probes guide the search toward contact preservation and perturbation tolerance. The natural-state benchmark quantifies cross-conformation performance, and the theoretical analysis establishes finite-probe score bounds and transfer enclosures under explicit target-proximity conditions.

Reproducibility at a glance. The reference implementation, default settings, and evaluation boundaries are documented separately from the numerical summaries. The complete core schedule contains 72 genome evaluations, all starting from the common screened pool. Appendix B.3 specifies the experimental records and aggregation rules for reproducible comparisons.

A ADDITIONAL DETAILS OF FLEXEVO

FLEXEVO separates the spatial prior that is optimized by the outer search from the candidate structures adapted by the inner editor. A genome specifies where and how editing is encouraged; evaluating that genome requires a fresh adaptation of the shared source pool. The following subsections specify this information flow and the reference implementation.

A.1 OBJECTS, INFORMATION FLOW, AND SCOPE

For a fixed observed target P^0, let $G = (\mathcal{R}_G, \ell, Z)$ denote a genome: $\mathcal{R}_G$ is the support of assigned residues, ℓ gives their stable, flexible, or forbidden roles, and $Z = (z_1, \ldots, z_R)$ stores a scale for every represented residue. Decoding G produces oriented FlexBoxes. The editor maps the screened pool $\mathcal{X}_0$ to an adapted pool $\mathcal{X}_G$, and the two genome fitnesses summarize scores in that pool. Table 12 distinguishes the objects in this process.

The source model supplies the initial candidates and remains frozen. Every genome evaluation restarts from the same $\mathcal{X}_0$ with the same inner seed. Parent-adapted structures supply feedback about exercised constraints; offspring adaptation starts from the shared source pool. The resulting comparison therefore evaluates each proposed prior through a common adaptation procedure. The final output contains the surviving spatial priors and their corresponding adapted candidate pools.

Molecular records and coordinate representation. A candidate couples an ordered coordinate array with its source molecular record and edit history. The reference implementation stores these as a candidate identifier, coordinates, metadata, and parent-linked edit records. The common representation retains atom-level indexing and coordinates. Structural calculations use the supplied heavy-atom coordinates; hydrogens are handled by the declared scoring-preparation procedure. An edited child receives a new candidate identifier while retaining the correspondence between coordinate indices and source atoms.

For SDF inputs, the loader retains the molecular block, atom identities, bond endpoints and orders, isomeric SMILES, and rotatable-bond annotations. For structure-file inputs, it retains the selected binder's atom names, elements, chains, residue identities, and coordinates. Covalent connectivity and polymer sequence annotations are separate metadata; they are not inferred from spatial prox-

Table 2: Small-molecule drugs. Black: base model. For directional metrics with FlexEvo, green: improved or tied; red: worse. Adapted logP is descriptive and shown in black; arrows indicate preferred directions where defined. QED, SA, Lipinski, logP, and PAINS summaries are presented under Reference. Binding scores are reported separately for Reference, Matched, and Divergent. Diversity summarizes the frozen candidate pool and is reported once under Reference; dashes indicate that it is not repeated for Matched or Divergent.

A. Binding and independent scores

Base model	Vina Mean↓		Vina Median↓		High affinity (%)↑		Vinardo↑		AutoDock4↑		RTMScore↑	
	Base	+FlexEvo	Base	+FlexEvo	Base	+FlexEvo	Base	+FlexEvo	Base	+FlexEvo	Base	+FlexEvo
Reference conformation												
DynamicFlow	-6.6062	-9.1775	-7.0317	-8.7901	51.2524	58.1689	0.4161	0.698	0.5782	0.5971	0.3716	0.6193
FlexSBDD	-6.906	-9.2245	-6.7984	-8.8448	47.0322	55.3719	0.4102	0.5524	0.3411	0.5847	0.3287	0.5322
YuelDesign	-5.7174	-9.1754	-5.4763	-8.8673	49.5198	62.5176	0.392	0.584	0.4317	0.6519	0.4674	0.6556
Matched conformation												
DynamicFlow	-6.4728	-8.3648	-6.573	-8.3392	39.9146	56.6844	0.4056	0.5368	0.4285	0.5696	0.4595	0.594
FlexSBDD	-6.6881	-8.2667	-6.8355	-8.3107	39.2887	57.1034	0.3887	0.6069	0.4422	0.6427	0.4622	0.5723
YuelDesign	-5.58	-8.4389	-5.4642	-8.3575	36.7877	61.2628	0.3737	0.6431	0.4396	0.6163	0.4138	0.5844
Divergent conformation												
DynamicFlow	-4.8519	-7.7991	-4.8465	-8.1933	26.3805	61.7445	0.3548	0.5868	0.3745	0.6723	0.4422	0.5711
FlexSBDD	-4.8552	-7.7438	-4.7988	-8.0889	38.8701	56.5146	0.3615	0.6033	0.4053	0.5875	0.365	0.5755
YuelDesign	-3.9456	-7.6653	-5.2145	-8.2671	31.2482	59.5304	0.3929	0.5776	0.3511	0.6188	0.3889	0.5866

B. Drug-likeness and synthesizability

Base model	QED Mean↑		QED Median↑		SA Mean↑		SA Median↑		Lipinski Mean↑		Lipinski Median↑	
	Base	+FlexEvo	Base	+FlexEvo	Base	+FlexEvo	Base	+FlexEvo	Base	+FlexEvo	Base	+FlexEvo
Reference conformation												
DynamicFlow	0.534	0.5935	0.5416	0.6024	0.6247	0.7097	0.6352	0.706	4.386	4.4522	3.1388	5.2364
FlexSBDD	0.3744	0.589	0.3856	0.5796	0.565	0.6888	0.5832	0.6883	1.7628	3.9447	2.2153	4.3738
YuelDesign	0.4212	0.6118	0.4257	0.5915	0.817	0.803	0.8227	0.8342	3.2247	4.8155	3.7201	4.1107
Matched conformation												
DynamicFlow	–	–	–	–	–	–	–	–	–	–	–	–
FlexSBDD	–	–	–	–	–	–	–	–	–	–	–	–
YuelDesign	–	–	–	–	–	–	–	–	–	–	–	–
Divergent conformation												
DynamicFlow	–	–	–	–	–	–	–	–	–	–	–	–
FlexSBDD	–	–	–	–	–	–	–	–	–	–	–	–
YuelDesign	–	–	–	–	–	–	–	–	–	–	–	–

C. Physicochemical properties and diversity

Base model	logP Mean		logP Median		PAINS pass (%)↑		Diversity Mean↑		Diversity Median↑	
	Base	+FlexEvo	Base	+FlexEvo	Base	+FlexEvo	Base	+FlexEvo	Base	+FlexEvo
Reference conformation										
DynamicFlow	3.2525	3.1108	4.6554	4.2811	90.2154	94.822	0.6719	0.6984	0.6711	0.6965
FlexSBDD	3.6414	4.0985	5.0487	5.3611	87.3202	89.3711	0.6677	0.7094	0.6668	0.7118
YuelDesign	4.3262	4.7377	3.9432	4.7952	99.1444	95.669	0.7213	0.7193	0.7076	0.7126
Matched conformation										
DynamicFlow	–	–	–	–	–	–	–	–	–	–
FlexSBDD	–	–	–	–	–	–	–	–	–	–
YuelDesign	–	–	–	–	–	–	–	–	–	–
Divergent conformation										
DynamicFlow	–	–	–	–	–	–	–	–	–	–
FlexSBDD	–	–	–	–	–	–	–	–	–	–
YuelDesign	–	–	–	–	–	–	–	–	–	–

imity. These records distinguish the candidate being adapted from the receptor used to evaluate it.

Topology-preserving adaptation denotes preservation of the input candidate's atom correspondence and supplied covalent connectivity during coordinate editing. The input molecular identity is retained throughout this procedure. Appendix A.3 specifies the permitted motions, their structural invariants, and the separate molecular-validity checks. Appendix B.2 connects this record to the exported structures read by the evaluators.

Table 3: Nonpeptidic macrocyclic drugs. Black: base model. For directional metrics with FlexEvo, green: improved or tied; red: worse. Adapted logP is descriptive and shown in black; arrows indicate preferred directions where defined. QED, SA, Lipinski, logP, and PAINS summaries are presented under Reference. Binding scores are reported separately for Reference, Matched, and Divergent. Diversity summarizes the frozen candidate pool and is reported once under Reference; dashes indicate that it is not repeated for Matched or Divergent.

A. Binding and independent scores

Base model	Vina Mean↓		Vina Median↓		High affinity (%)↑		Vinardo↑		AutoDock4↑		RTMScore↑	
	Base	+FlexEvo	Base	+FlexEvo	Base	+FlexEvo	Base	+FlexEvo	Base	+FlexEvo	Base	+FlexEvo
Reference conformation												
DynamicFlow	-5.9197	-8.3205	-6.3195	-7.8654	46.5292	51.6061	0.3698	0.6333	0.5181	0.5385	0.3295	0.5594
FlexSBDD	-6.2431	-8.2445	-6.0054	-8.0434	41.9284	50.0367	0.3668	0.4865	0.3012	0.5305	0.2965	0.4753
YuelDesign	-5.1565	-8.2256	-4.8197	-7.9627	44.0439	56.2662	0.3476	0.5204	0.3911	0.5904	0.4201	0.5806
Matched conformation												
DynamicFlow	-5.7327	-7.3665	-5.8834	-7.5453	35.2797	49.8652	0.3604	0.4855	0.3873	0.5082	0.4061	0.5381
FlexSBDD	-5.9464	-7.5012	-6.2064	-7.3676	34.6368	50.3398	0.3519	0.5343	0.4016	0.5837	0.4131	0.5035
YuelDesign	-4.9902	-7.4561	-4.8842	-7.5469	32.6268	54.9453	0.3339	0.5815	0.3911	0.5493	0.3752	0.5283
Divergent conformation												
DynamicFlow	-4.3465	-7.0453	-4.3321	-7.4284	23.1548	54.9721	0.3216	0.5288	0.3364	0.6011	0.3992	0.5162
FlexSBDD	-4.4148	-6.8324	-4.3513	-7.1253	34.9586	51.0609	0.3179	0.5404	0.3596	0.5234	0.3223	0.5134
YuelDesign	-3.5424	-6.8984	-4.6939	-7.5182	27.9735	53.6369	0.3478	0.5238	0.3101	0.5622	0.3445	0.5171

B. Drug-likeness and synthesizability

Base model	QED Mean↑		QED Median↑		SA Mean↑		SA Median↑		Lipinski Mean↑		Lipinski Median↑	
	Base	+FlexEvo	Base	+FlexEvo	Base	+FlexEvo	Base	+FlexEvo	Base	+FlexEvo	Base	+FlexEvo
Reference conformation												
DynamicFlow	0.4837	0.6162	0.4782	0.5802	0.5677	0.6287	0.5683	0.6343	3.8678	3.9499	2.8407	4.7433
FlexSBDD	0.3305	0.5345	0.3497	0.5183	0.5077	0.6071	0.5219	0.6174	1.6004	3.5383	1.9829	3.8934
YuelDesign	0.3705	0.5379	0.3852	0.5276	0.7262	0.7303	0.7246	0.7389	2.9073	4.3297	3.3178	3.6866
Matched conformation												
DynamicFlow	–	–	–	–	–	–	–	–	–	–	–	–
FlexSBDD	–	–	–	–	–	–	–	–	–	–	–	–
YuelDesign	–	–	–	–	–	–	–	–	–	–	–	–
Divergent conformation												
DynamicFlow	–	–	–	–	–	–	–	–	–	–	–	–
FlexSBDD	–	–	–	–	–	–	–	–	–	–	–	–
YuelDesign	–	–	–	–	–	–	–	–	–	–	–	–

C. Physicochemical properties and diversity

Base model	logP Mean		logP Median		PAINS pass (%)↑		Diversity Mean↑		Diversity Median↑	
	Base	+FlexEvo	Base	+FlexEvo	Base	+FlexEvo	Base	+FlexEvo	Base	+FlexEvo
Reference conformation										
DynamicFlow	3.7968	3.7042	4.1455	3.8082	81.0564	85.0685	0.6074	0.6293	0.6058	0.6165
FlexSBDD	3.9796	4.0176	4.5695	4.7844	79.1255	79.6318	0.6068	0.6323	0.6044	0.6438
YuelDesign	3.9256	4.2745	3.5421	4.2951	88.4262	86.7413	0.6374	0.6537	0.6315	0.6312
Matched conformation										
DynamicFlow	–	–	–	–	–	–	–	–	–	–
FlexSBDD	–	–	–	–	–	–	–	–	–	–
YuelDesign	–	–	–	–	–	–	–	–	–	–
Divergent conformation										
DynamicFlow	–	–	–	–	–	–	–	–	–	–
FlexSBDD	–	–	–	–	–	–	–	–	–	–
YuelDesign	–	–	–	–	–	–	–	–	–	–

A.2 Target Encoding, Screening, and Initialization

Residue descriptors and frames. Each target residue has an eight-dimensional descriptor

$$\mathbf{s}_r^P = [p_{r,u}, p_{r,v}, p_{r,w}, d_r, q_r, \theta_r, e_r, b_r]^\top. \tag{13}$$

The first three entries locate its centroid in a principal frame of the centered residue-centroid distribution. The remaining entries are the number of other residue centroids within 8 Å, radial distance from the mean residue centroid, side-chain/radial orientation angle, maximum residue-atom distance from Cα, and the number of target atoms within 6 Å of the residue centroid. The last quantity is a burial proxy. Each column is min–max normalized across residues; constant columns are set to

Table 4: Linear peptide drugs. Black: base model. With FlexEvo, green: improved or tied; red: worse. Arrows indicate the preferred direction. Diversity summarizes the frozen candidate pool and is reported once under Reference; dashes indicate that it is not repeated for Matched or Divergent.

A. Complex accuracy and interaction energy

Base model	DockQ Mean↑		DockQ Median↑		LRMSD < 2Å (%)↑		LRMSD Median (Å)↓		MM/GBSA (kcal/mol)↓		HADDOCK Score↓	
	Base	+FlexEvo	Base	+FlexEvo	Base	+FlexEvo	Base	+FlexEvo	Base	+FlexEvo	Base	+FlexEvo
Reference conformation												
BoltzGen	0.6566	0.8739	0.4699	0.7672	42.1851	40.1549	5.4802	5.3169	-40.3664	-44.5812	-127.0134	-126.5549
Chai-1	0.6434	0.7716	0.4811	0.6535	39.6334	39.6517	5.8455	4.9851	-40.563	-43.8717	-122.8868	-124.9151
Protenix	0.7154	0.8198	0.6704	0.7799	40.0431	39.7062	4.2893	4.5944	-45.0344	-46.2925	-123.1837	-123.3106
Matched conformation												
BoltzGen	0.3893	0.8402	0.1802	0.7719	26.9813	40.2237	7.5894	4.8185	-33.3442	-44.6877	-97.2486	-130.7503
Chai-1	0.403	0.832	0.1923	0.7695	27.578	38.5646	8.6852	4.5988	-33.0482	-42.4377	-111.289	-122.6793
Protenix	0.4692	0.8634	0.2118	0.7511	27.8251	38.8271	8.5333	4.7307	-34.4691	-43.2022	-110.0367	-123.4058
Divergent conformation												
BoltzGen	0.2838	0.8497	0.1074	0.7326	18.4272	36.8895	11.0691	5.5613	-19.1938	-45.7449	-69.1086	-122.2693
Chai-1	0.2895	0.8221	0.1522	0.7759	18.6759	35.8199	11.5766	5.9858	-20.8639	-40.1755	-77.8859	-122.0139
Protenix	0.3519	0.8363	0.1428	0.7943	18.9351	36.3792	12.2729	4.2467	-18.5328	-40.8286	-76.5662	-128.3253

B. Specificity, structural quality and diversity

Base model	On-target pAE↓		Off-target pAE↑		IDDT↑		Diversity Mean↑		Diversity Median↑	
	Base	+FlexEvo	Base	+FlexEvo	Base	+FlexEvo	Base	+FlexEvo	Base	+FlexEvo
Reference conformation										
BoltzGen	3.6716	3.5557	16.6227	17.5337	0.8546	0.8452	0.6249	0.7658	0.7199	0.7042
Chai-1	4.1231	4.3466	16.2316	16.7828	0.8655	0.8676	0.604	0.6031	0.6571	0.657
Protenix	3.2225	3.8616	17.0266	16.1164	0.8893	0.8587	0.6289	0.7613	0.6187	0.6729
Matched conformation										
BoltzGen	5.7389	3.4851	12.6004	17.4258	0.8746	0.8531	–	–	–	–
Chai-1	6.5374	3.7136	11.3432	16.4369	0.8529	0.8557	–	–	–	–
Protenix	4.2246	3.2168	11.9165	18.1324	0.8697	0.8699	–	–	–	–
Divergent conformation										
BoltzGen	6.8166	3.1114	10.8486	17.1859	0.8471	0.8538	–	–	–	–
Chai-1	6.2864	4.0256	11.8326	16.7861	0.8563	0.8665	–	–	–	–
Protenix	5.2334	3.0835	9.9361	17.8956	0.8731	0.8792	–	–	–	–

zero. Missing side-chain directions or Cα atoms use zero for the affected descriptor. The principal frame uses deterministic sign and handedness conventions. FlexBox frames Q_r are constructed from available target anchor triplets, with a residue-atom PCA fallback and an identity fallback for degenerate inputs. The descriptors, centroids, and frames remain fixed during search.

Initial prior. Using the normalized density and burial entries, let $w_r = (d_r + b_r)/2$. The seed assigns roles by the ordered rules

$$\ell_r^{(0)} = \begin{cases} \text{forbidden}, & w_r > 0.72, \\ \text{stable}, & w_r > 0.45 \ \wedge \ q_r < 0.55, \\ \text{flexible}, & q_r > 0.62. \end{cases} \tag{14}$$

Residues satisfying none of these rules lie outside the initial support. Seed scales are one. For the other initial genomes, each residue retains its seed assignment with probability 0.60; otherwise its support/role configuration is sampled uniformly from absence and the three roles. Scales are independently sampled from $[0.80, 1.25]$ and clipped to the allowed range. A scale is stored for every residue, including those outside the support, and is retained on subsequent activation.

The descriptors and thresholds define a heuristic initialization, not a fitted predictor of residue mobility. Stable assignments serve as geometric anchoring priors rather than identified binding-energy hot spots; buried surface area alone need not predict energetic importance (Bogan & Thorn, 1998). The outer search subsequently updates the assignments and scales.

Shared candidate screening. Screening rejects empty or non-finite coordinate sets, candidates with $c(x) < 0.35$, and candidates for which more than 25% of coordinates lie inside any seed forbidden box. Candidates are then processed in input order, retaining at most 64. For equal coordinate counts, geometric similarity is $[1 - \mathrm{RMSD}(x, y)/2.5]_+$, computed in the common target

Table 5: Cyclic peptide drugs. Black: base model. With FlexEvo, green: improved or tied; red: worse. Arrows indicate the preferred direction. Diversity summarizes the frozen candidate pool and is reported once under Reference; dashes indicate that it is not repeated for Matched or Divergent.

A. Complex accuracy and interaction energy

Base model	DockQ Mean↑		DockQ Median↑		LRMSD < 2Å (%)↑		LRMSD Median (Å)↓		MM/GBSA (kcal/mol)↓		HADDOCK Score↓	
	Base	+FlexEvo	Base	+FlexEvo	Base	+FlexEvo	Base	+FlexEvo	Base	+FlexEvo	Base	+FlexEvo
Reference conformation												
BoltzGen	0.6831	0.9018	0.3784	0.8291	43.2014	42.1781	5.6779	5.6422	-42.8373	-46.3974	-132.5755	-132.5663
Chai-1	0.856	0.8613	0.7376	0.7577	41.2663	40.9744	5.6419	4.7438	-46.2865	-44.6798	-129.835	-126.3605
Protenix	0.7567	0.8618	0.7788	0.7398	41.0919	42.0591	4.4517	4.7119	-47.4363	-48.4915	-129.0268	-131.1541
Matched conformation												
BoltzGen	0.3922	0.8802	0.1828	0.7748	27.6297	41.2035	7.7598	4.9429	-34.9468	-45.1455	-100.7753	-133.2734
Chai-1	0.4563	0.8672	0.2715	0.7666	34.8941	40.0107	8.8626	5.0465	-37.7359	-44.9197	-108.0846	-122.8666
Protenix	0.4746	0.8772	0.2183	0.7846	28.6124	40.1201	8.9939	4.9152	-36.6175	-45.9569	-114.0698	-130.4574
Divergent conformation												
BoltzGen	0.2951	0.8734	0.1157	0.7591	19.6438	39.0533	11.1111	5.8433	-20.1633	-46.9776	-72.7891	-128.2151
Chai-1	0.2574	0.8454	0.1628	0.7546	17.6747	36.9939	12.4423	6.0989	-22.781	-40.9813	-98.7699	-127.9845
Protenix	0.3544	0.8544	0.1477	0.7337	20.0467	37.8926	12.4269	4.3972	-19.8152	-41.3023	-80.4316	-130.0567

B. Specificity, structural quality and diversity

Base model	On-target pAE↓		Off-target pAE↑		IDDT↑		Diversity Median↑	
	Base	+FlexEvo	Base	+FlexEvo	Base	+FlexEvo	Base	+FlexEvo
Reference conformation								
BoltzGen	3.8508	3.7891	17.1871	18.2578	0.8879	0.8843	0.7328	0.7264
Chai-1	3.8432	4.1876	17.6561	17.2329	0.8659	0.8757	0.709	0.7147
Protenix	3.4114	4.0563	17.7846	16.9186	0.9152	0.9014	0.6428	0.6915
Matched conformation								
BoltzGen	6.1064	3.5126	13.1615	17.802	0.9026	0.9074	–	–
Chai-1	7.1278	4.0493	12.8907	17.5697	0.8875	0.8898	–	–
Protenix	4.4704	3.4145	12.4549	18.6172	0.9139	0.8991	–	–
Divergent conformation								
BoltzGen	6.9649	3.2892	11.3712	17.6497	0.8739	0.8854	–	–
Chai-1	6.6946	3.9826	11.8993	17.1375	0.8649	0.8735	–	–
Protenix	5.4252	3.2702	10.3118	17.9548	0.8775	0.9164	–	–

frame without an additional alignment. For unequal counts, it is $[1 - \|\bar{p}(x) - \bar{p}(y)\|_2/4]_+$, where $\bar{p}$ is the coordinate centroid. When both molecular fingerprints are available, redundancy requires Morgan-fingerprint Tanimoto similarity of at least 0.92 and geometric similarity of at least 0.90. A candidate without a usable fingerprint is checked by geometric similarity alone. Fingerprints use radius two and 1024 bits.

A.3 Candidate Editing and Structural Safeguards

Fixed-genome adaptation. Each genome evaluation creates a fresh editor with the same random seed. Boxes remain fixed, and candidate-level fusion is disabled in this inner loop. The population size is $N_x = \max\{4, \min\{12, 2|\mathcal{X}_0|\}\}$ for nonempty $\mathcal{X}_0$. It is initialized with the first available seeds, truncated to N_x, and filled with repaired variants when necessary. Each of two adaptation rounds selects up to six reproduction candidates by $f_{\text{pres}} + f_{\text{rob}}$ and generates two proposals per parent. Survivor selection retains the two-objective ranking described in Appendix A.5.

Molecular edits and representation-level operations. The dispatcher first addresses forbidden overlap, then missing anchor proximity, followed by flexible-region contacts and exploratory updates. The reference molecular pipeline selects `graph_preserving` mode. This mode retains the source atom ordering and supplied bond metadata, using whole-candidate translations, proper rotations, and annotated bond-axis torsions. Each proposal therefore edits coordinates associated with a fixed input molecular identity.

The coordinate-level operator family also contains point insertion, point removal, fragment-annotation replacement, and individual-coordinate displacement. These operations belong to the separate `coordinate_only` mode. In that mode, insertion and removal change represented coordinate entries, while replacement changes an annotation. For a selected coordinate and box frame,

Table 6: Non-antibody protein drugs. Black: base model. With FlexEvo, green: improved or tied; red: worse. Arrows indicate the preferred direction. Diversity summarizes the frozen candidate pool and is reported once under Reference; dashes indicate that it is not repeated for Matched or Divergent.

A. Complex accuracy and interaction energy

Base model	DockQ Mean↑		DockQ Median↑		LRMSD < 2Å (%)↑		LRMSD Median (Å)↓		MM/GBSA (kcal/mol)↓		HADDOCK Score↓	
	Base	+FlexEvo	Base	+FlexEvo	Base	+FlexEvo	Base	+FlexEvo	Base	+FlexEvo	Base	+FlexEvo
Reference conformation												
BoltzGen	0.6432	0.8617	0.4529	0.7936	38.5198	42.7348	5.4221	4.2243	-39.9824	-40.6781	-117.7648	-119.6118
Chai-1	0.731	0.8331	0.5799	0.7596	41.3627	44.8159	4.6758	3.8565	-38.3406	-39.3865	-119.3965	-111.0687
Protenix	0.6655	0.8178	0.5644	0.7984	44.2448	45.3202	3.814	3.7879	-40.5226	-41.3208	-119.1796	-113.8899
Matched conformation												
BoltzGen	0.3927	0.8498	0.1842	0.7733	25.807	39.6924	8.1786	4.8552	-29.0773	-42.6594	-85.4292	-112.1446
Chai-1	0.4725	0.8072	0.218	0.7115	22.4389	40.8204	9.2127	5.1052	-30.814	-43.4895	-98.2589	-117.2394
Protenix	0.3815	0.836	0.1979	0.7479	22.5321	41.7923	9.4713	5.0241	-34.9435	-41.1787	-95.0499	-111.7215
Divergent conformation												
BoltzGen	0.2969	0.8328	0.1162	0.7782	17.5496	38.9822	11.4243	5.8963	-17.016	-38.4681	-61.4787	-117.5396
Chai-1	0.3518	0.8147	0.149	0.7893	14.814	39.7527	12.8352	4.9678	-18.6552	-40.3316	-77.5558	-121.9582
Protenix	0.3397	0.817	0.1349	0.6964	16.9771	40.239	12.5131	4.594	-16.6165	-39.4594	-69.4591	-114.7956

B. Specificity, structural quality and diversity

Base model	On-target pAE↓		Off-target pAE↑		IDDT↑		Diversity Mean↑		Diversity Median↑	
	Base	+FlexEvo	Base	+FlexEvo	Base	+FlexEvo	Base	+FlexEvo	Base	+FlexEvo
Reference conformation										
BoltzGen	4.0651	3.8743	15.526	18.6578	0.7992	0.8328	0.706	0.6679	0.7156	0.6684
Chai-1	3.5522	4.2353	18.3426	15.0985	0.807	0.8686	0.6624	0.7584	0.7447	0.7517
Protenix	4.1921	3.7637	18.7125	18.9839	0.8369	0.8861	0.6266	0.6759	0.7084	0.6731
Matched conformation										
BoltzGen	5.3217	3.1726	11.7821	18.6465	0.7986	0.8452	–	–	–	–
Chai-1	3.9359	3.5522	13.1621	19.0971	0.8196	0.8216	–	–	–	–
Protenix	4.7208	3.6133	14.2961	19.2154	0.8254	0.8677	–	–	–	–
Divergent conformation										
BoltzGen	7.3659	2.9474	11.8123	15.5922	0.7645	0.8338	–	–	–	–
Chai-1	4.8468	3.3809	10.6624	18.9681	0.7846	0.8949	–	–	–	–
Protenix	6.8862	3.8313	11.9484	18.7633	0.792	0.8877	–	–	–	–

the displacement is

$$p_i' = p_i + Q_r \delta_i, \qquad \delta_i \sim \mathrm{Unif}([-\epsilon_o, \epsilon_o]^3), \tag{15}$$

with operation-dependent amplitude ϵ_o. This expression describes a coordinate primitive rather than a chemical graph-editing rule. The molecular pipeline disables point insertion, removal, and annotation replacement; its evaluation path rejects `coordinate_only` candidates.

Rigid and torsional structural invariants. The molecular pipeline preserves the source atom correspondence and supplied covalent graph throughout coordinate adaptation. For column-vector coordinates, a rigid proposal takes the form

$$p_i' = Rp_i + t, \qquad R^\top R = I_3, \qquad \det R = 1.$$

Consequently, every pair satisfies

$$\|p_i' - p_j'\|_2^2 = (p_i - p_j)^\top R^\top R (p_i - p_j) = \|p_i - p_j\|_2^2.$$

Whole-binder translations and proper rotations therefore preserve internal distances, bond angles, ring geometry, and the handedness of the input coordinates. Clash-relief repair applies a common translation to the whole binder, preserving these invariants during receptor-overlap relief. The rigid fallback provides the same internal-geometry preservation for structure-file inputs represented by ordered atom coordinates.

A torsional proposal acts on an explicitly annotated rotatable bond (a, b) in the supplied covalent graph. Removing this edge must separate its endpoints into different connected components. The editor rotates the component containing b around the axis through p_a and p_b, preserving all distances within each component. The bond endpoints lie on the rotation axis, and (a, b) is the sole covalent edge between the components. Therefore, the transformation preserves every supplied covalent bond

Table 7: Full-length antibody drugs. Black: base model. With FlexEvo, green: improved or tied; red: worse. Arrows indicate the preferred direction. Diversity summarizes the frozen candidate pool and is reported once under Reference; dashes indicate that it is not repeated for Matched or Divergent.

A. Complex accuracy and interaction energy

Base model	DockQ Mean↑		DockQ Median↑		LRMSD < 2Å (%)↑		LRMSD Median (Å)↓		MM/GBSA (kcal/mol)↓		HADDOCK Score↓	
	Base	+FlexEvo	Base	+FlexEvo	Base	+FlexEvo	Base	+FlexEvo	Base	+FlexEvo	Base	+FlexEvo
Reference conformation												
BoltzGen	0.6095	0.7995	0.3297	0.6427	36.2452	39.8647	5.1441	3.9774	-37.4543	-37.7412	-109.6532	-113.6292
Chai-1	0.6808	0.7909	0.4454	0.6112	39.1404	42.0743	4.3318	3.6037	-35.4734	-37.0303	-111.8543	-103.6798
Protenix	0.6178	0.7616	0.3397	0.5605	40.9499	42.9017	3.5419	3.5284	-37.6326	-39.1405	-110.3069	-106.1588
Matched conformation												
BoltzGen	0.3681	0.8001	0.1736	0.6377	24.2284	37.6849	7.7549	4.5763	-27.6118	-39.8732	-79.8202	-104.8587
Chai-1	0.4446	0.7553	0.2056	0.5835	21.0076	38.5872	8.7089	4.8166	-29.0343	-40.5187	-93.2411	-108.4916
Protenix	0.3579	0.7924	0.1837	0.5997	21.1998	38.8209	8.8142	4.7497	-32.9373	-38.2532	-88.2523	-104.4242
Divergent conformation												
BoltzGen	0.2761	0.7716	0.1101	0.5899	16.3655	37.0072	10.6048	5.5371	-15.9469	-35.6724	-57.4783	-109.9148
Chai-1	0.3305	0.7636	0.1387	0.5516	13.9525	37.4948	12.1503	4.7071	-17.4054	-38.0685	-71.8189	-115.1271
Protenix	0.3143	0.7702	0.1251	0.5559	15.9464	37.6242	11.8363	4.3461	-15.5366	-37.3434	-65.3604	-108.2527

B. Specificity, structural quality and diversity

Base model	On-target pAE↓		Off-target pAE↑		IDDT↑		Diversity Mean↑		Diversity Median↑	
	Base	+FlexEvo	Base	+FlexEvo	Base	+FlexEvo	Base	+FlexEvo	Base	+FlexEvo
Reference conformation										
BoltzGen	3.7784	3.5821	14.7108	17.4877	0.7422	0.7883	0.6694	0.6205	0.6772	0.6339
Chai-1	3.3538	3.9717	17.2236	14.1967	0.7577	0.8186	0.6273	0.7038	0.6962	0.7085
Protenix	3.8994	3.5152	17.5223	17.9927	0.7751	0.8295	0.5801	0.6293	0.6718	0.6342
Matched conformation										
BoltzGen	5.0256	3.0042	11.1672	17.4201	0.7503	0.7898	–	–	–	–
Chai-1	3.6862	3.3446	12.1739	17.6798	0.7719	0.7682	–	–	–	–
Protenix	4.4763	3.3697	13.5659	17.9164	0.7682	0.8113	–	–	–	–
Divergent conformation										
BoltzGen	6.8085	2.7801	11.1069	14.7692	0.7245	0.7794	–	–	–	–
Chai-1	4.5198	3.1794	9.8693	17.9056	0.7353	0.8479	–	–	–	–
Protenix	6.4495	3.5896	11.1696	17.4772	0.7407	0.8209	–	–	–	–

length and the graph connectivity. The input graph includes ring-closure and cross-link edges, including declared disulfides. Ring edges retain an alternative connecting path and are excluded from this torsional operation. The input adapter provides the complete covalent graph and chemically eligible rotatable-bond annotations. Candidates without an eligible annotation use whole-binder rigid motions, retaining their input internal geometry.

Proposal acceptance and geometric repair. FlexBoxes guide proposal locations through anchor proximity, adaptable regions, and forbidden-region avoidance. Molecular repair uses up to two whole-binder displacement passes to relieve receptor overlap while preserving the candidate's internal geometry. Each repaired proposal must pass the acceptance gate before survivor selection. Acceptance requires nonempty, finite coordinates, neighbor support $c(\widehat{x}) \geq 0.35$, and the same coordinate-array shape and supplied bond list as the parent. For every supplied covalent edge (i, j), acceptance additionally requires

$$|\|p_i(\widehat{x}) - p_j(\widehat{x})\|_2 - \|p_i(x) - p_j(x)\|_2| \leq 10^{-5}\ \text{Å}.$$

This tolerance enforces numerical preservation of the parent bond geometry. The gate rejects proposals violating these structural conditions before objective-based survivor selection. Accepted proposals are subsequently ranked using interaction preservation, perturbation robustness, and the spatial penalties in Appendix A.4.

Covalent-graph preservation is enforced by the permitted transformations and supplied-graph checks; $c(x)$ quantifies local spatial neighbor support. The default configuration applies these structural guards. Additional candidate-level tests are available through the configurable `validity_hook`, whose default value is `None`. The category-specific chemical and physical validation criteria are specified below, with their outcomes attached to the exported molecular record.

Table 8: Antibody fragments and nanobodies. Black: base model. With FlexEvo, green: improved or tied; red: worse. Arrows indicate the preferred direction. Diversity summarizes the frozen candidate pool and is reported once under Reference; dashes indicate that it is not repeated for Matched or Divergent.

A. Complex accuracy and interaction energy

Base model	DockQ Mean↑		DockQ Median↑		LRMSD < 2Å (%)↑		LRMSD Median (Å)↓		MM/GBSA (kcal/mol)↓		HADDOCK Score↓	
	Base	+FlexEvo	Base	+FlexEvo	Base	+FlexEvo	Base	+FlexEvo	Base	+FlexEvo	Base	+FlexEvo
Reference conformation												
BoltzGen	0.6169	0.8239	0.3372	0.6666	37.1832	41.5067	5.2336	4.0466	-38.1485	-39.1582	-112.0832	-114.8399
Chai-1	0.6969	0.8022	0.4589	0.6322	39.8487	43.1363	4.4744	3.7165	-36.8421	-37.5865	-113.4421	-105.6989
Protenix	0.6342	0.7832	0.3542	0.5768	42.3102	43.6115	3.6699	3.5971	-38.8703	-40.1934	-115.6324	-108.4364
Matched conformation												
BoltzGen	0.3797	0.8216	0.1778	0.6494	24.9698	38.0403	7.9436	4.6858	-28.2069	-41.0014	-81.4757	-106.9819
Chai-1	0.4487	0.7698	0.2102	0.5987	21.7351	39.3352	8.9106	4.9298	-29.7421	-41.8244	-93.9058	-112.8884
Protenix	0.3643	0.814	0.1877	0.652	21.7921	39.8784	9.0096	4.8803	-33.9544	-39.4471	-91.432	-108.4257
Divergent conformation												
BoltzGen	0.2855	0.81	0.1111	0.648	16.6732	37.1021	11.0622	5.6455	-16.2307	-37.3532	-58.8742	-111.7947
Chai-1	0.3337	0.7935	0.1419	0.5987	14.1025	37.9284	12.4077	4.7659	-17.7388	-38.9375	-74.7328	-117.2265
Protenix	0.327	0.796	0.1273	0.6219	16.4976	38.4876	11.973	4.4619	-16.0826	-38.4459	-66.5255	-110.9338

B. Specificity, structural quality and diversity

Base model	On-target pAE↓		Off-target pAE↑		IDDT↑		Diversity Mean↑		Diversity Median↑	
	Base	+FlexEvo	Base	+FlexEvo	Base	+FlexEvo	Base	+FlexEvo	Base	+FlexEvo
Reference conformation										
BoltzGen	3.9393	3.7144	14.9833	17.9637	0.7597	0.8076	0.6838	0.6469	0.6817	0.6396
Chai-1	3.4169	4.0604	17.5468	14.5896	0.7727	0.8251	0.6333	0.7226	0.7202	0.7158
Protenix	4.0851	3.633	18.2213	18.3084	0.8082	0.8495	0.6028	0.6494	0.6795	0.6545
Matched conformation										
BoltzGen	5.1033	3.0136	11.2482	17.7218	0.7736	0.8093	–	–	–	–
Chai-1	3.7385	3.4188	12.5764	18.1712	0.7959	0.7994	–	–	–	–
Protenix	4.5795	3.4569	13.8595	18.6504	0.7866	0.8303	–	–	–	–
Divergent conformation										
BoltzGen	7.0902	2.8493	11.3796	14.9092	0.7446	0.7929	–	–	–	–
Chai-1	4.6378	3.2444	10.2362	18.0941	0.7453	0.8518	–	–	–	–
Protenix	6.6159	3.7161	11.525	18.0938	0.7646	0.8427	–	–	–	–

Category-specific molecular-validity specification. Molecular validation is defined on the exported candidate's chemical identity and three-dimensional structure. For fixed-identity adaptation, identity is compared with that candidate's own source ancestor, not with a chemically different native binder. The validation record separates graph identity, stereochemical identity, internal geometry, and receptor-dependent overlaps. The applicable checks are as follows.

Conventional small molecules and nonpeptidic macrocycles. The identity comparison includes atom identities, bond orders, formal charges, and stereochemical assignments. Geometry checks include bond lengths and angles, internal clashes, and the geometry of each declared ring closure. For supported ligand chemistries, PoseBusters provides chemical-consistency and physical-plausibility checks, including stereochemistry, bond geometry, planarity, and protein–ligand clashes (Buttenschoen et al., 2024). The tool version, configuration, and per-check outcomes define the reported validation result.

Linear and cyclic peptides. The molecular record preserves residue identity and order, peptide connectivity, termini, and declared modifications. Validation covers stereochemistry, peptide-bond geometry, backbone continuity, and internal overlaps. Cyclic peptides additionally require their declared closure bonds and closure geometry to be checked. Their ring bonds remain subject to the torsional restriction described above.

Non-antibody proteins, full-length antibodies, and antibody fragments or nanobodies. The identity specification includes chain and residue correspondence, covalent connectivity, and declared disulfide or other cross-links. Structural checks cover backbone continuity, local stereochemistry, bond geometry, and steric overlaps. Local side-chain or backbone motions require explicit chemistry-aware atom partitions and rotatable-bond annotations; otherwise, the reference editor preserves the complete input binder through rigid motions.

Table 9: Oligonucleotide drugs. Black: base model. With FlexEvo, green: improved or tied; red: worse. Arrows indicate the preferred direction. Diversity summarizes the frozen candidate pool and is reported once under Reference; dashes indicate that it is not repeated for Matched or Divergent.

A. Complex accuracy and interaction energy

Base model	DockQ Mean↑		DockQ Median↑		LRMSD < 2Å (%)↑		LRMSD Median (Å)↓		MM/GBSA (kcal/mol)↓		HADDOCK Score↓	
	Base	+FlexEvo	Base	+FlexEvo	Base	+FlexEvo	Base	+FlexEvo	Base	+FlexEvo	Base	+FlexEvo
Reference conformation												
BoltzGen	0.5812	0.7918	0.3225	0.6375	35.4166	38.8602	4.8907	3.8668	-36.9099	-37.2423	-108.2698	-108.9199
Chai-1	0.6744	0.7572	0.4361	0.5988	37.3752	40.4985	4.3123	3.5139	-35.0184	-36.1978	-107.8819	-100.5058
Protenix	0.6139	0.7419	0.3305	0.5449	39.9387	40.8061	3.4706	3.4549	-37.2316	-38.0442	-108.3421	-105.2553
Matched conformation												
BoltzGen	0.3579	0.7723	0.1701	0.6208	23.6166	36.0407	7.3954	4.4653	-26.7155	-38.7042	-78.6222	-100.9389
Chai-1	0.4294	0.7453	0.1997	0.5583	20.3407	37.7066	8.4591	4.6118	-28.1806	-40.0486	-88.5841	-106.4813
Protenix	0.3524	0.7594	0.1782	0.5953	20.3365	38.4897	8.6653	4.5631	-32.0128	-38.0237	-87.2657	-101.5998
Divergent conformation												
BoltzGen	0.2703	0.7675	0.1044	0.5704	15.8563	35.5326	10.2976	5.3444	-15.7288	-35.3631	-56.4482	-108.3156
Chai-1	0.3235	0.7369	0.1365	0.5358	13.6982	36.7154	11.7157	4.4727	-17.2152	-36.6873	-71.2834	-110.0478
Protenix	0.3061	0.7554	0.1216	0.5373	15.4408	36.2917	11.3569	4.1536	-15.0455	-36.0455	-62.6656	-103.9151

B. Specificity, structural quality and diversity

Base model	On-target pAE↓		Off-target pAE↑		IDDT↑		Diversity Mean↑		Diversity Median↑	
	Base	+FlexEvo	Base	+FlexEvo	Base	+FlexEvo	Base	+FlexEvo	Base	+FlexEvo
Reference conformation										
BoltzGen	3.6597	3.4954	14.3358	17.0931	0.7219	0.7683	0.6439	0.6115	0.6561	0.6024
Chai-1	3.2037	3.8848	16.6055	13.6782	0.7365	0.7961	0.6047	0.6959	0.6827	0.6917
Protenix	3.7768	3.4593	17.1694	17.0991	0.7654	0.8125	0.5747	0.6079	0.6453	0.6162
Matched conformation										
BoltzGen	4.8252	2.9219	10.6104	16.7826	0.7184	0.7681	–	–	–	–
Chai-1	3.5689	3.2574	11.9516	17.4595	0.7563	0.7527	–	–	–	–
Protenix	4.3303	3.3073	13.1951	17.4349	0.7532	0.7836	–	–	–	–
Divergent conformation										
BoltzGen	6.7059	2.6769	10.7163	14.1959	0.7047	0.7679	–	–	–	–
Chai-1	4.3856	3.0669	9.8296	17.1495	0.7066	0.8257	–	–	–	–
Protenix	6.3374	3.4789	11.0271	17.2584	0.7149	0.8029	–	–	–	–

Oligonucleotides and mRNA or other long nucleic acids. The identity specification includes nucleotide order, chain identity, sugar and base identities, modifications, and phosphodiester connectivity. Structural checks cover sugar stereochemistry, backbone continuity, bond geometry, base geometry, and internal overlaps. A small-molecule validation configuration is not treated as a universal polymer validator.

Validation records and frozen-candidate evaluation. The validation specification assigns a pass when every applicable test in the declared configuration succeeds. Each record retains the outcome and applicability of individual checks, alongside the structural acceptance conditions defined above. Intrinsic checks apply to the frozen candidate, and receptor-dependent overlaps are evaluated for Reference, Matched, and Divergent separately. The same representative and coordinates are retained throughout these state-specific evaluations. Geometry-changing refinement receives its own structure identifier, validation record, and score record. Appendices B.2 and B.3 specify the export interfaces and reporting conventions.

Coordinate preservation in Rerank-only. Rerank-only operates on candidate rankings and performs no translation, rotation, torsional update, local minimization, or geometric repair. The selected structure retains the source candidate's atom ordering, coordinates, and molecular identity. Its output record preserves the source identifier and coordinate checksum, linking each evaluation input to the selected source structure. The validation and failure-handling policy is shared across the three arms. Evaluation checks preserve the frozen representative and do not trigger candidate replacement or target-state-specific repair.

Table 10: mRNA and long nucleic acid drugs. Black: base model. With FlexEvo, green: improved or tied; red: worse. Arrows indicate the preferred direction. Diversity summarizes the frozen candidate pool and is reported once under Reference; dashes indicate that it is not repeated for Matched or Divergent.

A. Complex accuracy and interaction energy

Base model	DockQ Mean↑		DockQ Median↑		LRMSD < 2Å (%)↑		LRMSD Median (Å)↓		MM/GBSA (kcal/mol)↓		HADDOCK Score↓	
	Base	+FlexEvo	Base	+FlexEvo	Base	+FlexEvo	Base	+FlexEvo	Base	+FlexEvo	Base	+FlexEvo
Reference conformation												
BoltzGen	0.6435	0.8616	0.3526	0.7935	41.2659	45.2835	5.7891	4.4419	-42.5524	-43.7174	-126.0592	-126.9724
Chai-1	0.7316	0.8332	0.4798	0.7594	44.0388	47.8695	4.9724	4.0948	-40.3513	-41.9244	-125.9534	-117.1707
Protenix	0.6651	0.8176	0.3642	0.6987	47.3985	47.6928	4.0711	4.0487	-43.3554	-43.9896	-127.3749	-119.7496
Matched conformation												
BoltzGen	0.3928	0.8493	0.1849	0.7736	27.2126	42.0513	8.7107	5.1881	-30.9788	-44.9613	-90.9291	-118.4181
Chai-1	0.472	0.8078	0.2188	0.6159	23.6169	43.3125	9.7664	5.3827	-33.0369	-45.6925	-105.4453	-125.7659
Protenix	0.3811	0.836	0.1977	0.7471	23.9538	44.4931	9.9701	5.3765	-36.7379	-44.1818	-99.8345	-117.4869
Divergent conformation												
BoltzGen	0.2962	0.8325	0.1163	0.7286	18.8527	41.3252	11.9946	6.2245	-18.2159	-41.0712	-65.6305	-125.9276
Chai-1	0.3515	0.8144	0.1496	0.6898	15.6241	42.2652	13.6901	5.2412	-19.8488	-43.1475	-81.5273	-129.8867
Protenix	0.3398	0.8172	0.1342	0.6866	17.8574	42.8658	13.1507	4.8256	-17.6284	-41.9619	-73.6176	-123.0036

B. Specificity, structural quality and diversity

Base model	On-target pAE↓		Off-target pAE↑		IDDT↑		Diversity Mean↑		Diversity Median↑	
	Base	+FlexEvo	Base	+FlexEvo	Base	+FlexEvo	Base	+FlexEvo	Base	+FlexEvo
Reference conformation										
BoltzGen	4.2939	4.1355	16.4981	19.7682	0.8398	0.8902	0.7457	0.7047	0.7517	0.7162
Chai-1	3.7891	4.5393	19.4026	16.1871	0.8576	0.9162	0.6956	0.8026	0.7985	0.8059
Protenix	4.4709	3.9701	20.0858	20.3136	0.8913	0.9303	0.6676	0.7105	0.7538	0.7118
Matched conformation										
BoltzGen	5.6365	3.3301	12.4661	19.8712	0.8514	0.8984	–	–	–	–
Chai-1	4.1981	3.7439	13.8841	20.3667	0.8666	0.8646	–	–	–	–
Protenix	5.0484	3.7991	15.3096	20.4576	0.8822	0.9301	–	–	–	–
Divergent conformation										
BoltzGen	7.7522	3.1443	12.6249	16.4044	0.8052	0.8855	–	–	–	–
Chai-1	5.0844	3.6057	11.4036	20.0854	0.8252	0.9585	–	–	–	–
Protenix	7.3069	4.0782	12.6093	19.7727	0.8434	0.9398	–	–	–	–

A.4 Single-State Perturbation Scoring

Target probes. Let y_a^0 denote target atom a, $r(a)$ its residue, and o the input pocket center. Radial directions are $\widehat{\mathbf{r}}_a = (y_a^0 - o)/\|y_a^0 - o\|_2$. For each atom, its tangential direction is obtained by projecting the centered target atoms onto the plane orthogonal to $\widehat{\mathbf{r}}_a$ and normalizing the projection with largest norm. At the center, a farthest-target direction supplies the radial fallback; a degenerate tangential direction is set to zero. Thus directions are atom dependent, while the perturbation strength is assigned at residue level.

For flexible residues, $\phi_r = 0.85$. The displacement amplitude is $A_r = \min\{1.65, 0.45 + 1.95\phi_r\}$. For atoms in these residues, the six displacement modes are

$$\begin{aligned} \Delta_a^{(1,2)} &= \pm A_{r(a)}\widehat{\mathbf{r}}_a, \\ \Delta_a^{(3,4)} &= \pm 0.75 A_{r(a)}\widehat{\mathbf{t}}_a, \\ \Delta_a^{(5)} &= 0.55 A_{r(a)}(\widehat{\mathbf{r}}_a + \widehat{\mathbf{t}}_a), \\ \Delta_a^{(6)} &= 0.55 A_{r(a)}(-\widehat{\mathbf{r}}_a + \widehat{\mathbf{t}}_a). \end{aligned} \tag{16}$$

Other atoms remain unchanged. If no flexible residue is present, non-anchor residues provide a fallback with $\phi_r = 0.22$. The reference target and six perturbed targets form the probe collection; candidate coordinates are fixed throughout scoring. All distances and displacement amplitudes below are numerical values in angstroms.

Table 11: Ablations across nine binder categories. Prior denotes the flexibility prior, and FLEXEVO denotes the full model with all components enabled. Green/bold and light green indicate the best and second-best results, respectively, within each category, metric, and target conformation; displayed ties share the same rank. We evaluate three one-component-removal variants by removing the prior, FlexBox, or evolution module from the full model. QED is reported only for the Reference conformation because the molecular graph is unchanged across target conformations; repeated Matched and Divergent values are omitted.

A. De novo ligand generation

Variant	Vina↓			High affinity (%)↑			Vinardo↑			AutoDock4↑			RTMScore↑			QED↑		
	Reference	Matched	Divergent	Reference	Matched	Divergent	Reference	Matched	Divergent	Reference	Matched	Divergent	Reference	Matched	Divergent	Reference	Matched	Divergent
Small molecule																		
w/o prior	-7.50	-7.31	-5.98	50.2	45.6	46.1	.522	.549	.503	.618	.593	.575	.487	.496	.476	.534	–	–
w/o FlexBox	**-8.03**	-7.45	-5.88	**52.7**	38.0	41.6	.536	.547	.502	.537	.567	.590	.494	.505	.481	.533	–	–
w/o evolution	-7.92	-7.00	-5.92	47.8	39.6	43.9	.482	.492	.527	.594	.531	.532	.493	.439	.441	.527	–	–
FlexEvo	-8.01	**-7.88**	**-6.66**	48.0	**48.1**	**49.8**	**.621**	**.627**	**.534**	**.668**	**.655**	**.675**	**.551**	**.525**	**.516**	**.558**	–	–
Nonpeptidic macrocycle																		
w/o prior	-8.80	-8.82	-7.02	58.7	54.5	55.8	.617	.665	.594	.623	.679	.672	.583	.614	.566	.651	–	–
w/o FlexBox	-9.55	-8.93	-7.04	**62.8**	45.5	48.9	.620	.646	.590	.641	.668	.627	.589	.593	**.682**	.629	–	–
w/o evolution	-9.33	-8.33	-6.94	56.2	46.8	50.9	.567	.581	.619	.583	.640	.629	.593	.527	.523	.627	–	–
FlexEvo	**-9.63**	**-9.36**	**-8.02**	56.4	**56.5**	**58.5**	**.731**	**.749**	**.639**	**.721**	**.712**	**.704**	**.664**	**.724**	.608	**.686**	–	–

B. Biomolecular complex design

Variant	DockQ↑			< 2Å (%)↑			LRMSD (Å)↓			MM/GBSA↓			HADDOCK↓			On-target pAE↓		
	Reference	Matched	Divergent	Reference	Matched	Divergent	Reference	Matched	Divergent	Reference	Matched	Divergent	Reference	Matched	Divergent	Reference	Matched	Divergent
Linear peptide																		
w/o prior	.734	.568	.348	40.3	28.5	17.3	**4.18**	7.88	12.41	-44.3	**-45.4**	-42.1	-115.1	-118.3	-114.8	3.59	4.38	4.84
w/o FlexBox	.752	.516	.356	40.4	25.0	19.6	4.38	8.77	11.06	-42.0	-36.0	-34.7	-115.2	-117.3	-86.4	3.95	4.37	4.82
w/o evolution	.660	.418	.308	37.4	28.5	17.2	5.02	7.81	12.28	-40.7	-37.0	-31.2	-115.1	-114.6	-99.8	3.76	3.87	4.81
FlexEvo	**.792**	**.780**	**.760**	**40.9**	**40.4**	**37.8**	4.87	**4.28**	**5.01**	**-44.8**	-44.5	**-42.4**	**-117.4**	**-121.4**	**-124.9**	**3.29**	**3.11**	**2.89**
Cyclic peptide																		
w/o prior	.765	.571	.357	41.5	29.3	17.9	**4.22**	8.02	12.49	-45.0	-45.1	-42.6	-117.1	-119.2	-115.4	3.63	4.38	4.85
w/o FlexBox	.765	.534	.362	41.5	25.7	20.2	4.49	8.92	11.04	-42.6	-36.6	-35.7	-116.6	-117.8	-87.0	4.00	4.40	4.85
w/o evolution	.684	.417	.311	38.4	28.6	17.6	5.12	7.94	12.58	-41.8	-38.4	-31.9	-116.8	-115.3	-100.7	3.89	3.98	4.86
FlexEvo	**.802**	**.799**	**.772**	**42.2**	**41.4**	**38.2**	5.05	**4.35**	**5.20**	**-46.0**	**-45.4**	**-43.5**	**-121.1**	**-123.8**	**-127.6**	**3.33**	**3.10**	**2.92**
Non-antibody protein																		
w/o prior	.807	.606	.384	43.5	30.6	18.9	**4.53**	8.73	13.29	**-49.3**	-48.3	**-46.2**	-127.3	**-130.6**	-122.7	3.89	4.83	5.14
w/o FlexBox	.827	.571	.386	43.8	27.9	21.3	4.84	9.88	11.73	-46.4	-40.3	-37.9	-127.9	-125.7	-94.0	3.41	4.79	5.30
w/o evolution	.736	.444	.340	41.1	30.9	19.2	5.35	8.51	13.41	-44.0	-41.6	-35.0	**-128.2**	-121.7	-107.8	4.14	4.18	5.13
FlexEvo	**.860**	**.862**	**.842**	**44.6**	**45.6**	**41.6**	5.40	**4.75**	**5.55**	-49.1	**-50.2**	-45.8	-128.0	-129.6	**-136.5**	**3.26**	**3.32**	**3.20**
Full-length antibody																		
w/o prior	.723	.547	.337	39.8	27.4	17.1	**4.02**	7.57	12.08	-42.5	-43.7	-41.1	-112.5	-114.6	-112.1	3.44	4.29	4.63
w/o FlexBox	.715	.506	.340	39.1	24.1	19.2	4.32	8.59	10.66	-41.1	-35.1	-33.8	-112.2	-113.3	-85.2	3.86	4.16	4.63
w/o evolution	.648	.404	.303	36.6	27.4	16.9	4.79	7.68	11.92	-39.8	-36.5	-31.0	-112.6	-110.3	-98.0	3.67	3.75	4.71
FlexEvo	**.763**	**.762**	**.746**	**40.7**	**40.1**	**37.0**	4.84	**4.19**	**4.95**	**-43.0**	**-44.4**	**-41.5**	**-116.7**	**-117.4**	**-122.6**	**3.17**	**3.03**	**2.85**
Antibody fragment and nanobody																		
w/o prior	.742	.569	.347	40.8	28.2	17.5	4.15	7.92	12.38	-44.5	-44.3	**-42.3**	-114.7	-119.0	-115.6	3.56	4.40	4.76
w/o FlexBox	.741	.524	.350	40.1	25.0	19.7	4.37	8.89	10.94	-42.0	-36.5	-34.5	-117.1	-116.7	-87.6	3.93	4.29	4.77
w/o evolution	.674	.410	.305	37.7	28.3	17.3	4.98	7.81	12.47	-40.5	-37.9	-31.4	-115.9	-114.1	-100.8	3.79	3.89	4.74
FlexEvo	**.784**	**.778**	**.766**	**41.7**	**40.8**	**38.0**	**3.93**	**4.25**	**5.06**	**-44.9**	**-44.9**	-41.9	**-118.6**	**-122.4**	**-124.4**	**3.29**	**3.13**	**2.94**
Oligonucleotide																		
w/o prior	.671	.527	.322	38.7	25.6	16.5	4.84	7.23	11.30	-40.0	-40.5	-39.7	-105.4	-108.1	-106.7	3.26	4.17	4.39
w/o FlexBox	.685	.470	.328	37.6	23.0	18.0	4.15	8.29	10.35	-38.9	-33.5	-32.1	-104.3	-105.8	-82.9	3.67	3.88	4.42
w/o evolution	.619	.387	.284	34.7	26.4	15.9	4.49	7.44	11.39	-37.9	-35.3	-29.7	-104.6	-106.2	-92.6	3.48	3.59	4.38
FlexEvo	**.708**	**.714**	**.714**	**39.6**	**38.2**	**35.0**	**3.76**	**4.01**	**4.82**	**-41.2**	**-42.0**	**-40.4**	**-111.7**	**-113.1**	**-118.8**	**2.94**	**2.82**	**2.65**
mRNA / long nucleic acid																		
w/o prior	.785	.600	.371	43.7	29.9	18.7	4.43	8.44	13.16	**-46.9**	-47.0	-43.6	-120.7	-127.2	-122.0	3.71	4.58	5.01
w/o FlexBox	.784	.559	.373	43.1	26.4	20.6	4.68	9.39	11.53	-44.2	-38.7	-37.0	-124.5	-125.3	-92.0	3.14	4.56	5.01
w/o evolution	.697	.435	.324	39.7	30.3	18.1	5.33	8.31	13.14	-43.5	-39.5	-33.2	-121.3	-121.4	-104.7	3.98	4.08	5.13
FlexEvo	**.845**	**.838**	**.805**	**44.2**	**42.8**	**40.5**	**4.24**	**4.60**	**5.38**	-46.8	**-48.4**	**-44.2**	**-124.6**	**-128.1**	**-131.3**	**3.07**	**3.30**	**3.09**

Table 12: Principal objects used in the implementation and analysis.

Symbol	Meaning
P^0, R	Observed target coordinates and number of represented target residues.
G, $\mathcal{R}_G$	Spatial-prior genome and its residue support.
c_r, Q_r, z_r	Fixed residue centroid, fixed local frame, and evolvable box scale.
x, $n(x)$, $\mathcal{H}_x$	Candidate record, represented coordinate count, and edit history.
$\mathcal{X}_0$, $\mathcal{X}_G$	Shared screened source pool and pool adapted under G.
$\widetilde{P}_G^m$, K	Target probe m and total probe count; $K = 7$ includes Reference.
f_{pres}, f_{rob}	Candidate-level preservation and robustness objectives.
F_{pres}, F_{rob}	Corresponding genome-level pool summaries.
$\mathcal{H}^g$	Outer genome population at generation g.

Contact and clash. For candidate coordinate $p_i(x)$, let $d_{im} = \min_a \|p_i(x) - y_a^{(m)}\|_2$. For $n = n(x) > 0$, the contact and clash descriptors are

$$q_m = \text{clip}\left(\frac{1}{n}\sum_i \left[\mathbf{1}_{\{2.0\le d_{im}\le 4.5\}} + 0.35\mathbf{1}_{\{1.2\le d_{im}<2.0\}} - 1.5\mathbf{1}_{\{d_{im}<1.15\}}\right], 0, 1\right),$$
$$\chi_m = \frac{1}{n}\sum_i \left[\mathbf{1}_{\{d_{im}<1.15\}} + 0.35\mathbf{1}_{\{1.15\le d_{im}<1.45\}}\right]. \tag{17}$$

Algorithm 1 FLEXEVO core search

Require: Target P^0, source candidates, population N_G, generations $g_{\max}$, offspring per parent n_o
Compute fixed target descriptors, centroids, and frames
Construct seed genome G_0 and screen candidates to obtain $\mathcal{X}_0$
Initialize $\mathcal{H}^0$ with G_0 and repaired randomized variants
for each $G \in \mathcal{H}^0$ **do**
    Evaluate G from $\mathcal{X}_0$; store $\mathcal{X}_G$, both fitnesses, and feedback
end for
for $g = 0, \ldots, g_{\max} - 1$ **do**
    $\mathcal{Q}^g \leftarrow \emptyset$
    for each primary parent $G_a \in \mathcal{H}^g$ **do**
        for $v = 1, \ldots, n_o$ **do**
            Sample one of Pruning, Fusion, or Perturbation to obtain $\widetilde{G}$
            Apply G_a's feedback to $\widetilde{G}$ and repair the result to obtain G^+
            Decode G^+; reset the inner seed and adapt the shared $\mathcal{X}_0$
            Store $\mathcal{X}_{G^+}$, both genome fitnesses, and G^+'s own feedback
            $\mathcal{Q}^g \leftarrow \mathcal{Q}^g \cup \{G^+\}$
        end for
    end for
    $\mathcal{H}^{g+1} \leftarrow \mathcal{S}_G(\mathcal{H}^g \cup \mathcal{Q}^g; N_G)$
end for
return $\{(G, \mathcal{X}_G) : G \in \mathcal{H}^{g_{\max}}\}$

Finite-probe profile. Let $K = 7$, with probe zero equal to Reference, and define

$$\begin{aligned} \bar{q} &= K^{-1} \sum_{m=0}^{K-1} q_m, \qquad q_{\mathrm{rob}} = \rho \min_m q_m + (1-\rho)\bar{q}, \\ \Delta_q &= \max_m q_m - \min_m q_m, \qquad \chi_{\max} = \max_m \chi_m, \\ \kappa_q &= \begin{cases} K^{-1} \sum_{m=0}^{K-1} \mathbf{1}_{\{q_m \geq \tau q_0\}}, & q_0 > \varepsilon_q, \\ 0, & q_0 \leq \varepsilon_q. \end{cases} \end{aligned} \tag{18}$$

The reference configuration uses $\rho = 0.60$, $\tau = 0.75$, and $\varepsilon_q = 10^{-9}$. These quantities summarize aggregate contact quality across the evaluated target probes.

Anchor proximity and internal geometry. Let $\mathcal{S}_G$ and $\mathcal{F}_G$ denote the stable and flexible residue sets, respectively, and define $d_r(x) = \min_i \|p_i(x) - c_r\|_2$. For nonempty $\mathcal{S}_G$,

$$a(x; G) = \frac{1}{|\mathcal{S}_G|} \sum_{r \in \mathcal{S}_G} \exp\left[-\frac{(d_r(x) - 3.1)^2}{4.5} \right]; \tag{19}$$

otherwise $a = 0$. For finite candidates with $n \geq 2$, let $d_i^{\mathrm{self}} = \min_{j \neq i} \|p_i - p_j\|_2$. The internal-geometry and neighbor-support descriptors are

$$g(x) = \mathrm{clip}\left(1 - 0.16 \sum_i \mathbf{1}_{\{d_i^{\mathrm{self}} < 0.55\}} - 0.03[n - 12]_+, 0, 1 \right), \tag{20}$$

$$c(x) = \frac{1}{n} \sum_i \mathbf{1}_{\{\exists j \neq i:\ 0.65 \leq \|p_i - p_j\|_2 \leq 2.70\}}. \tag{21}$$

The core sets $g = c = 1$ for a single coordinate and $g = c = 0$ for an empty candidate. The size penalty is defined on the number of represented coordinates.

Spatial penalties and edit cost. Write $t_{ir} = \|Q_r^\top(p_i - c_r)\|_\infty/(b_{\ell_r} z_r)$ and $w_r^{\mathrm{box}} = \mathrm{clip}(z_r, 0.1, 1)$. The flexible-region exposure and forbidden occupancy are

$$r_{\mathrm{flex}}(x;G) = \frac{0.85\sum_{r\in\mathcal{F}_G}[1-d_r(x)/3]_+ + 0.05\sum_{r\in\mathcal{F}_G} w_r^{\mathrm{box}}\sum_i \mathbf{1}_{\{t_{ir}\le 0.55\}}}{\max(1,|\mathcal{F}_G|)},$$
$$\pi_{\mathrm{forb}}(x;G) = \frac{1}{n}\sum_i \sum_{r:\ell_r=\mathrm{forbidden}} \left[\mathbf{1}_{\{t_{ir}\le 0.5\}} + 0.2\mathbf{1}_{\{0.5<t_{ir}<0.7\}}\right]. \tag{22}$$

The edit cost is $e(x) = \sum_{o\in\mathcal{H}_x} w_o$, where $\mathcal{H}_x$ is the recorded edit history. Replacement, growth, removal, and candidate-fusion records cost 0.06; torsional, linker, pose-refinement, and side-chain-torsion records cost 0.025; outward-motion records cost 0.015; other records cost zero. Thus e measures the accumulated operation-weighted edit-history cost. The combined penalty is

$$\Pi = 0.95 r_{\mathrm{flex}} + 1.25\pi_{\mathrm{forb}} + 0.70\chi_{\max} + 0.40e. \tag{23}$$

The remaining objective coefficients are

$$(\alpha_a, \alpha_0, \alpha_r, \alpha_\chi) = (1.50, 1.00, 0.80, 1.40),$$
$$(\beta_g, \beta_c, \beta_k, \beta_d) = (0.90, 0.80, 0.75, 0.75). \tag{24}$$

Candidate objectives and genome aggregation. Both candidate objectives are maximized:

$$f_{\mathrm{pres}}(x;G) = 1.50a + q_0 + 0.80q_{\mathrm{rob}} - 1.40\chi_{\max},$$
$$f_{\mathrm{rob}}(x;G) = 0.90g + 0.80c + 0.75\kappa_q - 0.75\Delta_q - \Pi. \tag{25}$$

For a nonempty adapted pool, the genome objectives are

$$F_j = \lambda \max_{x\in\mathcal{X}_G} f_j(x;G) + \frac{1-\lambda}{|\mathcal{X}_G|}\sum_{x\in\mathcal{X}_G} f_j(x;G) - \eta_j\frac{|\mathcal{R}_G|}{R},$$
$$j\in\{\mathrm{pres},\mathrm{rob}\}, \qquad \eta_{\mathrm{pres}} = 0, \quad \eta_{\mathrm{rob}} = \eta. \tag{26}$$

Genome aggregation uses $\lambda = 0.70$ and $\eta = 0.10$. The two coordinate-wise maxima can be attained by different candidates. The coverage penalty applies only to the robustness objective. The core returns $(0, 0)$ if no boxes or no adapted candidates are available. No external docking or affinity estimator is queried by these scores.

Proxy ranking for the no-edit baseline. Let G_0 denote the initial genome after the deterministic repair used before the first genome evaluation of FLEXEVO. Rerank-only holds G_0 fixed and constructs its seven Reference-derived probes according to Eq. 16. Every candidate in $\mathcal{X}_0$ is scored using the two objectives in Eq. 25, with the original coefficients and penalty definitions. Its adaptation edit history is empty, giving $e(x) = 0$. Define

$$J_0(x) = f_{\mathrm{pres}}(x;G_0) + f_{\mathrm{rob}}(x;G_0).$$

The representative is

$$x_{\mathrm{rerank}} = \arg\max_{x\in\mathcal{X}_0} J_0(x).$$

Ties are resolved by decreasing $f_{\mathrm{pres}}(x;G_0)$ and then ascending candidate identifier, matching the final candidate-ranking rule of FLEXEVO. This baseline evaluates the entire screened pool under one fixed genome, without invoking the inner editor or outer genome evolution. Selection uses the geometric proxy scores; Vina and DockQ remain external evaluation endpoints.

A.5 GENOME VARIATION, FEEDBACK, AND SURVIVAL

Variation and scale storage. Pruning selects $\max\{1, \mathrm{round}(0.15|\mathcal{E}|)\}$ entries from the stable/flexible set $\mathcal{E}$, prioritizing low usage when available and sampling uniformly otherwise; it does nothing when $\mathcal{E}$ is empty. Fusion samples a target residue and transfers the donor configuration within an 8 Å neighborhood, copying all scale entries there, including those of residues outside the donor support. The donor is sampled from the half of alternative parents most distant in residue-role space.

Perturbation samples $\max\{1, \mathrm{round}(0.18|\mathcal{N}|)\}$ residues without replacement from a 7 Å neighborhood. For a selected covered residue, it chooses uniformly between removing its box and changing to either other role; for an uncovered residue, it adds one of the three roles uniformly. A separate subset of $\max\{1, \mathrm{round}(0.25R)\}$ residues is sampled from all R residues for scale mutation, with $\sigma_z = 0.12$. Removing a box leaves its stored scale unchanged.

Feedback and repair. Feedback uses the primary parent's surviving edit histories and adapted coordinates, including for fusion offspring. Let u_r count uses of residue r in the surviving edit histories. For a nonempty parent-adapted pool, the signal combines usage and proximity:

$$D_r = \min_{x \in \mathcal{X}_{G_a}, i} \|p_i(x) - c_r\|_2,$$
$$s_r = \max\left\{ \frac{u_r}{\max(1, \sum_k u_k)}, \left[1 - \frac{D_r}{d_c}\right]_+ \right\}, \qquad d_c = 4.5. \tag{27}$$

For a nonempty parent pool, the minimum is taken over its represented coordinates. For an empty pool, the reference implementation initializes nearest-coordinate distances to $+\infty$ and edit-use counts to zero. The resulting proximity contribution and feedback signal are zero. Uncovered residues with $s_r \geq 0.35$ are activated as flexible boxes, retaining their proposal scales. Let $\mathcal{E}^+$ be the stable/flexible set after activation. Among entries with $s_r \leq 0.02$, retirement selects the lowest-signal entries up to the budget

$$B_- = \max\left\{0, \min\left(\left\lfloor 0.20|\mathcal{E}^+| \right\rfloor, |\mathcal{E}^+| - \lceil 0.25R \rceil\right)\right\}. \tag{28}$$

The $0.25R$ term limits this retirement step; it does not impose a global minimum support on an incoming genome. Pruning and feedback retirement exclude forbidden assignments, although fusion and role mutation may change them. Repair clips scales and attempts to restore missing stable and flexible roles using density and radial descriptors when eligible residues exist.

Diversity-aware selection. For two genomes, the residue-role distance is

$$D_G(G_a, G_b) = |\mathcal{R}_{G_a} \triangle \mathcal{R}_{G_b}| + \sum_{r \in \mathcal{R}_{G_a} \cap \mathcal{R}_{G_b}} \mathbf{1}_{\{\ell_r^a \neq \ell_r^b\}}, \tag{29}$$

where $\triangle$ is symmetric difference; scales do not enter this measure. Both selectors visit nondominated fronts in increasing rank and order each front by decreasing crowding, followed by objective sum and identifier. Outer selection defers a genome within distance 2 of any selected genome; inner selection uses a candidate-centroid separation of 1 Å. After front traversal, deferred entries fill remaining slots in traversal order. Consequently, diversity can admit a lower-front entry before a deferred higher-front entry, and backfilling does not enforce a hard pairwise-distance bound.

A.6 Reference Configuration and Execution Schedule

Table 13 gives the default core configuration. Algorithm 1 keeps the inner adaptation and outer genome updates separate; each offspring evaluation also produces its own feedback statistics for the next generation.

Execution order and budget. One complete run computes the fixed target descriptors, constructs the initial prior, screens the source pool, evaluates the initial genomes, and then alternates offspring generation, feedback, fresh inner adaptation, and survivor selection. With the defaults in Table 13, the scheduled genome-evaluation count is $8 + 4 \times 8 \times 2 = 72$. This is a search-budget count, not a sample size or replication count. The returned records should retain the selected genomes, adapted coordinates, edit histories, and both candidate and genome scores so that the decisions can be reconstructed.

Table 13: Default configuration of the coordinate-level FlexEvo core.

Component	Parameter	Default
Randomization	Seed	31
Screening	Pool cap / neighbor threshold	$64/0.35$
Screening	Molecular / geometric similarity	$0.92/0.90$
FlexBox	Scale bounds	$[0.55, 1.75]$
FlexBox	Stable / flexible / forbidden extent	$3.0/4.5/2.9$ Å
Inner adaptation	Rounds / maximum population	$2/12$
Inner adaptation	Reproduction cap / proposals per parent	$6/2$
Inner selection	Centroid niche radius	1.0 Å
Outer evolution	Generations / population	$4/8$
Outer evolution	Offspring per parent	2
Variation	Fusion / pruning / perturbation	$0.35/0.20/0.45$
Variation	Fusion / mutation radius	$8.0/7.0$ Å
Variation	Relabel / resize fraction	$0.18/0.25$
Variation	Scale noise / pruning fraction	$0.12/0.15$
Feedback	Proximity radius	4.5 Å
Feedback	Activation / retirement threshold	$0.35/0.02$
Feedback	Retirement cap / retention-floor fraction	$0.20/0.25$
Genome fitness	Maximum weight / coverage penalty	$0.70/0.10$
Genome selection	Role-distance threshold	2

B Experimental Protocol and Reproducibility

This section specifies candidate selection, evaluation, and reporting for the single-conformation adaptation protocol. The coordinate-level search is defined in Appendix A, with its reference settings in Table 13. Implementation details below describe the reference protocol.

B.1 Evaluation design

Benchmark sample size. Distinct protein targets define the benchmark sample size. The nine binder categories have the following within-category target counts.

Binder category	Targets
Conventional small molecules	9,393
Nonpeptidic macrocycles	10,322
Linear peptides	4,694
Cyclic peptides	5,164
Non-antibody proteins	1,260
Full-length antibodies	1,386
Antibody fragments and nanobodies	1,525
Oligonucleotides	1,677
mRNA and other long nucleic acids	1,878
Total target–category memberships	37,299

Each target contributes once within each category in which it is evaluated. The total therefore comprises 37,299 target–category memberships. The globally distinct target set is the union of the nine category-specific target sets. All observations from the same target retain their shared target identity throughout aggregation.

The reference protocol specifies three natural conformations per target: one Reference, one Matched, and one Divergent. Reference supplies the structure for source generation and adaptation; Matched and Divergent are held out for evaluation. A complete eligible pair contains base and adapted observations for the same case, source model, seed, state, and metric. Paired observations are aggregated within targets, and targets receive equal weight under Appendix B.3.

Scope and candidate sources. The benchmark covers nine binder categories: conventional small molecules, nonpeptidic macrocycles, linear peptides, cyclic peptides, non-antibody proteins, full-length antibodies, antibody fragments and nanobodies, oligonucleotides, and mRNA or other long nucleic acids. The reference protocol assigns DynamicFlow, FlexSBDD, and YuelDesign as the ligand-generation candidate sources. Linear peptides and cyclic peptides are evaluated with BoltzGen, Chai-1, and Protenix; the other five biomolecular categories use the same three sources. For every configured pair of source and category, the protocol compares the original output with the corresponding output after FLEXEVO adaptation and fixes the source-model parameters.

Baseline execution and source pools. The reference protocol requests up to 64 candidates per case, source model, and seed using Reference alone. The source adapter retains the native output order, truncates excess outputs, and records shortfalls without duplicating candidates to fill the pool. Imported predictions require provenance for the source-model revision, checkpoint checksum, inference settings, candidate coordinates, and Reference-frame alignment. The task mode distinguishes de novo ligand generation from binder design or fixed-sequence conditioned structure prediction for other modalities; the executed mode must accompany each source output.

Final representative selection. Across all 27 source–category combinations, FLEXEVO improved the primary evaluation metric under the held-out Divergent conformation. These gains were obtained using only the Reference conformation during adaptation, with source-model parameters fixed and each selected candidate evaluated without state-specific redesign. Together, these results demonstrate the cross-conformation benefits of the complete FLEXEVO pipeline across nine binder categories. The base representative is the first candidate in the validated source-native order. Adaptation first screens the source pool in input order, rejecting empty or non-finite coordinates, insufficient neighbor support, and forbidden-region overlap. Screening treats a candidate as redundant when its geometric similarity to an accepted candidate is at least 0.90 and, when both fingerprints are available, its fingerprint similarity is at least 0.92. With missing fingerprints, only the geometric condition is used. At most 64 accepted candidates form the shared adaptation pool. This additional geometric screening belongs to adaptation; the base representative is selected from the validated source pool before that screen.

After outer-loop survival, the reference implementation chooses the genome with the largest sum of its two proxy fitness values. Ties are resolved by the preservation objective and then the genome identifier. Within that genome's candidate pool, candidates are ranked by the sum of their two candidate objectives, with preservation and candidate identifier breaking ties. The first candidate supplies the adapted representative. This rule selects one winning genome's pool; it does not merge all survivor pools. Both ranking stages use Reference-derived proxy objectives only.

Pipeline and source-selection comparisons. The nine-category benchmark evaluates the complete FLEXEVO workflow against the source-native representative. This comparison includes screening, adaptive search, and final proxy-based selection. The supplementary three-arm comparisons report Vina scores averaged across DynamicFlow, FlexSBDD, and YuelDesign for conventional small-molecule drugs (Table 14), and DockQ scores averaged across BoltzGen, Chai-1, and Protenix (Table 15). Rerank-only and FLEXEVO receive the same screened pool $\mathcal{X}_0$, containing up to 64 source candidates. Rerank-only scores every member of this pool under the fixed initial genome G_0; FLEXEVO applies its full adaptation search to the shared input pool. Both return one representative selected using Reference-derived information. The Rerank-only comparison assesses the additional benefit of the complete adaptive-search workflow over fixed-G_0 source selection.

Source + Proxy Rerank (No Edit). The comparison comprises three arms: Base, Rerank-only, and FLEXEVO. All three reuse the same source-model generation run. Base retains the first validated candidate in source-native order. Rerank-only selects one unedited candidate from the same screened source pool $\mathcal{X}_0$ supplied to FLEXEVO, using the proxy-ranking rule specified in Appendix A.4. The pool contains up to 64 accepted source candidates, with their original coordinates retained. FLEXEVO returns the representative from its full adaptation search.

All representative selection uses Reference-derived information. Each representative is frozen before evaluation on Reference, Matched, and Divergent with the same alignment and evaluator settings. Base versus Rerank-only measures the contribution of source-pool screening and proxy rank-

ing. Rerank-only versus FLEXEVO measures the additional benefit of adaptive search over the source pool. Under the reference protocol, each combination of case, model, seed, and variant contributes one representative to pose and interaction evaluation. Set-level diversity uses the corresponding frozen pool. The same representative is evaluated across all three natural states. The two base candidates and three adapted candidates specified for the cross-conformation display are the first available members of their respective frozen orderings, without padding. These displayed subsets are distinct from the evaluation unit that uses one representative per combination.

Natural target-state conditions and alignment. Reference denotes the single structure supplied to the source workflow and to FLEXEVO. Matched and Divergent denote alternative natural structures of the same target and binding site. Matched preserves the regulatory state assigned to Reference. Divergent contains a propeptide-associated change in that state and localized structural changes. These groups follow biological annotation rather than a post hoc RMSD threshold. Matched and Divergent structures are held out until evaluation; their scores do not enter candidate screening, editing, search fitness, or representative selection.

The protocol specifies a declared one-to-one correspondence between receptor Cα atoms, with at least 90% mapping coverage. The alignment routine uses the supplied atom pairs and requires at least three non-collinear pairs. A proper Kabsch rotation and translation place each alternative receptor in the Reference coordinate frame. Candidate coordinates remain in that frame throughout. Thus the receptor is transformed around a fixed candidate; no redocking, local minimization, or state-specific adaptation is performed. The Reference state uses the identity transform. Coordinate and full-record checksums are recorded before evaluation and checked after each state. Alignment receipts retain the receptor identifiers, correspondence, rotation, and translation.

Synthetic probes and fixed candidates. The held-out natural states are distinct from the synthetic probes used by the search. For each evaluated genome, FLEXEVO constructs six local perturbations from Reference and scores them together with the unperturbed Reference target. Candidate coordinates remain fixed across all seven probes. Probe construction and geometric objectives are defined in Appendix A.4. These synthetic probes provide the geometric robustness signal used during search. Natural-state evaluation measures performance on the held-out conformations.

Matched comparisons and secondary analyses. Within a comparison of one source and one category, base and adapted outputs use the same target-state definitions and the same evaluator. Percentage differences are percentage-point changes when the underlying quantity is a percentage. The three natural states are evaluations of one-time adaptation, rather than independent adaptation tasks. Statistical units and missing-value handling are specified in Appendix B.3.

The geometry–mobility analysis specifies 200 ATLAS proteins. It averages residue RMSF across three genuine molecular-dynamics trajectories, and it normalizes each quartile mean by the corresponding whole-chain mean RMSF. The lowest and highest within-protein quartiles of packing sparsity, and the lowest and highest within-protein quartiles of radial position, are compared using proteins as the analysis units. This analysis assesses the rationale for geometry-guided initialization. FlexBox roles specify geometric editing constraints. PDB and UniProt entry and target counts characterize the composition and annotation coverage of the wider structure pool.

The reference-similarity analysis specifies 10 generated samples and 47 randomly sampled public references in each scenario. It evaluates all 470 pairs of a generated sample and a reference, and it reports the maximum and the median similarity. Morgan fingerprints of radius 2 and length 1024 bits, compared by Tanimoto similarity, are specified only for molecular modalities. Polymer scenarios require a separately labelled sequence-based similarity. These comparisons quantify resemblance under the chosen representation within the sampled reference collection. The reporting protocol associates each value with its reference collection, sample identifiers, sampling seed, and modality-specific similarity definition.

B.2 EVALUATION METRICS

Structural export and evaluator inputs. The object passed to an evaluator is the frozen molecular candidate defined in Appendix A.1, with the structural safeguards in Appendix A.3. The reference pipeline writes a complex mmCIF using the candidate's atom, element, chain, residue, and

coordinate records. It requires one atom record per candidate coordinate and rejects candidates marked `coordinate_only`. The accompanying molecular record retains the supplied connectivity and other chemical metadata independently of the coordinate-file representation. Writing atom and residue coordinates to mmCIF does not reconstruct unsupplied covalent bonds or polymer sequence annotations.

For a candidate with an SDF-derived molecular block, the exporter reconstructs that molecular object and writes the adapted coordinates into its existing atom order. It checks that the molecular and coordinate atom counts agree. The resulting SDF supplies the ligand object for PDBQT preparation and fixed-pose Vina-family scoring. For native-complex comparisons, DockQ reads the exported candidate complex together with the declared native structure and chain/interface mapping. For MM/GBSA and other preparation-dependent evaluators, the reporting specification requires prepared topology, coordinate inputs, and preparation records alongside the score. Their candidate identity must refer to the same frozen molecular record. The reference package integrates MM/GBSA through imported external score records linked to their candidate and preparation provenance.

Coordinate and complete-record checksums identify the representative before evaluation across the three receptor states. Exported structures and prepared inputs have separate checksums to preserve the preparation history. Preparation that changes heavy-atom geometry defines a separate refinement result, rather than the original fixed-pose measurement. Molecular-validity outcomes, evaluator success, and numerical scores occupy separate reporting fields. The checks in Appendix A.3 assess molecular validity, while Vina, DockQ, and MM/GBSA quantify their respective scoring endpoints.

Fixed-pose ligand scoring. The reference scoring implementation evaluates Vina, Vinardo, and AutoDock4 in `score_only` mode, taking the total score returned for the frozen pose. It does not invoke docking or optimization. Meeko prepares PDBQT inputs; ligand hydrogens may be added without optimizing heavy-atom coordinates. The paired native ligand is placed in the Reference frame using the receptor transform. Candidate and native ligand share the same receptor preparation and the same grid. The grid center is the midpoint of their combined coordinate extrema. Each side is the larger of two lengths: the corresponding coordinate span plus 12 Å, and 20 Å. AutoDock4 requires AutoGrid4 maps. Preparation commands, input checksums, package versions, grid parameters, CPU count, and scoring seed are retained with each score.

Vina is the principal ligand-generation score, with lower values preferred. The reference implementation defines its high-affinity indicator as a candidate Vina score strictly below the paired native ligand score under the same scoring conditions. This high-affinity indicator is a comparative docking-score criterion defined against the paired native ligand under the same computational conditions. Vinardo, AutoDock4, and RTMScore provide complementary views. Raw Vinardo and AutoDock4 energies are ordinarily minimized. External scores are separate from the geometric search objectives in Eqs. equation 25 and equation 26.

Molecular and set-level properties. QED and SA summarize drug-likeness and synthetic-accessibility heuristics; Lipinski descriptors, logP, and PAINS screening characterize complementary molecular properties. With the molecular graph and descriptor settings fixed, each candidate retains the same intrinsic descriptor values across receptor conformations. Tables 2 and 3 present these summaries under Reference. Diversity characterizes the frozen candidate pool, separately from representative-pose metrics. With candidate-pool membership, coordinates, and the diversity definition fixed, changing the receptor state does not change this pool-level summary. Across the nine category tables, diversity is reported once under Reference; dashes in Matched and Divergent indicate that the pool-level summary is not repeated. LogP is reported descriptively, with black values and a neutral column heading.

Native-complex comparison. DockQ is favorable when higher; LRMSD is favorable when lower, with LRMSD below 2 Å defining pose success. The reference evaluator requires a matching native binder identity, an explicit native-to-model chain map, and a declared list of native interfaces between target and binder. It evaluates the mapped interfaces with DockQ and takes the arithmetic mean of their DockQ and LRMSD values when several interfaces are requested. The native complex, mapping, interface list, and evaluator version are retained in the score record. MM/GBSA and HADDOCK are complementary scoring terms, with lower values preferred in the supplied tables. RTMScore, MM/GBSA, HADDOCK, and other external measures must use validated, versioned

score records that declare the modality to which they apply. The scoring record specifies input preparation, evaluator version and settings, and the identities of the evaluated structures.

Confidence-derived interaction metrics. For a declared target-token set T and binder-token set B, the reference pAE aggregation is the arithmetic mean of the two directional interchain means, $\frac{1}{2}(\overline{\mathrm{PAE}_{T,B}} + \overline{\mathrm{PAE}_{B,T}})$. The sets must be nonempty and disjoint. The confidence array must be tied to the evaluated candidate, predictor version, and input checksum. Confidence from an unmodified source output cannot be substituted for that of an edited candidate. The off-target summary is the minimum of this symmetric mean across the complete declared off-target panel. Lower on-target pAE and higher off-target pAE are the reported preferences; these quantities are not affinity measurements. The exact predictor and token subsets remain part of each run's provenance.

Primary endpoints for the three-arm comparison. Base, Rerank-only, and FLEXEVO are compared using the primary endpoints of Table 1: Vina in kcal/mol for small molecules and nonpeptidic macrocycles, and DockQ for the remaining seven binder categories. Comparisons are resolved by source model, binder category, and natural target state. Raw scores retain their original units and directions: lower Vina and higher DockQ indicate better performance. For effect calculations, $Y = -\mathrm{Vina}$ for the two ligand categories and $Y = \mathrm{DockQ}$ for the other categories, so positive differences consistently indicate improvement.

Supplementary comparison and display conventions. Tables 2, 3, 4, 5, 6, 7, 8, 9, and 10 retain modality-specific metrics. Under the reference reporting protocol, main-table values and changes are rounded independently, with changes calculated before rounding. The reference cross-conformation plotting protocol defines the molecular margin as native Vina minus candidate Vina and the biomolecular margin as $\mathrm{DockQ} - 0.49$. Centroids summarize the candidate subsets displayed in each configuration. The reference radar protocol shares bounds across displayed methods and states within a binder category. Metrics shown with a downward arrow are reverse-scaled. Each category has its own normalization bounds, and individual radial axes provide the metric-specific comparisons. On-target and off-target diagnostic regions are defined by an on-target score below 5 and a minimum off-target score above 10. These thresholds are specific to the plot.

Reference-structure illustrations describe molecular architectures. The optimization trajectories in Appendix C.10 report recorded metric evolution across binder categories during FlexEvo adaptation. Figure 4 reports measured adaptation dynamics, and Appendix C.5 reports computational costs.

B.3 REPRODUCIBILITY

Core search budget. The manuscript reference configuration uses seed 31, a screened-pool cap of 64, two inner adaptation rounds, an outer population of eight, four generations, and two offspring per parent. Other parameters are listed in Table 13. Each genome evaluation creates a fresh editor, resets the inner random seed, and starts from the same screened pool $\mathcal{X}_0$. Parent-adapted coordinates contribute feedback about where to search but are not inherited as the offspring's initial candidates in the full FlexBox-based method.

Eight initial evaluations plus $4 \times 8 \times 2$ offspring evaluations give 72 scheduled genome evaluations under the complete schedule. This counts genome evaluations, rather than distinct genomes, candidate-objective calls, or independent experimental replicates. Resetting the inner seed controls comparisons among genomes; it does not estimate run-to-run variability. Source-generation, adaptation, and evaluator randomness are recorded separately.

Independent repeats and statistical units. Each experiment is repeated three times using random seeds 31, 47, and 71. Within a category, model, state, and seed, the reporting procedure first averages observations within each case and then averages cases within each target. Targets receive equal weight; means and medians summarize these target-level values. Candidate pools, conformational states, and repeated score calls are not counted as independent targets.

Paired contrasts match the same case, model, seed, and state across base and adapted variants. Eligible complete pairs are averaged within targets before computing the equal-target-weight contrast. The family-bootstrap interval procedure resamples target families with replacement 10,000 times and uses the 2.5th and 97.5th percentiles for a 95% interval. Each draw retains all targets and paired

cases within a sampled family, with equal target weights. Fewer than two eligible families yields no interval.

Model-averaged three-arm comparisons. Tables 14 and 15 report model-averaged scores for each arm and natural target state. The Vina comparison averages DynamicFlow, FlexSBDD, and YuelDesign; the DockQ comparison averages BoltzGen, Chai-1, and Protenix. Score differences are computed within each table and state. Let $\bar{E}_s^R$ and $\bar{E}_s^F$ denote the reported Vina means for Rerank-only and FLEXEVO, and let $\bar{Q}_s^R$ and $\bar{Q}_s^F$ denote their DockQ means. Improvements over Rerank-only are $\bar{E}_s^R - \bar{E}_s^F$ for Vina and $\bar{Q}_s^F - \bar{Q}_s^R$ for DockQ, so positive differences favor FLEXEVO in both comparisons.

Paired effects of reranking and adaptive search. Three-arm contrasts match the same case, source model, seed, and natural target state. All three-arm summaries are restricted to the common set of complete observations across Base, Rerank-only, and FLEXEVO. Repeated observations are averaged within cases and cases within targets, with equal target weighting. For a fixed source–category combination and state s, let $Y_{t,s}^B$, $Y_{t,s}^R$, and $Y_{t,s}^F$ denote target-level scores for Base, Rerank-only, and FLEXEVO, respectively. On their common target set $\mathcal{T}_s$, define

$$\Delta_{\text{select},s} = \frac{1}{|\mathcal{T}_s|} \sum_{t \in \mathcal{T}_s} \left(Y_{t,s}^R - Y_{t,s}^B \right),$$

$$\Delta_{\text{adapt},s} = \frac{1}{|\mathcal{T}_s|} \sum_{t \in \mathcal{T}_s} \left(Y_{t,s}^F - Y_{t,s}^R \right),$$

$$\Delta_{\text{total},s} = \Delta_{\text{select},s} + \Delta_{\text{adapt},s}.$$

Here $\Delta_{\text{adapt},s}$ is the incremental effect of the full adaptive search relative to source-pool reranking. All three contrasts use the same 10,000 target-family bootstrap draws, retaining the paired records of every sampled family. The 2.5th and 97.5th percentiles define pointwise 95% intervals for each contrast.

Failure handling, coverage, and denominators. Coverage records retain every requested case. The reporting implementation distinguishes measured values, executed failures, unsupported metrics, and unrun requests. Finite measured continuous scores enter their estimates; failed continuous scores remain missing. An executed failure contributes zero to applicable success indicators, including high-affinity fraction, LRMSD success, validity, and PAINS pass. Unsupported and unrun requests remain missing rather than becoming failure zeros. Consequently, requested coverage and eligible metric denominators are reported separately.

For each comparison defined by category, model, state, seed, and metric, the coverage record contains the number of requested cases, the number of eligible base cases, the number of eligible adapted cases, the number of complete pairs, the number of unpaired cases, the number of targets, the number of families, and the status counts. Continuous paired contrasts use only complete eligible pairs. Success-rate denominators retain executed failures, and the case-to-target aggregation above determines the weighting. Empty source pools, screening rejection, empty adapted pools, unavailable native references, and evaluator failures retain distinct reasons.

Optimizer and ablation budgets. The reference optimizer comparison shares the screened pool, initial prior, proxy objectives, and seed, resetting the random stream for each method. Its matched budget is 72 genome evaluations. Candidate-objective evaluation counts and wall-clock time are recorded separately because equal genome counts need not imply equal candidate-level work or runtime. Shared setup is timed separately from each optimizer's search.

Reranking budget and provenance. For a completed Rerank-only run, each of the $|\mathcal{X}_0|$ source candidates receives one candidate-objective evaluation under G_0, comprising seven probe evaluations. The run therefore uses $|\mathcal{X}_0|$ candidate-objective calls and $7|\mathcal{X}_0|$ probe evaluations. The record retains the screened-pool size, source order and identifiers, G_0, score components, selected-candidate checksum, elapsed time, and execution status. Source-pool shortfalls retain their actual cardinality, and an empty screened pool is recorded as a failure.

Hyperparameter specification. Table 13 gives the shared reference defaults, while the run configuration records any override. The core defaults do not define separate values for individual binder categories. Source-model weights are fixed, and adaptation does not retrain them. The search itself accesses Reference only. The hyperparameter reporting protocol distinguishes fixed defaults from validation-selected overrides. For each tuned configuration, its record specifies the parameters varied, search ranges, selection criterion, validation targets, and category-specific choices. Dataset and split identifiers document the separation between parameter selection and held-out evaluation.

Run reconstruction. A complete run manifest must retain Reference coordinates, chain assignments, pocket center, ordered source records, fingerprints where available, source-model settings, evaluator settings, and the identifiers of the Matched and Divergent structures. Candidate order must be preserved because screening precedes pool truncation. Separate run artifacts must identify the original source pool, the screened pool $\mathcal{X}_0$, the genome pools, the winning genome, and the representative frozen before held-out evaluation. Checksums must connect candidates, alignment receipts, native references, score records, and table exports. The structural-provenance specification additionally requires each representative's source ancestor, executed edit mode, ordered atom identities, supplied covalent graph, polymer annotations, and edit history. Each validation record must identify the exported structure checksum, validator version, configuration, applicable tests, outcomes, and failure reasons. The records must link the candidate to its SDF or complex mmCIF export and to each evaluator's prepared inputs. The protocol requires the same declared validation policy for base and adapted candidates. It reports validation coverage and observed pass counts separately from successful score counts, using the denominators defined above. A pass is assigned only to a candidate with a recorded successful validation result.

Runtime reporting. The reference timing unit is one execution of a case, a model, and a seed, processing its source pool through adaptation to a representative, rather than one timing observation per candidate in that pool. The pipeline measures stage elapsed wall-clock time, including the overhead inside each stage, and separates source generation, adaptation, and external evaluation. The runtime summarizer reports source-generation plus adaptation time as the total; external evaluation is reported separately. Importing existing source predictions is not a measurement of native source-model inference time.

The auxiliary timing protocol configures one warm-up and ten measured executions with GPU synchronization, while the summarizer excludes warm-up and unsuccessful records and averages the successful observations. The timing record specifies hardware, software revision, parallelism, batch size, numerical precision, preprocessing boundaries, and the number of successful measured executions. Downloads and setup are recorded separately under the reference timing protocol.

C Supplementary Results and Analyses

This section presents the experimental results and accompanying analyses. The detailed tables report base and adapted results for each evaluated model, binder category, and target conformation. Maintext values are rounded from these same results, and changes are calculated before rounding.

C.1 Structure-pool coverage and biological context

Figures 5 and 6 characterize the structural and biological coverage of the source pool. The small-molecule subset contains 19,715 PDB entries, and the UniProt analysis covers 1,774 mapped targets. These counts describe the source pool at the structure-entry and annotated-target levels, respectively. The category-specific benchmark target counts appear in Appendix B.1; Appendix B.3 defines the metric-specific pairing, coverage denominators, and family-level resampling procedure.

C.2 Detailed results across nine binder categories

FlexEvo improves the Divergent-state primary endpoint across all 27 source–category combinations in Tables 2–10. The primary endpoint is Vina for conventional small molecules and non-peptidic macrocycles, and DockQ for the remaining seven categories. Each table reports Base and FlexEvo under Reference, Matched, and Divergent conditions. Additional metrics characterize

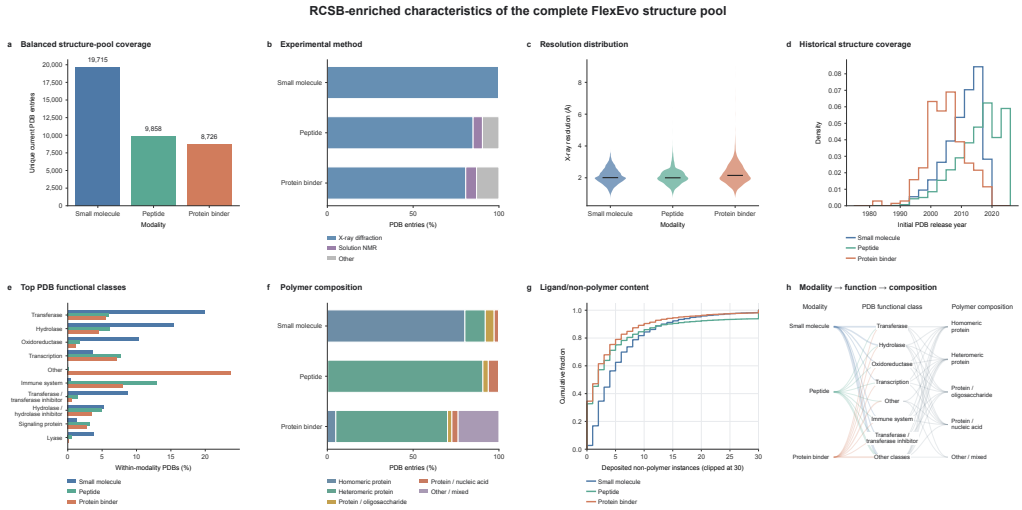

Figure 5: **Structural and functional characteristics of the FlexEvo structure pool.** RCSB Protein Data Bank (PDB) annotations characterize the small-molecule, peptide and protein-binder subsets. **a,** Numbers of unique current PDB entries in each subset: 19,715 for small molecules, 9,858 for peptides and 8,726 for protein binders. **b,** Proportions of structures determined by X-ray diffraction, solution nuclear magnetic resonance (NMR) and other experimental methods. **c,** Distributions of reported resolution for X-ray structures. **d,** Density distributions of initial PDB release years. **e,** Prevalence of the displayed PDB functional classes, expressed as percentages of entries within each modality. **f,** Polymer composition, categorized as homomeric protein, heteromeric protein, protein–oligosaccharide, protein–nucleic acid and other or mixed assemblies. **g,** Cumulative distributions of deposited non-polymer instance counts per entry, displayed up to 30 instances. **h,** Alluvial representation linking modality, PDB functional class and polymer composition. Blue, teal and orange distinguish the small-molecule, peptide and protein-binder subsets, respectively, in panels a, c–e and g.

pose agreement, molecular properties, model confidence, and diversity. Colors encode the direction of change under each metric's stated preference.

Model-averaged rerank-only controls. Tables 14 and 15 compare Base, Rerank-only, and FLEXEVO across Reference, Matched, and Divergent. The conventional small-molecule Vina results are averaged across DynamicFlow, FlexSBDD, and YuelDesign. The DockQ results are averaged across BoltzGen, Chai-1, and Protenix. Rerank-only follows the fixed-G_0, unchanged-source-pool protocol in Appendices B.1 and A.4. FLEXEVO achieves the best model-averaged score in all three states for both metric groups. For conventional small molecules, FLEXEVO improves the model-

Table 14: Three-state rerank-only comparison for conventional small-molecule drugs. Vina scores are reported on the common three-arm evaluation set and aggregated across DynamicFlow, FlexSBDD, and YuelDesign. Scores are in kcal/mol; lower values are better. Rerank-only selects an original source candidate without coordinate editing. Bold values indicate the best model-averaged score in each state.

Method	Reference ↓	Matched ↓	Divergent ↓
Base	−8.260	−7.440	−6.540
Rerank-only	−8.560	−7.370	−7.750
FLEXEVO	**−9.740**	**−8.678**	**−8.698**

averaged Vina score over Rerank-only by 1.180, 1.308, and 0.948 kcal/mol on Reference, Matched, and Divergent, respectively. In the DockQ comparison, the corresponding gains are 0.1909, 0.4401, and 0.4801, with the largest gain on Divergent. These comparisons demonstrate improved model-

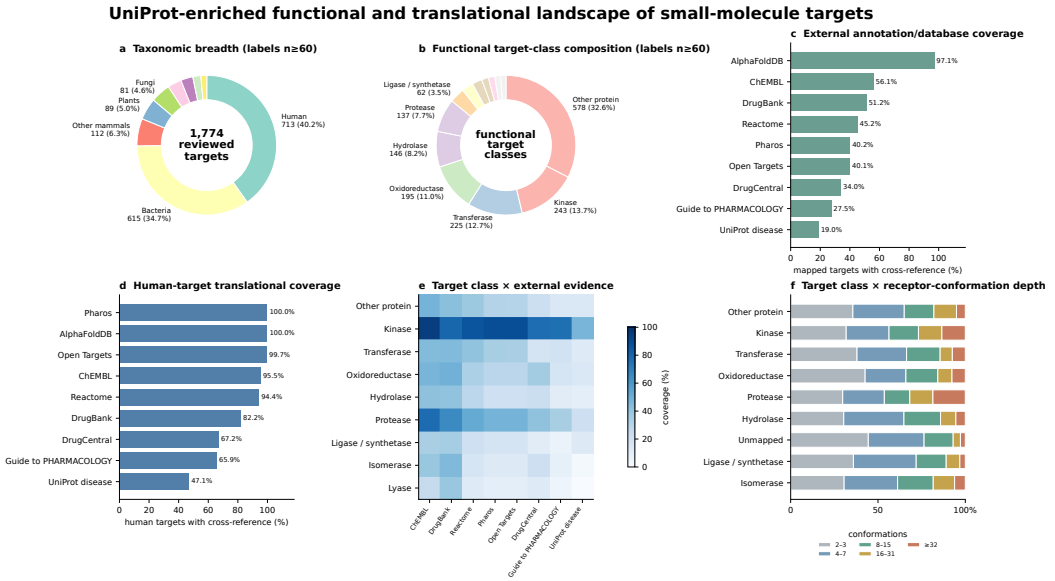

Figure 6: **Taxonomic, functional and translational annotation landscape of small-molecule targets in FlexEvo. a,** Taxonomic composition of 1,774 targets mapped to reviewed UniProt entries, including 713 human targets (40.2%) and 615 bacterial targets (34.7%). **b,** Functional target-class composition. In panels a and b, categories containing at least 60 targets are labelled with their counts and percentages. **c,** Coverage of external database cross-references and UniProt disease annotations among mapped targets, including AlphaFoldDB, ChEMBL, DrugBank, Reactome, Pharos, Open Targets, DrugCentral and Guide to PHARMACOLOGY. **d,** Corresponding annotation coverage within the human-target subset. **e,** Annotation coverage stratified by functional target class; darker blue indicates a higher percentage of targets with the corresponding cross-reference or annotation. **f,** Within-class distributions of receptor-conformation counts per target, grouped into 2–3, 4–7, 8–15, 16–31 and ≥ 32 conformations, with an additional unmapped category. Database coverage quantifies the availability of target annotations and external cross-references.

Table 15: Three-state rerank-only comparison using DockQ. DockQ scores are reported on the common three-arm evaluation set and aggregated across BoltzGen, Chai-1, and Protenix. Higher values are better. Rerank-only selects an original source candidate without coordinate editing. Bold values indicate the best model-averaged score in each state.

Method	Reference ↑	Matched ↑	Divergent ↑
Base	0.6270	0.3960	0.2760
Rerank-only	0.6630	0.4213	0.3372
FLEXEVO	**0.8539**	**0.8614**	**0.8173**

averaged performance beyond source-pool reranking in both metric groups across all three natural target states.

Category-balanced cross-conformation degradation. We quantified Reference-to-Divergent degradation using the primary metrics in Tables 2–10. For category c, source model m, and configuration $v \in \{\text{base}, \text{FlexEvo}\}$, let $S^{(v)}_{cm,R}$ and $S^{(v)}_{cm,D}$ denote the Reference and Divergent scores. We used negative mean Vina scores for small molecules and nonpeptidic macrocycles, and mean DockQ for the remaining seven categories. Thus, higher S indicates better performance within each

category. We computed the relative degradation and its category-balanced mean as

$$d_{cm}^{(v)} = 100\% \times \frac{S_{cm,R}^{(v)} - S_{cm,D}^{(v)}}{|S_{cm,R}^{(v)}|}, \tag{30}$$

$$\bar{d}^{(v)} = \frac{1}{9} \sum_{c=1}^{9} \frac{1}{|\mathcal{M}_c|} \sum_{m \in \mathcal{M}_c} d_{cm}^{(v)}, \tag{31}$$

where $\mathcal{M}_c$ contains all three source models reported for category c. This includes Chai-1 for both linear and cyclic peptides, giving 27 source–category combinations in total. Each configuration was normalized by its own Reference score, and all Reference denominators were nonzero. Negative degradation values indicate improved Divergent performance and were retained without clipping to zero. Ratios were calculated from the reported state-wise means before averaging within categories and then equally across the nine categories. This statistic summarizes category-balanced relative degradation using the reported state-wise means.

For example, DynamicFlow small molecules had mean Vina scores of -6.6062 and -4.8519 kcal/mol under Reference and Divergent, respectively, giving 26.5554% degradation. After FLEXEVO, the corresponding scores were -9.1775 and -7.7991 kcal/mol, giving 15.0193% degradation. Across all nine categories and 27 source–category combinations, the complete FLEXEVO selection-and-adaptation workflow reduced category-balanced mean relative degradation from 47.8376% to 4.3854%, a reduction of 43.4522 percentage points. Rounded to one decimal place, these values are 47.8% and 4.4%. This benchmark summary quantifies cross-conformation score retention for the complete workflow relative to the source-native representative. Tables 14 and 15 complement this benchmark summary with model-averaged three-arm comparisons, quantifying the additional Vina and DockQ improvements over source-pool reranking.

C.3 JOINT CROSS-CONFORMATION COMPARISONS

Figures 7–9 place two target conditions on separate axes. Their joint margins distinguish candidates meeting both plotted thresholds from those meeting only one. Centroid shifts describe the displayed subsets of two Base and three FLEXEVO candidates, respectively.

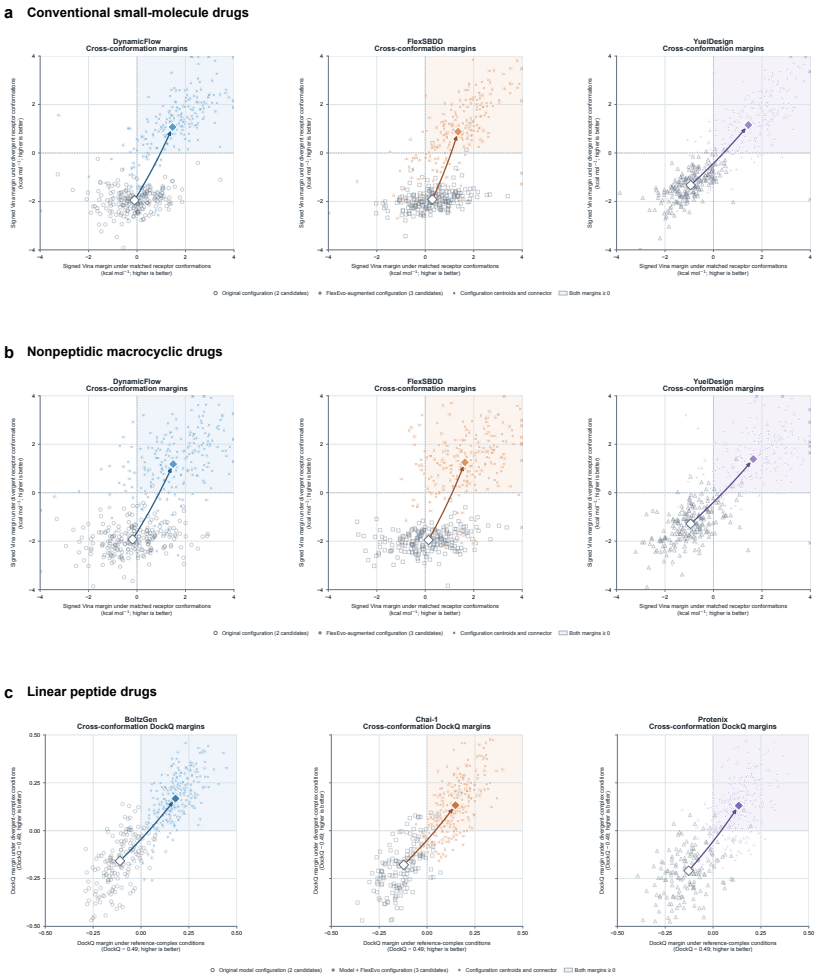

Figure 7: **Cross-conformation evaluation of conventional small molecules, nonpeptidic macrocycles and linear peptides.** **a**, Conventional small-molecule drugs. **b**, Nonpeptidic macrocyclic drugs. **c**, Linear peptide drugs. In **a** and **b**, columns show DynamicFlow, FlexSBDD and YuelDesign from left to right. The horizontal and vertical axes show signed Vina margins under matched and divergent receptor conformations, respectively (kcal mol^{-1}; higher is better). In **c**, columns show BoltzGen, Chai-1 and Protenix from left to right. The horizontal and vertical axes show DockQ margins under reference-complex and divergent-complex conditions, respectively, defined as $\mathrm{DockQ} - 0.49$. Grey open symbols and coloured points denote the original configuration (two candidates) and the FlexEvo-augmented configuration (three candidates), respectively. Diamond markers indicate configuration centroids, with connecting lines showing the displacement between configurations. Dashed lines mark zero margins, and the shaded upper-right quadrant indicates non-negative margins under both conditions.

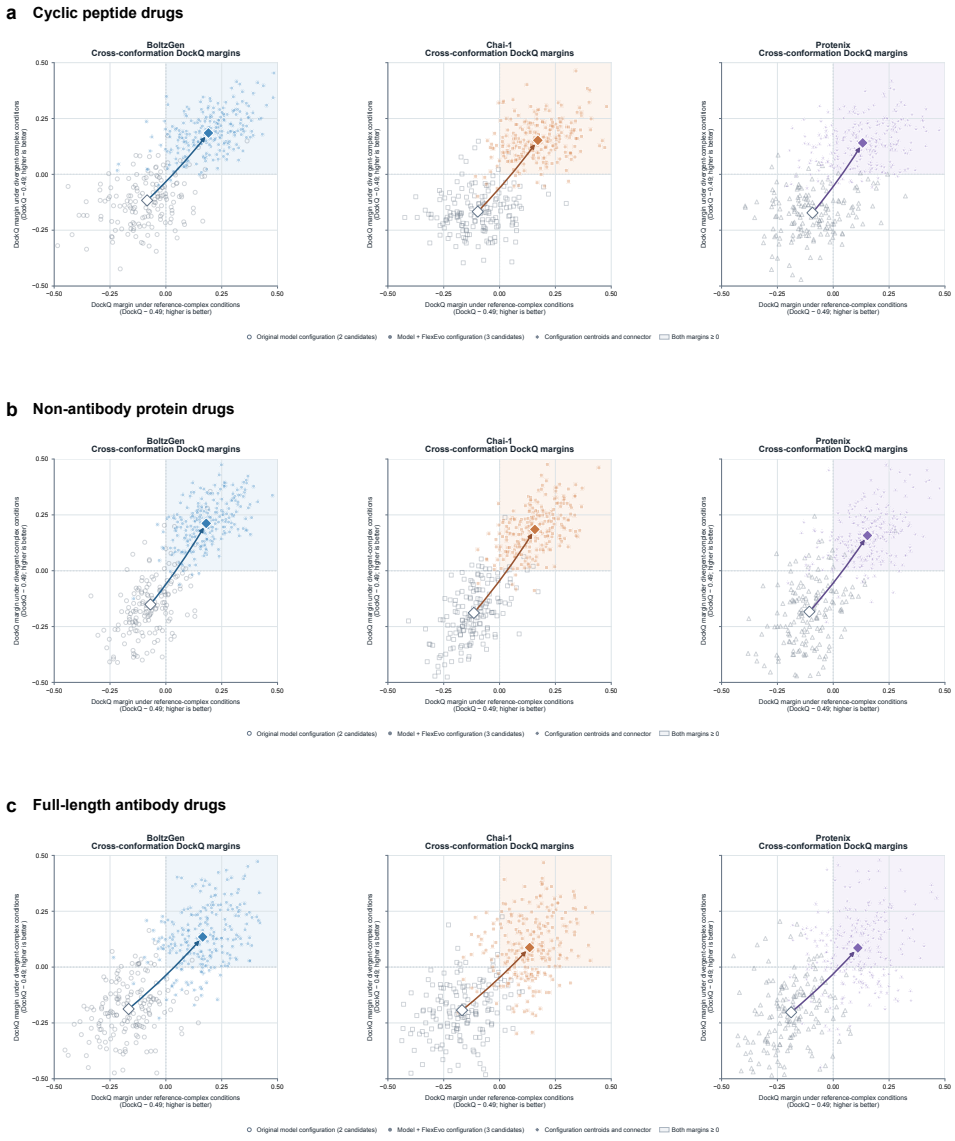

Figure 8: **Cross-conformation evaluation of cyclic peptides, non-antibody proteins and full-length antibodies. a**, Cyclic peptide drugs. **b**, Non-antibody protein drugs. **c**, Full-length antibody drugs. Columns show BoltzGen, Chai-1 and Protenix from left to right in each panel. The horizontal and vertical axes show DockQ margins under reference-complex and divergent-complex conditions, respectively. Margins are defined as $\mathrm{DockQ} - 0.49$, with higher values indicating better structural agreement with the reference complex under the corresponding condition. Grey open symbols and coloured points denote the original model configuration (two candidates) and the model supplemented with FlexEvo (three candidates), respectively. Diamond markers indicate configuration centroids, with connecting lines showing the displacement between configurations. Dashed lines mark zero margins, corresponding to $\mathrm{DockQ} = 0.49$. The shaded upper-right quadrant indicates $\mathrm{DockQ} \geq 0.49$ under both conditions.

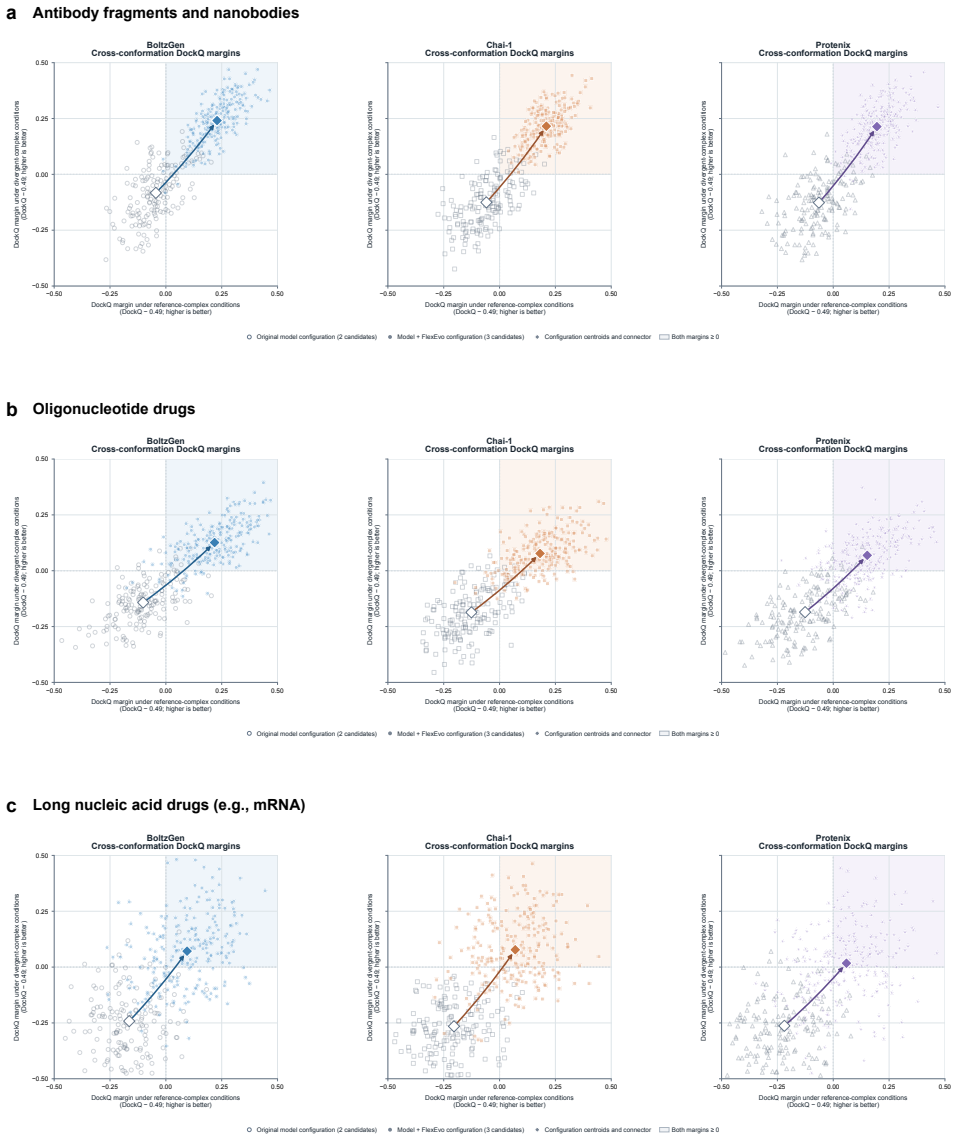

Figure 9: **Cross-conformation evaluation of antibody fragments, nanobodies and nucleic acid drugs.** **a**, Antibody fragments and nanobodies. **b**, Oligonucleotide drugs. **c**, Long nucleic acid drugs, including mRNA. Columns show BoltzGen, Chai-1 and Protenix from left to right in each panel. The horizontal and vertical axes show DockQ margins under reference-complex and divergent-complex conditions, respectively. Margins are defined as $\mathrm{DockQ} - 0.49$, with higher values indicating better structural agreement with the reference complex under the corresponding condition. Grey open symbols and coloured points denote the original model configuration (two candidates) and the model supplemented with FlexEvo (three candidates), respectively. Diamond markers indicate configuration centroids, with connecting lines showing the displacement between configurations. Dashed lines mark zero margins, corresponding to $\mathrm{DockQ} = 0.49$. The shaded upper-right quadrant indicates $\mathrm{DockQ} \geq 0.49$ under both conditions.

C.4 ALTERNATIVE OPTIMIZERS AND COMPONENT ABLATIONS

Figure 10 compares six optimization strategies under the displayed HV-AUC and B-pocket summaries. The two criteria should be read separately because their scales differ. The ranking summarizes the reported scores. The budget-matching protocol specifies initialization, objective-evaluation count, stopping rule, hypervolume reference point, and repetition procedure for each optimizer.

Comparison with Best-of-N. Table 16 summarizes the Best-of-N and FLEXEVO values displayed in Figure 10. The reported HV-AUC is 8.482 for FLEXEVO and 0.959 for Best-of-N, corresponding to an approximately 8.84-fold ratio. The corresponding B-pocket top-10 task scores are -31.409 and -32.497, giving a 1.088-point advantage for FLEXEVO.

Table 16: Best-of-N comparison using the values displayed in Figure 10. Higher values are preferred for both metrics.

Method	HV-AUC ↑	B-pocket top-10 ↑
Best-of-N	0.959	-32.497
FLEXEVO	8.482	-31.409

Table 11 displays the full method and three component-removal variants. Component-removal comparisons show that the full method achieves the best Divergent-state primary metric across all nine binder categories. The table includes trade-offs: for example, the small-molecule Reference Vina is slightly more favorable without FlexBox (-8.03 versus -8.01), whereas the full method has the more favorable Divergent Vina (-6.66 versus -5.88).

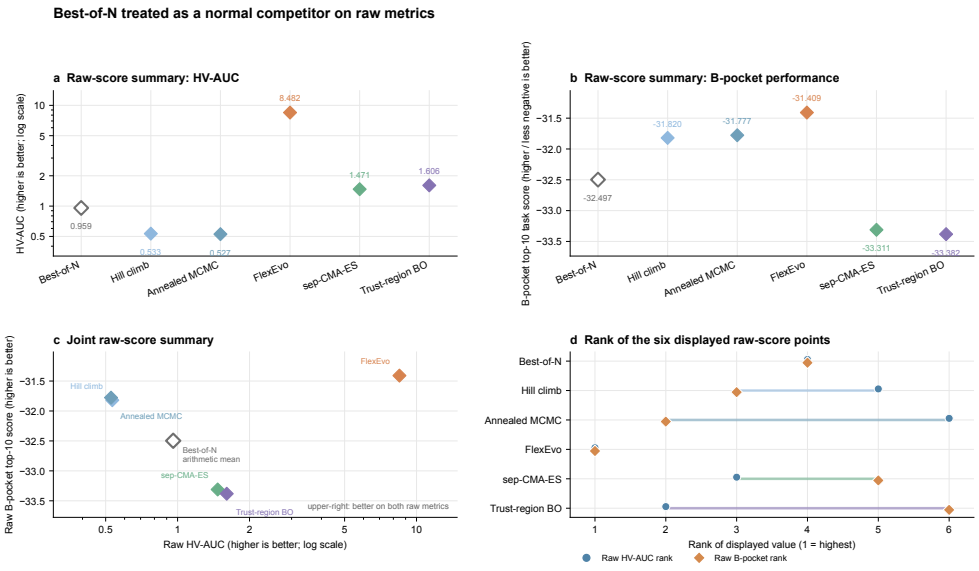

Figure 10: **Comparison of FlexEvo and alternative optimization strategies using raw performance metrics.** Six methods are compared: Best-of-N, hill climbing, annealed Markov chain Monte Carlo (MCMC), FlexEvo, separable covariance matrix adaptation evolution strategy (sep-CMA-ES) and trust-region Bayesian optimization (BO). **a,** Hypervolume area under the curve (HV-AUC), displayed on a logarithmic scale; higher values indicate better performance. **b,** B-pocket top-10 task scores, for which higher (less negative) values indicate better performance. **c,** Joint comparison of the two metrics, with the upper-right region indicating better performance on both. The open grey diamond denotes the Best-of-N arithmetic-mean summary. **d,** Rankings of the six displayed values for HV-AUC (blue circles) and B-pocket score (orange diamonds), with rank 1 indicating the highest value. Connecting lines link the two metric-specific ranks for each method. FlexEvo achieves the highest displayed value on both metrics, with an HV-AUC of 8.482 and a B-pocket top-10 score of -31.409.

C.5 RUNTIME AND STATE-WISE SUMMARY PROFILES

Figure 11 separates source generation from the additional adaptation cost under a common timing boundary. Across the nine workflows, FLEXEVO adds 1.393–3.141 min per sample, yielding total generation-plus-adaptation times of 6.861–12.966 min per sample. The reference search schedule contains 72 genome evaluations, with execution and timing definitions given in Appendix B.3.

Time-resolved adaptation and structural progression. Figure 4 links the displayed adaptation trajectories to elapsed computation time. In panel (a), the Vina-defined high-affinity fraction reaches 60.0% at approximately two minutes and remains near that level late in the run. Panels (b,c) show an overall decrease in on-target pAE and an increase in off-target pAE over the displayed 2–3 min intervals. In the ligand example in panels (d–f), Vina improves from -7.900 through -10.491 to -12.666 kcal/mol across the early, intermediate, and late structures. These observations demonstrate progressive docking-score improvement in the illustrated structural sequence and favorable interaction-confidence trends in the displayed runs.

State-wise summary profiles. Figures 12–14 summarize supplied percentage profiles across Reference, Matched, and Divergent states. The displayed "performance" percentage is a figure-level summary, distinct from DockQ, high-affinity fraction, and LRMSD success. Metric-specific comparisons use their own definitions, selection rules, and denominators.

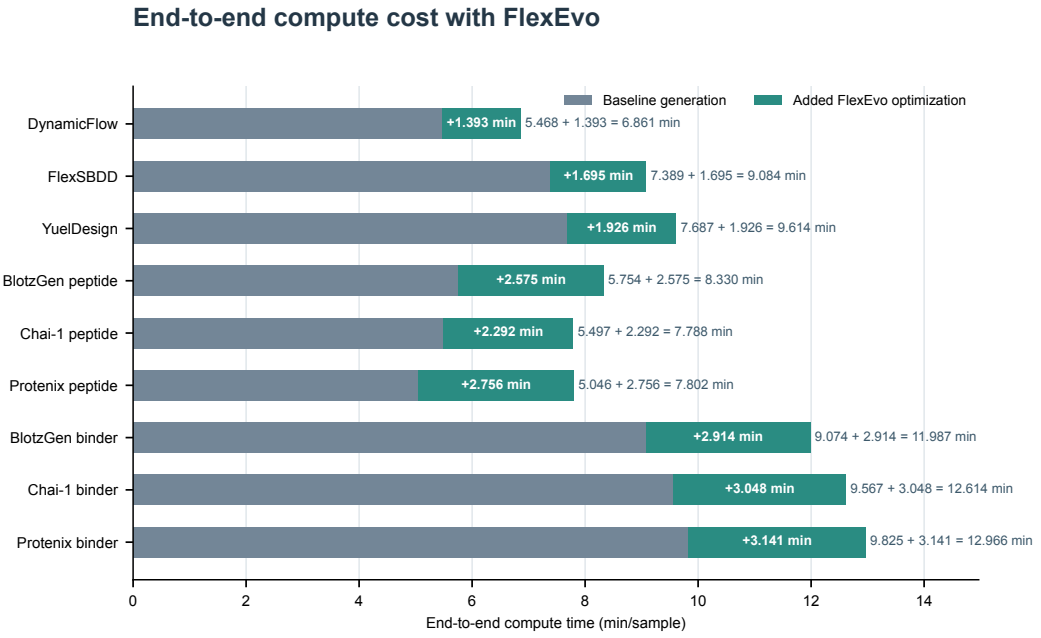

Figure 11: End-to-end computational cost of incorporating FlexEvo into generation workflows. Per-sample compute time is shown for DynamicFlow, FlexSBDD, YuelDesign, and the peptide and binder workflows of BoltzGen, Chai-1 and Protenix. Stacked bars separate baseline generation time (grey) from the additional time required for FlexEvo optimization (teal). Labels within the teal segments indicate the added optimization time, and annotations beside each bar report baseline, optimization and total times. Across the nine workflows, FlexEvo adds 1.393–3.141 min per sample, resulting in total compute times of 6.861–12.966 min per sample.

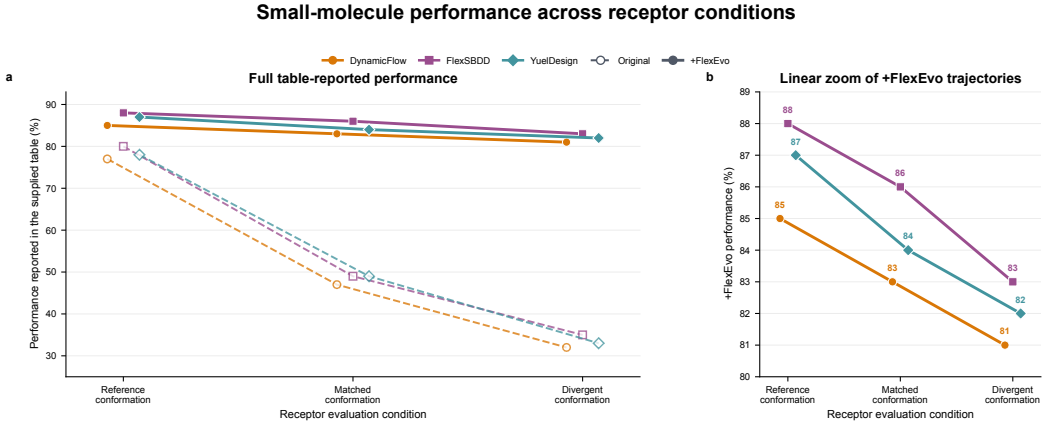

Figure 12: **Small-molecule performance across receptor conformations with and without FlexEvo. a,** Comparison of DynamicFlow (orange circles), FlexSBDD (purple squares) and YuelDesign (teal diamonds) under reference, matched and divergent receptor conformations. Dashed lines with open markers indicate the original methods; solid lines with filled markers indicate their FlexEvo-optimized counterparts. **b,** Expanded view of the FlexEvo trajectories on a linear scale, with labels indicating performance values (%). FlexEvo-optimized DynamicFlow, FlexSBDD and YuelDesign retain performance of 81%, 83% and 82%, respectively, under the divergent condition, with decreases of 4–5 percentage points relative to the reference condition.

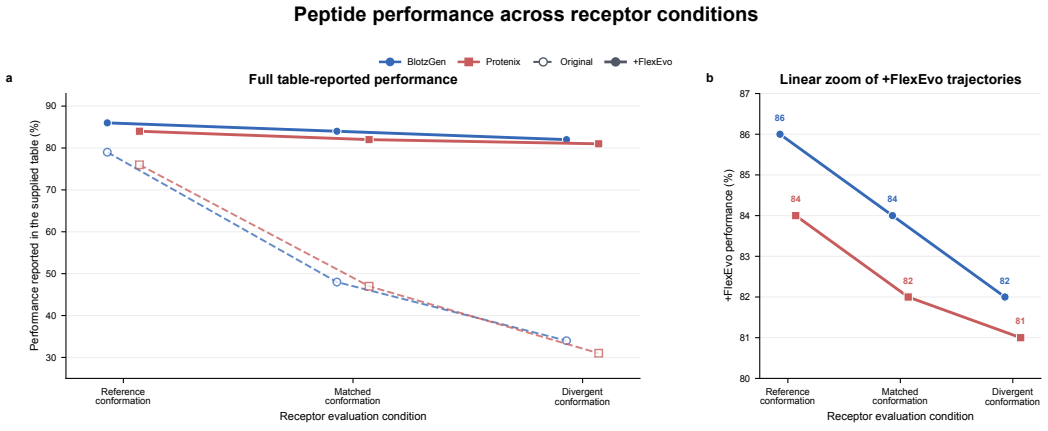

Figure 13: **Peptide performance across receptor conformations with and without FlexEvo. a,** Comparison of BoltzGen (blue circles) and Protenix (red squares) under reference, matched and divergent receptor conformations. Dashed lines with open markers indicate the original methods; solid lines with filled markers indicate their FlexEvo-optimized counterparts. **b,** Expanded view of the FlexEvo trajectories on a linear scale, with labels indicating performance values (%). BoltzGen achieves 86%, 84% and 82%, and Protenix achieves 84%, 82% and 81%, respectively, across the three conditions. FlexEvo reduces the decline in performance associated with changes in receptor conformation.

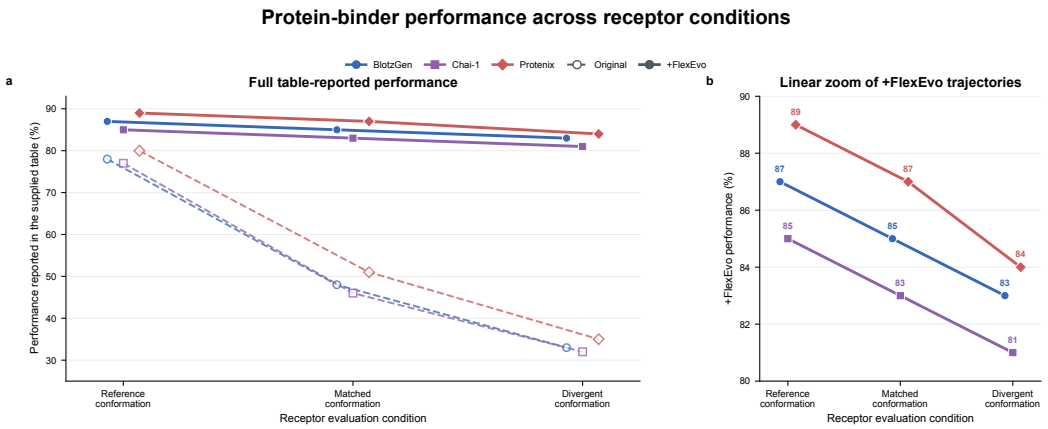

Figure 14: **Protein-binder performance across receptor conformations with and without FlexEvo. a,** Comparison of BoltzGen (blue circles), Chai-1 (purple squares) and Protenix (red diamonds) under reference, matched and divergent receptor conformations. Dashed lines with open markers indicate the original methods; solid lines with filled markers indicate their FlexEvo-optimized counterparts. **b,** Expanded view of the FlexEvo trajectories on a linear scale, with labels indicating performance values (%). Performance decreases from 87% to 83% for BoltzGen, from 85% to 81% for Chai-1 and from 89% to 84% for Protenix between the reference and divergent conditions, showing smaller declines than those observed for the original methods.

C.6 Similarity to a sampled reference library

Table 17 and Figures 15–17 characterize generated molecules relative to a sampled reference library. Each scenario comprises 470 pairwise comparisons between 10 generated and 47 reference entries, with each entry participating in multiple comparisons. Low similarity indicates limited resemblance within the sampled library under the molecular representation used.

Table 17: Reference-library similarity across nine binder categories. For each category, 10 evaluated candidates are compared with 47 randomly sampled public reference entries. Maximum and median similarity are reported over the resulting 470 pairwise comparisons using the modality-specific similarity definition.

Evaluation scenario	Maximum T_c	Median T_c
Antibody fragments and nanobodies	0.222	0.052
Conventional small-molecule drugs	0.159	0.052
Cyclic peptide drugs	0.262	0.055
Full-length antibody drugs	0.245	0.059
Linear peptide drugs	0.226	0.064
Long nucleic acid drugs, including mRNA	0.254	0.071
Non-antibody protein drugs	0.270	0.052
Nonpeptidic macrocyclic drugs	0.145	0.044
Oligonucleotide drugs	0.314	0.039

T_c, Tanimoto coefficient. Values summarize fingerprint similarity to the sampled reference molecules under the specified molecular representation.

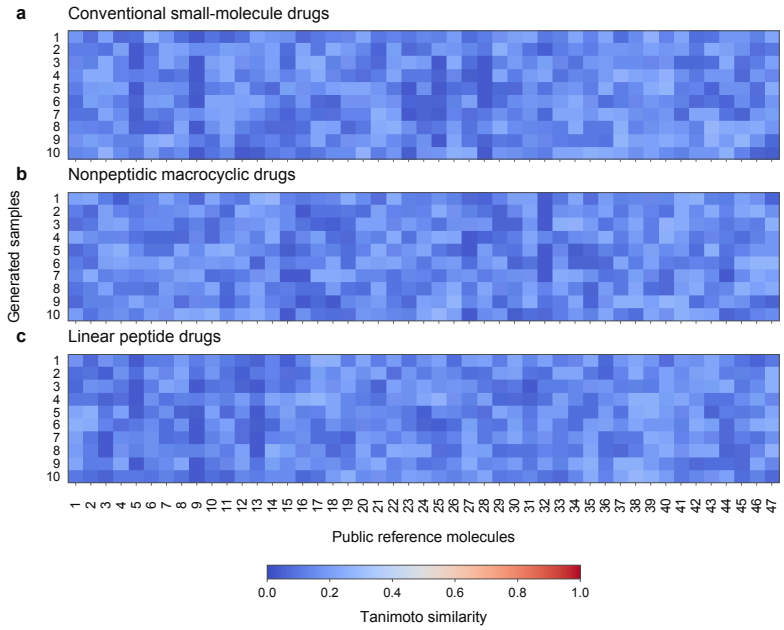

Figure 15: Tanimoto similarity between generated samples and public reference molecules for three drug modalities. **a**, Conventional small-molecule drugs. **b**, Nonpeptidic macrocyclic drugs. **c**, Linear peptide drugs. Each heatmap compares 10 generated samples (rows) with 47 public reference molecules (columns). Each cell represents the Tanimoto similarity of the corresponding pair, with colours indicating values from 0 to 1. Higher values indicate greater similarity, and all panels share the same colour scale.

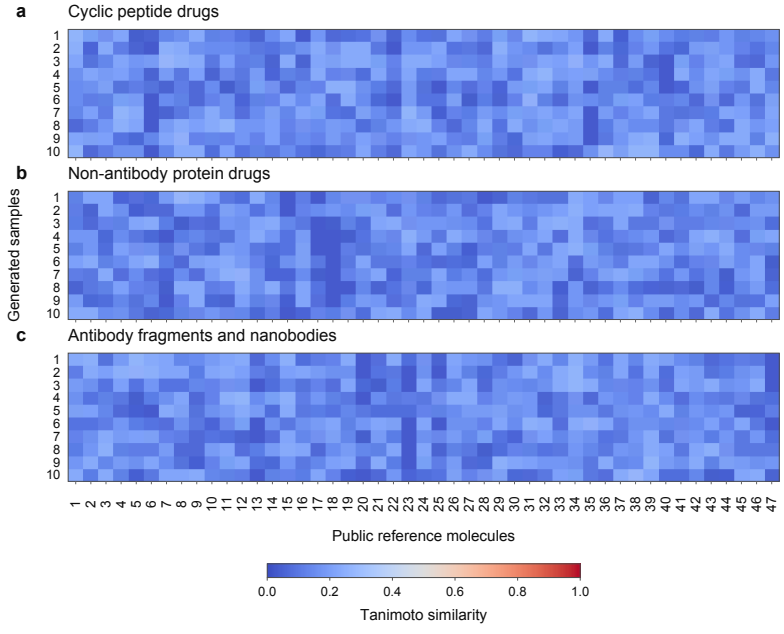

Figure 16: Tanimoto similarity between generated samples and public reference molecules for cyclic peptides and protein-based drug modalities. **a**, Cyclic peptide drugs. **b**, Non-antibody protein drugs. **c**, Antibody fragments and nanobodies. Each heatmap compares 10 generated samples (rows) with 47 public reference molecules (columns). Each cell represents the Tanimoto similarity of the corresponding pair, with colours indicating values from 0 to 1. Higher values indicate greater similarity, and all panels share the same colour scale.

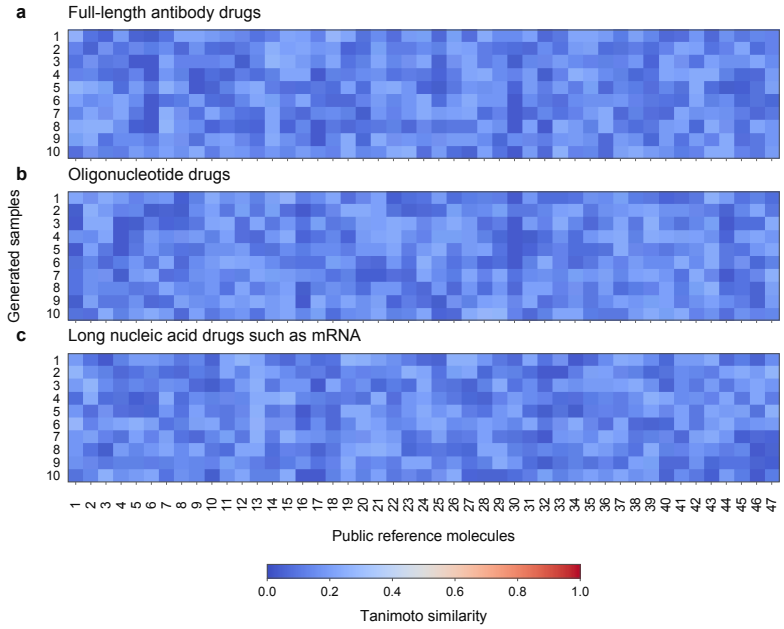

Figure 17: Tanimoto similarity between generated samples and public reference molecules for antibody and nucleic acid drug modalities. **a**, Full-length antibody drugs. **b**, Oligonucleotide drugs. **c**, Long nucleic acid drugs, such as mRNA. Each heatmap compares 10 generated samples (rows) with 47 public reference molecules (columns). Each cell represents the Tanimoto similarity of the corresponding pair, with colours indicating values from 0 to 1. Higher values indicate greater similarity, and all panels share the same colour scale.

C.7 Structural examples and physicochemical context

Figures 18–20 illustrate the architectures of the nine binder categories using publicly deposited reference structures and explicitly identified RNA schematics. Figure 21 summarizes physicochemical descriptors for 116 reference structures arranged in two illustrative sets of 58. Guideline lines provide context for interpreting the descriptor profiles.

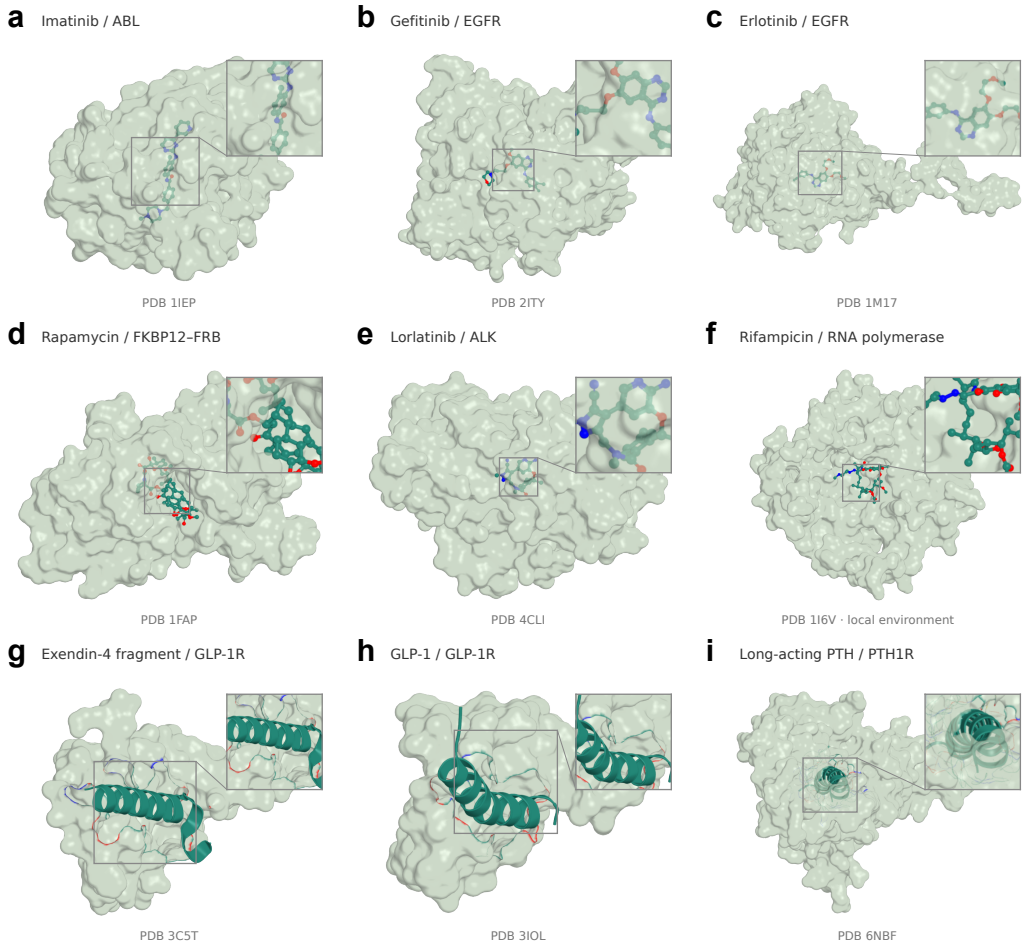

Figure 18: **Representative structures of conventional small molecules, nonpeptidic macrocycles and linear peptides. a–c**, Conventional small-molecule drugs: imatinib bound to ABL (PDB: 1IEP), gefitinib bound to EGFR (2ITY), and erlotinib bound to EGFR (1M17). **d–f**, Nonpeptidic macrocyclic drugs: rapamycin in the FKBP12–FRB complex (1FAP), lorlatinib bound to ALK (4CLI), and rifampicin bound to RNA polymerase (1I6V). Panel **f** shows the local protein environment surrounding rifampicin. **g–i**, Linear peptide examples: an exendin-4 fragment bound to the GLP-1 receptor extracellular domain (3C5T), GLP-1 bound to the same domain (3IOL), and a long-acting parathyroid hormone analogue bound to PTH1R (6NBF). Target proteins are shown as pale-green surfaces, with bound molecules represented as sticks or cartoons. Grey boxes identify regions reproduced in the enlarged insets. Structures are rendered from deposited coordinates and independently scaled.

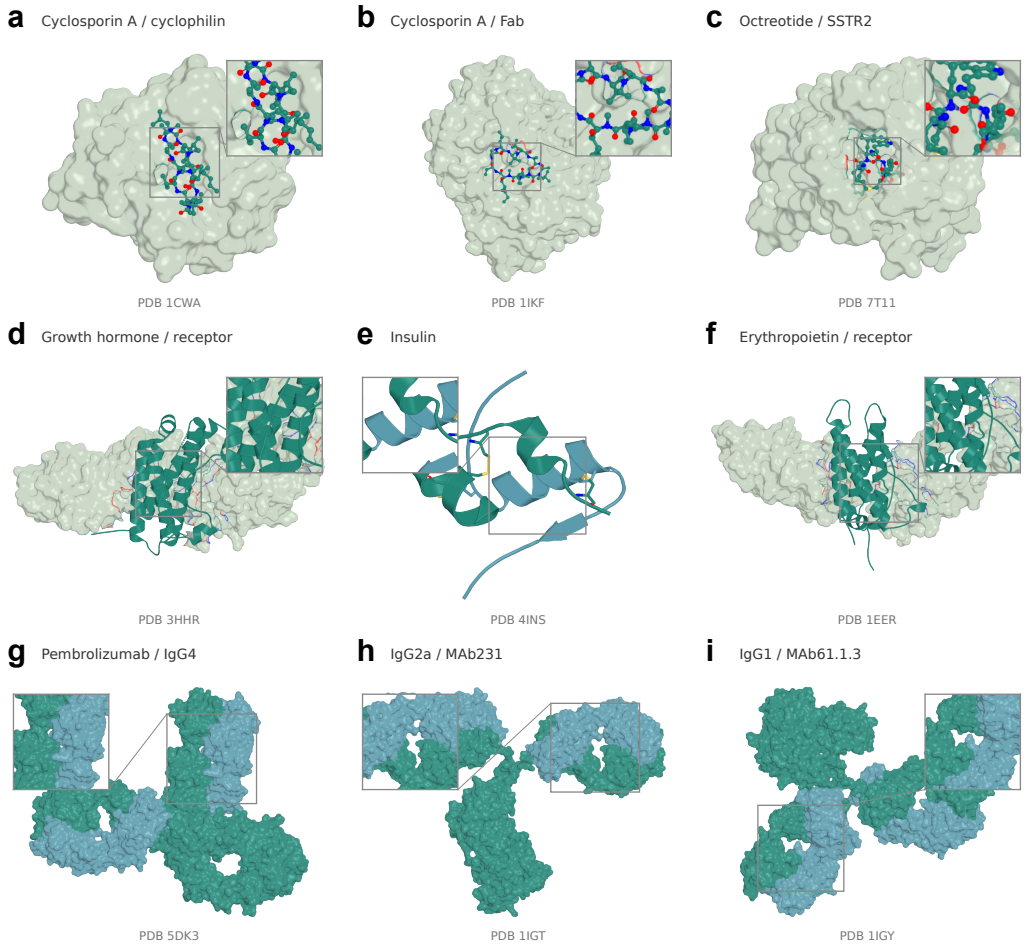

Figure 19: **Representative structures of cyclic peptides, non-antibody proteins and full-length antibodies. a–c**, Cyclic peptide examples: cyclosporin A bound to cyclophilin A (PDB: 1CWA), cyclosporin A bound to an antibody Fab fragment (1IKF), and octreotide bound to somatostatin receptor 2 (7T11). **d–f**, Non-antibody protein examples: human growth hormone bound to its receptor (3HHR), porcine insulin (4INS), and human erythropoietin bound to its receptor (1EER). **g–i**, Full-length antibody examples: pembrolizumab IgG4 (5DK3), IgG2a MAb231 (1IGT), and IgG1 MAb61.1.3 (1IGY). The latter two antibodies illustrate intact immunoglobulin architectures. Where present, target proteins are shown as pale-green surfaces. Protein ligands are shown as cartoons, whereas intact antibodies are shown as surfaces with heavy and light chains distinguished by colour. Grey boxes identify regions reproduced in the enlarged insets. Structures are rendered from deposited coordinates and independently scaled.

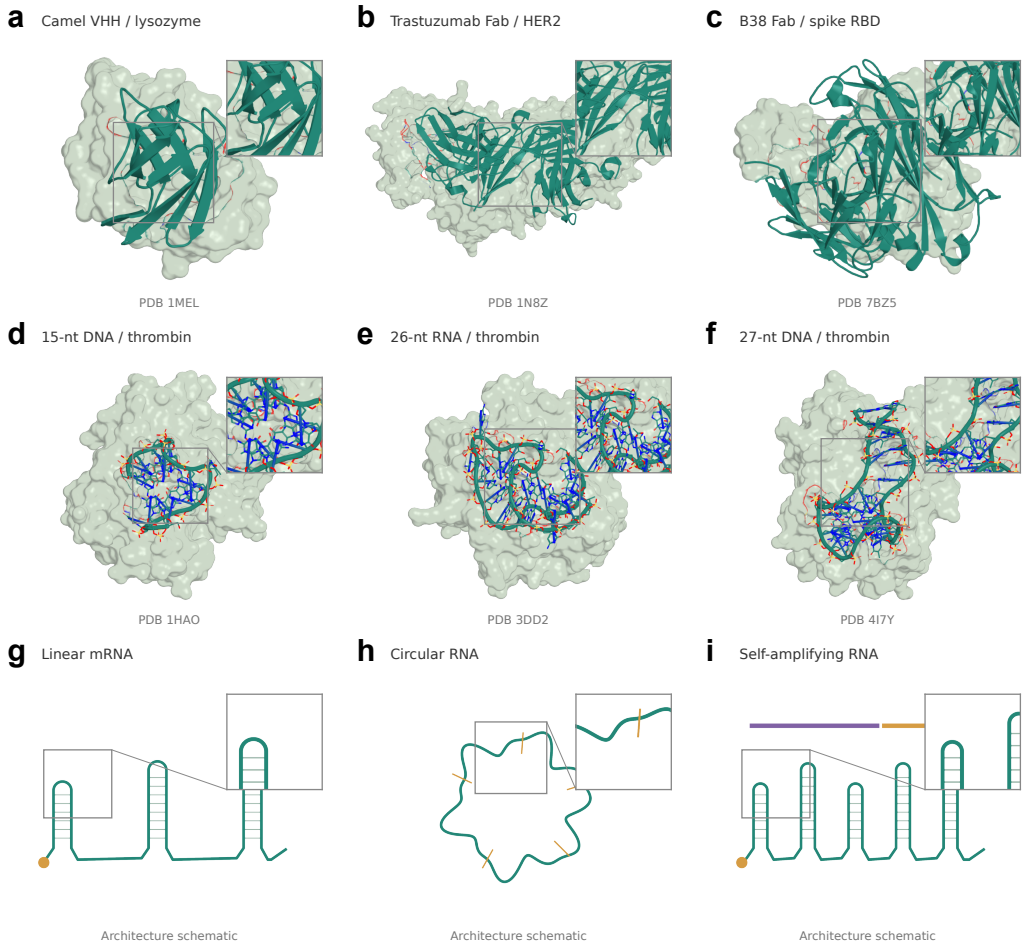

Figure 20: **Representative antibody-fragment and oligonucleotide structures, and long-RNA architectures. a–c**, Antibody fragments and nanobodies: a camel single-domain antibody bound to lysozyme (PDB: 1MEL), trastuzumab Fab bound to HER2 (1N8Z), and B38 Fab bound to the SARS-CoV-2 spike receptor-binding domain (7BZ5). **d–f**, Oligonucleotide examples: a 15-nucleotide DNA aptamer (1HAO), a 26-nucleotide RNA aptamer (3DD2), and a 27-nucleotide DNA aptamer (4I7Y), each bound to thrombin. These aptamers serve as structural representatives of the oligonucleotide modality. Target proteins are shown as pale-green surfaces, with antibody fragments and nucleic acids represented as cartoons or sticks. **g–i**, Schematic architectures of linear mRNA, circular RNA and self-amplifying RNA, respectively. RNA schematics illustrate molecular organization without specifying sequence, chain length or experimentally determined folding. Grey boxes identify regions reproduced in the enlarged insets. Molecular views are independently scaled.

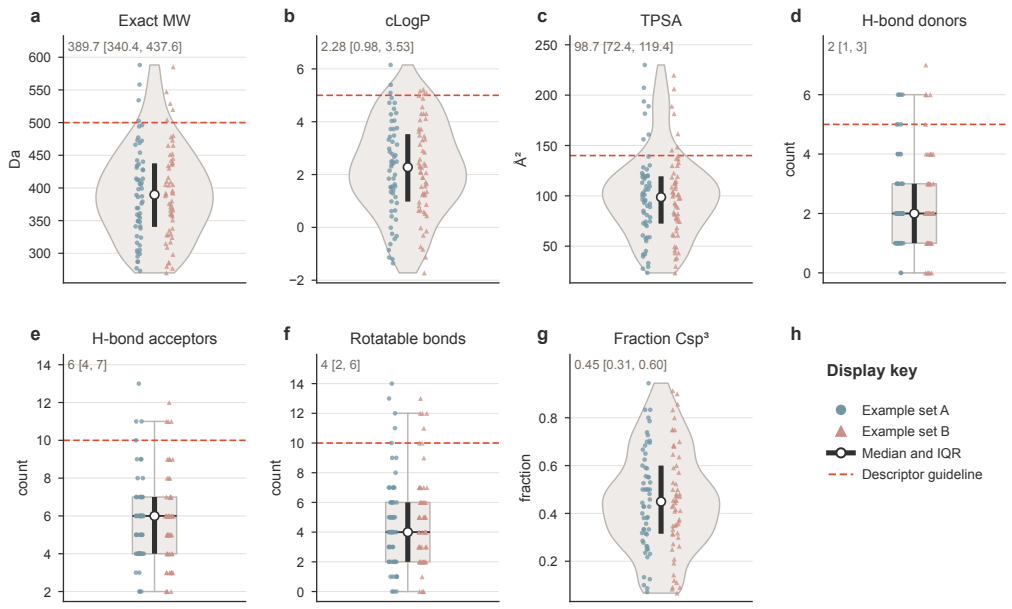

Figure 21: **Physicochemical profiles of selected reference molecules.** The figure summarizes seven molecular descriptors for 116 reference structures selected for illustration and divided into two example sets ($n = 58$ each), shown as blue circles and terracotta triangles. **a–g,** Exact molecular mass (Da), calculated octanol–water partition coefficient (cLogP), topological polar surface area (TPSA; Å^2), hydrogen-bond donor count, hydrogen-bond acceptor count, rotatable bond count and the fraction of carbon atoms with sp^3 hybridization, respectively. Grey violins show pooled distributions of continuous descriptors; boxes show the first-to-third-quartile intervals for count descriptors, with whiskers extending to the most extreme observations within 1.5 interquartile ranges (IQRs) of the box boundaries. Open black circles and thick black intervals indicate pooled medians and IQRs. Numerical annotations report median [Q1, Q3], and each coloured marker represents one molecule. Red dashed lines mark commonly used physicochemical guideline upper bounds: molecular mass of 500 Da, cLogP of 5, TPSA of 140 Å^2, 5 hydrogen-bond donors, 10 hydrogen-bond acceptors and 10 rotatable bonds. For all six descriptors with reference lines, the pooled median and upper quartile lie below the corresponding guideline, although individual molecules exceed these values. The median fraction of sp^3-hybridized carbon atoms is 0.45, with an IQR of 0.31–0.60. **h,** Key to the graphical encodings.

C.8 Multimetric profiles

Figures 22–24 visualize metric trade-offs using common normalization bounds across the 18 method–state combinations within each modality. Each radial axis supports comparison within its category-specific bounds. Interpretation follows the individual metric axes; the logP axis displays hydrophobicity according to the stated plotting convention.

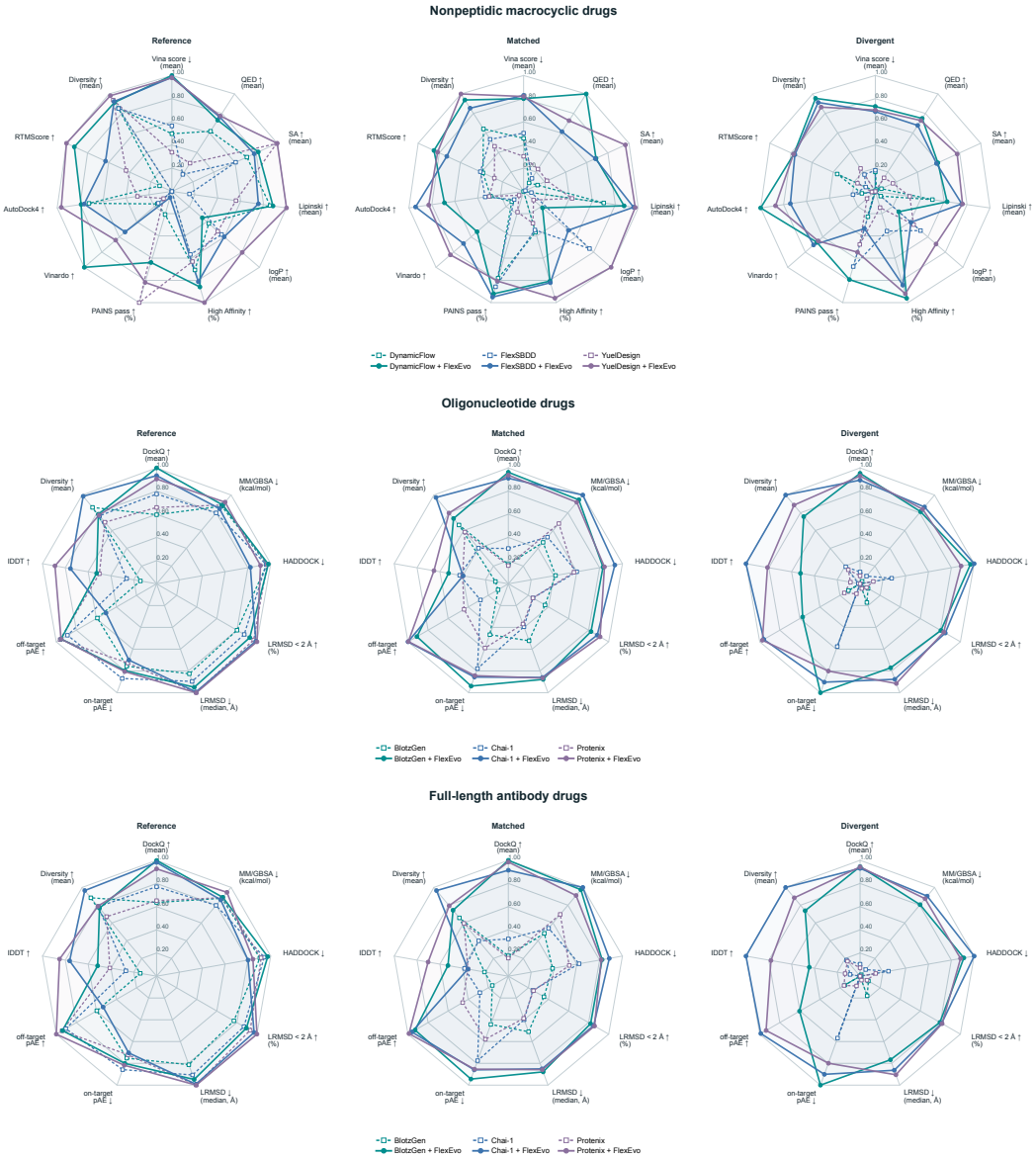

Figure 22: **Performance comparisons for nonpeptidic macrocycles, oligonucleotides and full-length antibodies.** Rows show nonpeptidic macrocyclic drugs, oligonucleotide drugs and full-length antibody drugs, respectively. Columns correspond to reference, matched and divergent conformations, from left to right. DynamicFlow, FlexSBDD and YuelDesign are compared with their FlexEvo variants for macrocycles; BoltzGen, Chai1 and Protenix are compared with their FlexEvo variants for oligonucleotides and antibodies. Colors identify model families within each row. Dashed lines with open squares indicate base models, whereas solid lines with filled circles indicate the corresponding FlexEvo variants. Each metric is min–max normalized to $[0, 1]$ across all 18 method–conformation combinations within the same drug modality. Metrics marked with downward arrows are reverse-scaled, so outward values follow the indicated preferred direction. Normalization bounds are shared across the three columns within each row but differ between modalities. Means are used where indicated; LRMSD success rates below 2 Å and median LRMSD are shown separately. The logP axis displays increasing hydrophobicity according to the stated plotting convention. Performance comparisons follow the individual metric axes.

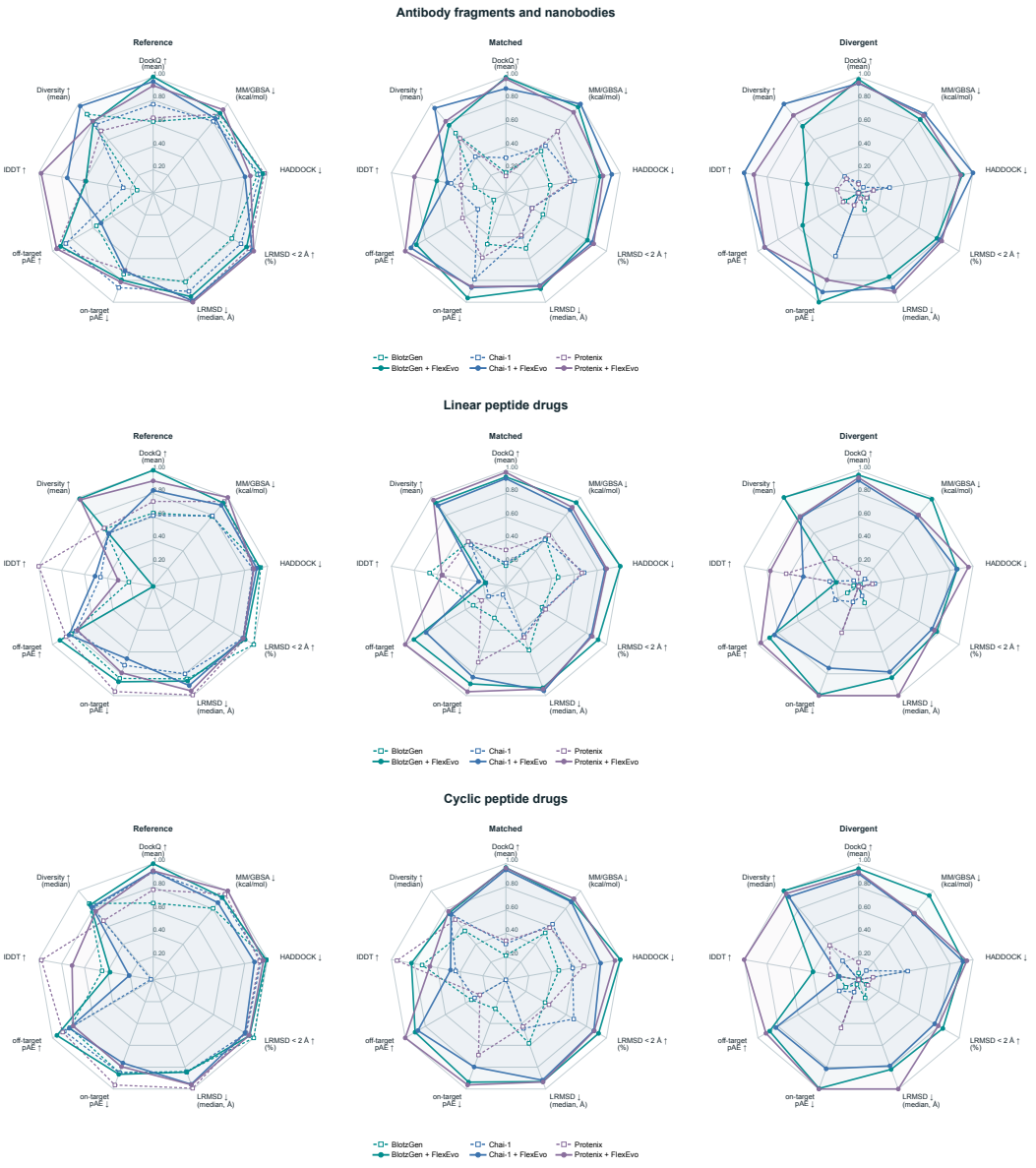

Figure 23: **Performance comparisons for antibody fragments, nanobodies and peptide drugs.** Rows show antibody fragments and nanobodies, linear peptide drugs and cyclic peptide drugs, respectively. Columns correspond to reference, matched and divergent conformations, from left to right. Each panel compares BoltzGen, Chai1 and Protenix with their corresponding FlexEvo variants. Colors identify model families within each row. Dashed lines with open squares indicate base models, whereas solid lines with filled circles indicate the corresponding FlexEvo variants. Each metric is min–max normalized to $[0, 1]$ across all 18 method–conformation combinations within the same drug modality. Metrics marked with downward arrows are reverse-scaled, so outward values follow the indicated preferred direction. Normalization bounds are shared across the three columns within each row but differ between modalities. Means are used where indicated; LRMSD success rates below 2 Å and median LRMSD are shown separately. Diversity is represented by the median for cyclic peptides and by the mean for the other two modalities. Performance comparisons follow the individual metric axes.

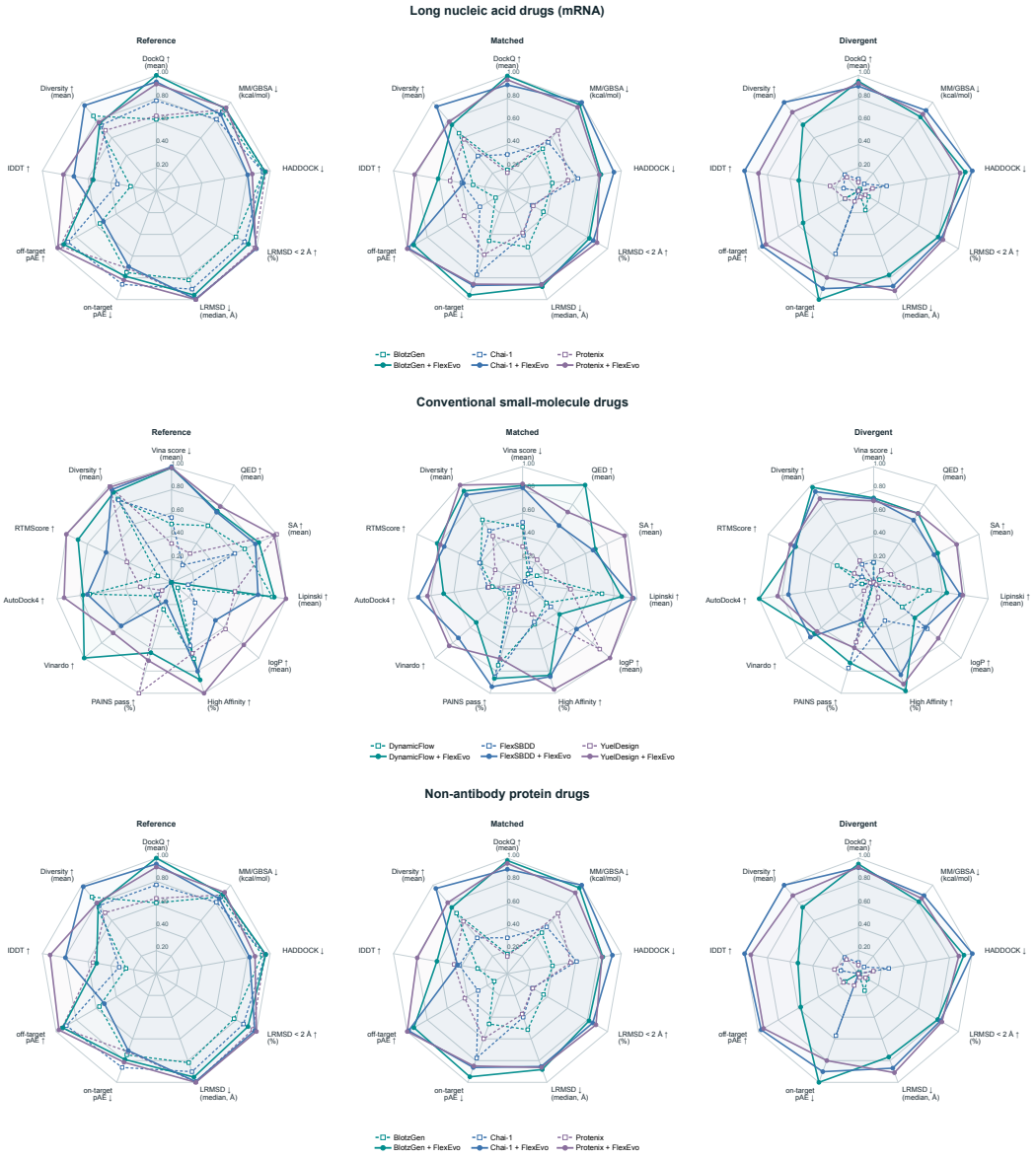

Figure 24: **Performance comparisons for long nucleic acids, small molecules and non-antibody proteins.** Rows show long nucleic acid drugs such as mRNA, conventional small-molecule drugs and non-antibody protein drugs, respectively. Columns correspond to reference, matched and divergent conformations, from left to right. BoltzGen, Chai1 and Protenix are compared with their FlexEvo variants for nucleic acids and proteins; DynamicFlow, FlexSBDD and YuelDesign are compared with their FlexEvo variants for small molecules. Colors identify model families within each row. Dashed lines with open squares indicate base models, whereas solid lines with filled circles indicate the corresponding FlexEvo variants. Each metric is min–max normalized to $[0, 1]$ across all 18 method–conformation combinations within the same drug modality. Metrics marked with downward arrows are reverse-scaled, so outward values follow the indicated preferred direction. Normalization bounds are shared across the three columns within each row but differ between modalities. Means are used where indicated; LRMSD success rates below 2 Å and median LRMSD are shown separately. The logP axis displays increasing hydrophobicity according to the stated plotting convention. Performance comparisons follow the individual metric axes.

C.9 On-target and off-target score diagnostics

Figures 25–27 summarize on/off-target pAE diagnostics using the score definitions in App. B.2. Highlighted candidates have on-target pAE below 5 and minimum off-target pAE above 10 across the evaluated off-target panel. Annotated percentages report the fraction of plotted candidates meeting both criteria. These model-confidence diagnostics summarize the on/off-target contrast within the evaluated panel.

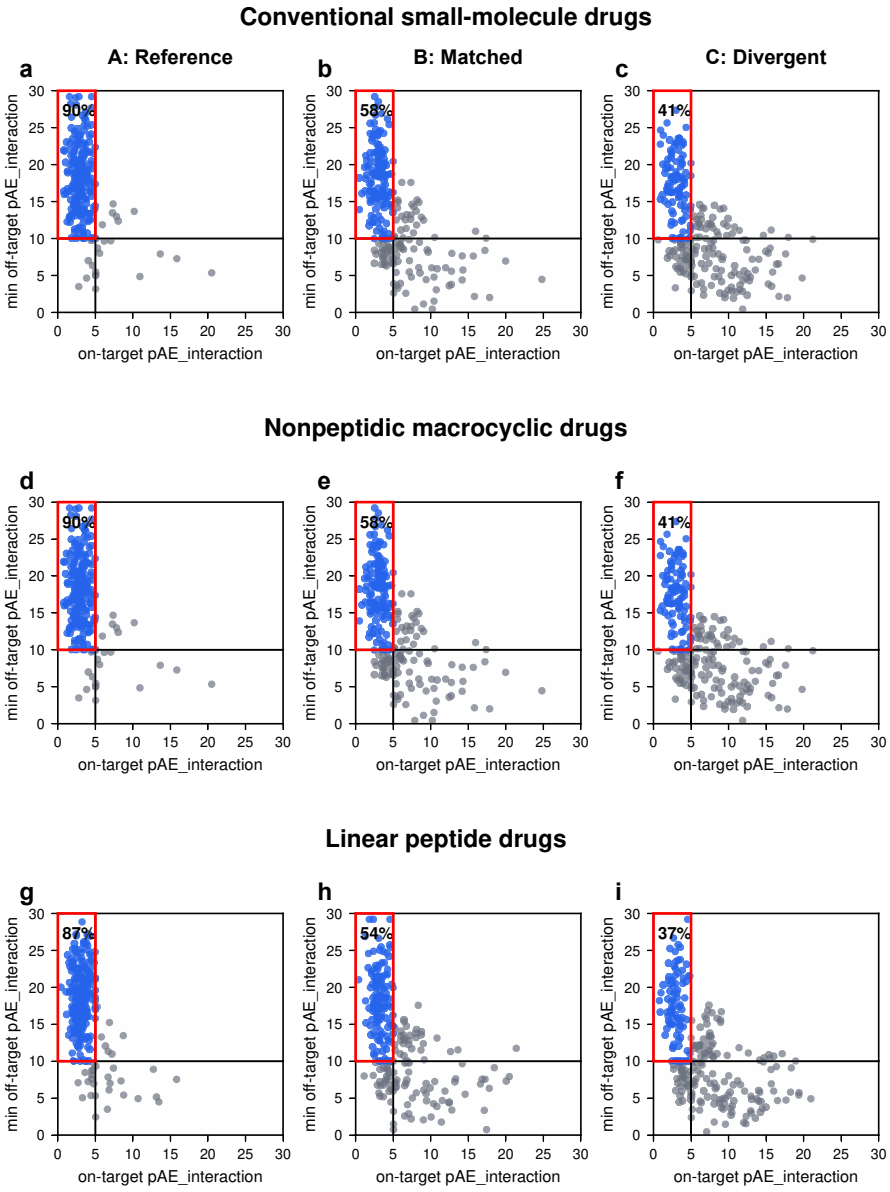

Figure 25: On-target and off-target interaction scores for small-molecule, macrocyclic and linear peptide drug candidates. Rows show **a**, conventional small-molecule drugs; **b**, nonpeptidic macrocyclic drugs; and **c**, linear peptide drugs. Columns retain the original A–C designations. Each point represents a candidate, plotted by its on-target `pAE_interaction` score and the minimum score across the evaluated off-targets. Reference lines mark an on-target score of 5 and a minimum off-target score of 10. Red boxes highlight the region with on-target scores below 5 and minimum off-target scores above 10. Blue points fall within this region, whereas grey points fall outside it. Annotated percentages indicate the proportion of candidates within the highlighted region.

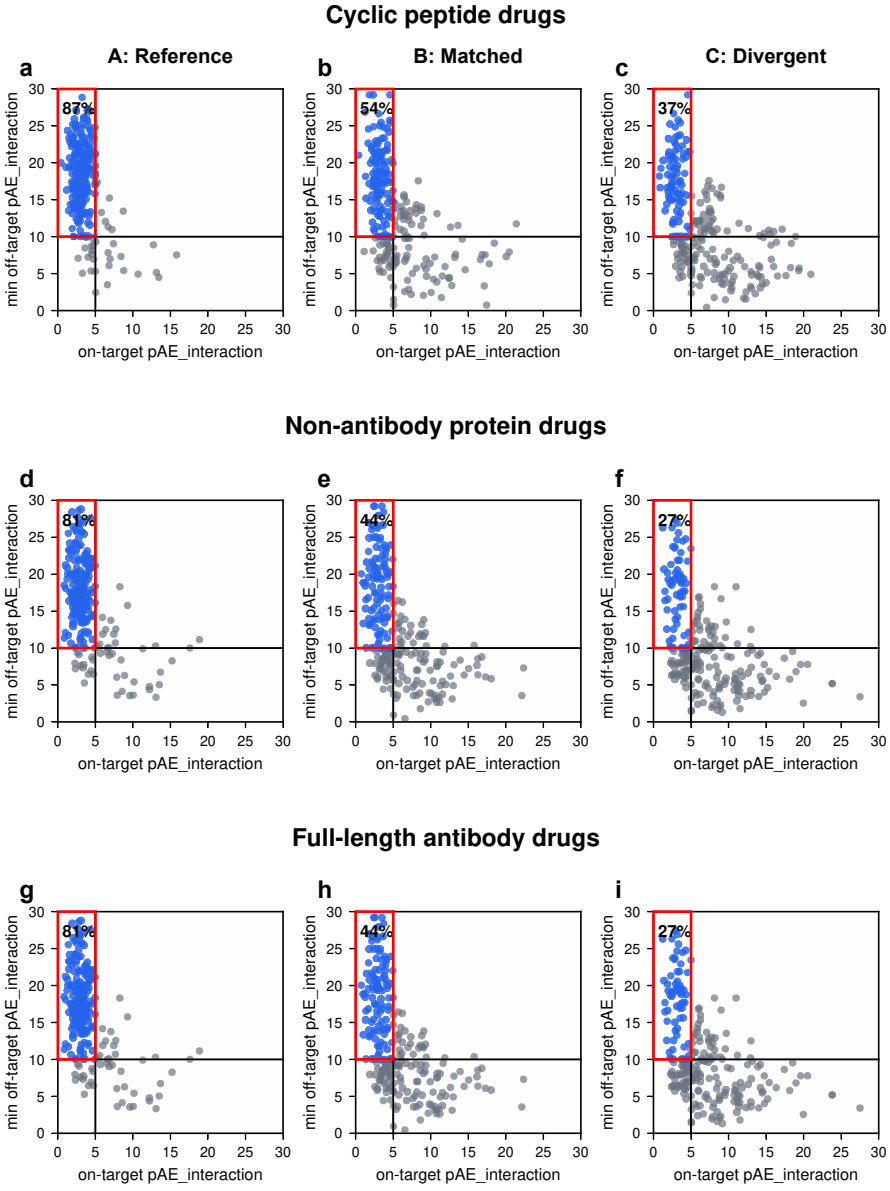

Figure 26: On-target and off-target interaction scores for cyclic peptide, protein and full-length antibody drug candidates. Rows show **a**, cyclic peptide drugs; **b**, non-antibody protein drugs; and **c**, full-length antibody drugs. Columns retain the original A–C designations. Each point represents a candidate, plotted by its on-target `pAE_interaction` score and the minimum score across the evaluated off-targets. Reference lines mark an on-target score of 5 and a minimum off-target score of 10. Red boxes highlight the region with on-target scores below 5 and minimum off-target scores above 10. Blue points fall within this region, whereas grey points fall outside it. Annotated percentages indicate the proportion of candidates within the highlighted region.

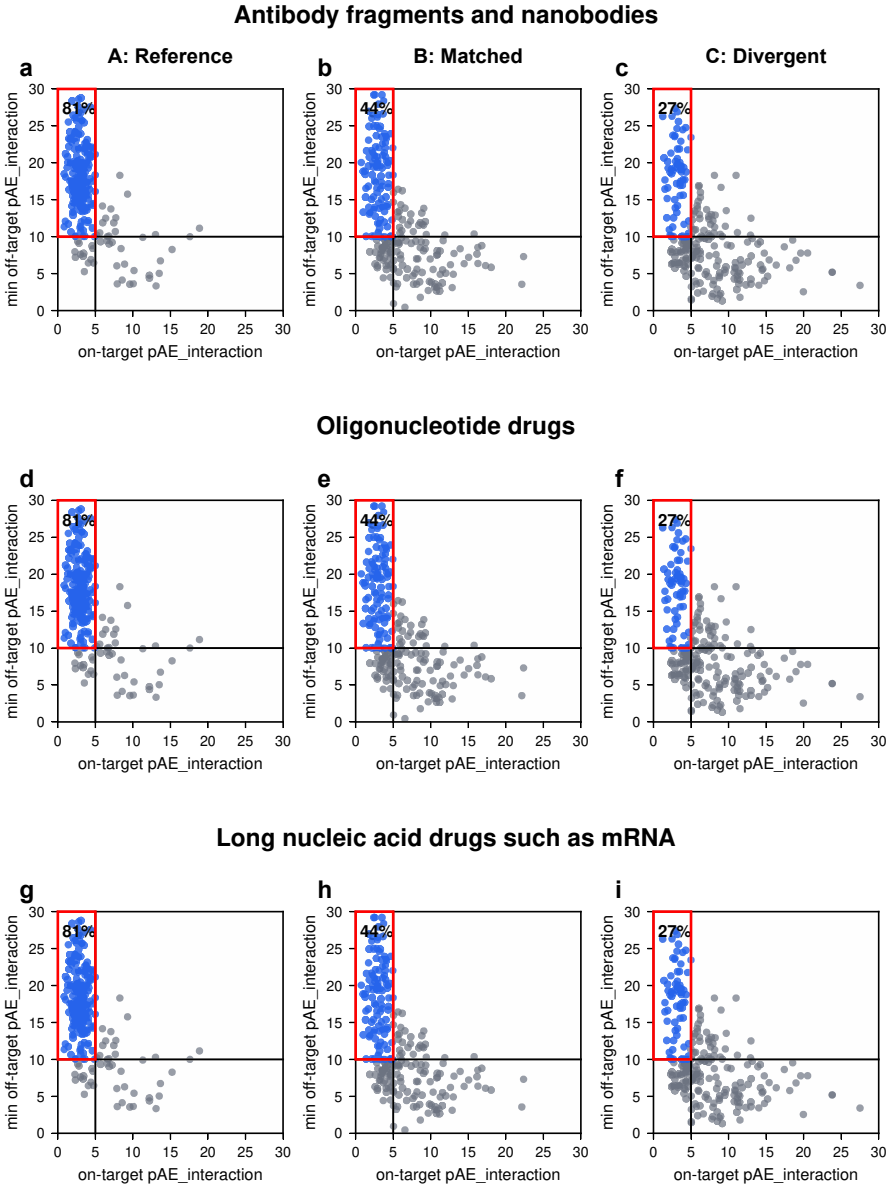

Figure 27: On-target and off-target interaction scores for antibody fragment and nucleic acid drug candidates. Rows show **a**, antibody fragments and nanobodies; **b**, oligonucleotide drugs; and **c**, long nucleic acid drugs, such as mRNA. Columns retain the original A–C designations. Each point represents a candidate, plotted by its on-target `pAE_interaction` score and the minimum score across the evaluated off-targets. Reference lines mark an on-target score of 5 and a minimum off-target score of 10. Red boxes highlight the region with on-target scores below 5 and minimum off-target scores above 10. Blue points fall within this region, whereas grey points fall outside it. Annotated percentages indicate the proportion of candidates within the highlighted region.

C.10 Optimization trajectories across binder categories

Figures 28–30 report the recorded optimization trajectories across the nine binder categories. The horizontal axes show cumulative FlexEvo optimization time per sample. The curves summarize metric evolution during adaptation, including local fluctuations and late-stage stabilization. Shaded bands summarize variability across the recorded experimental runs. Figure 4 presents representative adaptation dynamics in the main paper, while Appendix C.5 reports the computational cost analysis.

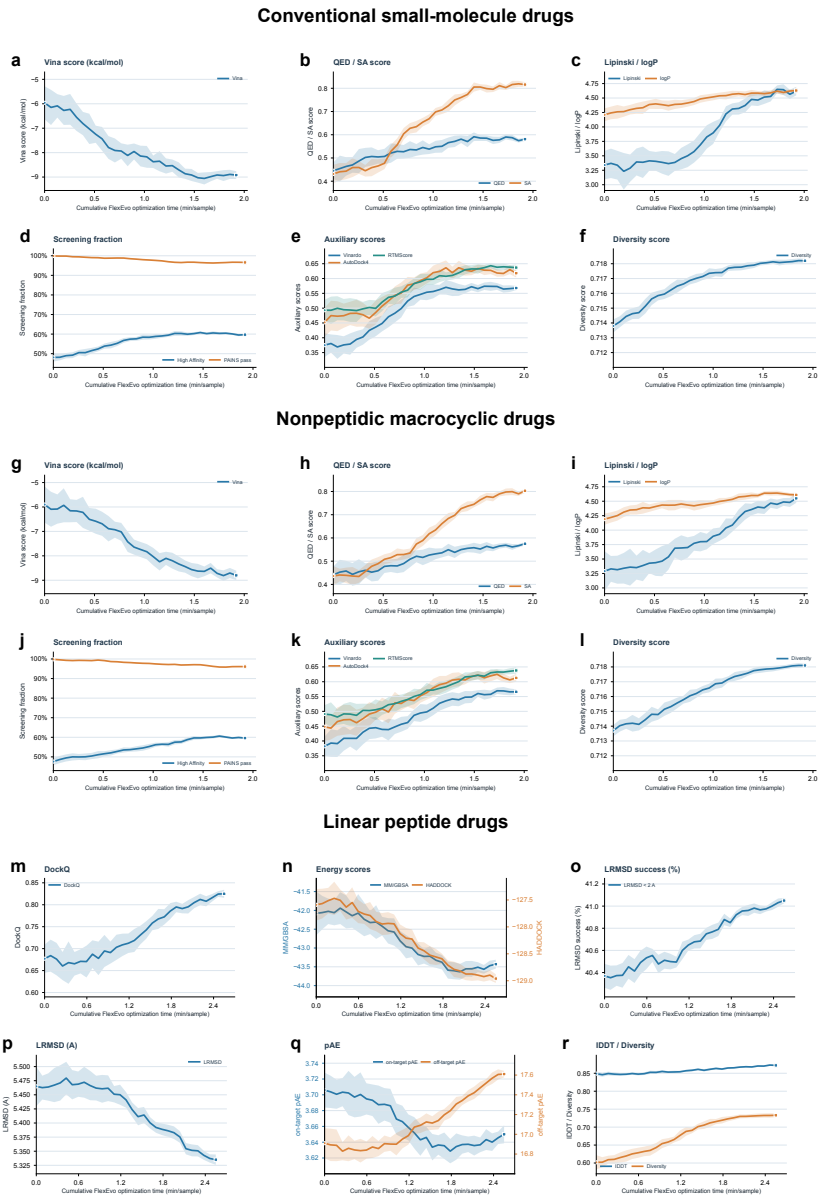

Figure 28: Optimization trajectories across conventional small-molecule drugs, nonpeptidic macrocyclic drugs, and linear peptide drugs. Panels a–f, g–l, and m–r correspond to the three categories, respectively. For the two small-molecule categories, panels a and g show Vina scores; b and h, QED and SA scores; c and i, Lipinski scores and logP; d and j, high-affinity and PAINS-pass fractions; e and k, auxiliary scoring functions; and f and l, diversity. For linear peptides, panels m–r show DockQ, MM/GBSA and HADDOCK scores, LRMSD success rate, LRMSD, on-target and off-target pAE, and IDDT and diversity, respectively. The horizontal axes indicate cumulative FlexEvo optimization time per sample. The recorded trajectories show metric evolution during adaptation, including local fluctuations and late-stage stabilization. Shaded bands summarize variability across the recorded experimental runs.

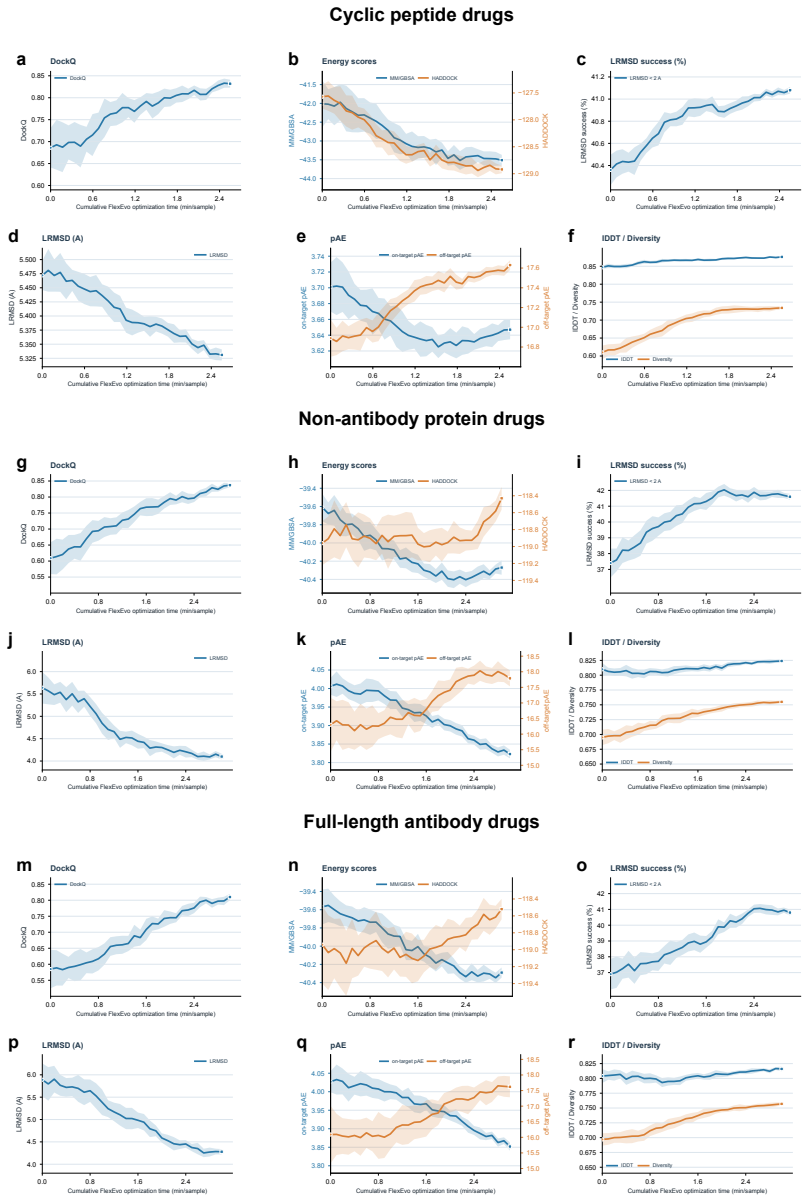

Figure 29: Optimization trajectories across cyclic peptide drugs, non-antibody protein drugs, and full-length antibody drugs. Panels a–f, g–l, and m–r correspond to the three categories, respectively. Within each category, the six panels report DockQ, MM/GBSA and HADDOCK scores, LRMSD success rate, LRMSD, on-target and off-target pAE, and IDDT and diversity, in that order. The horizontal axes indicate cumulative FlexEvo optimization time per sample. The recorded trajectories show increasing DockQ and LRMSD success rates, decreasing LRMSD and on-target pAE, and increasing diversity over optimization, with local fluctuations and late-stage stabilization. Differences in turning points and convergence profiles illustrate category-dependent optimization behavior. Energy-based metrics are interpreted according to their metric-specific preferred directions. Panels displaying paired metrics use separate left and right axes, with colors matched to the corresponding legends.

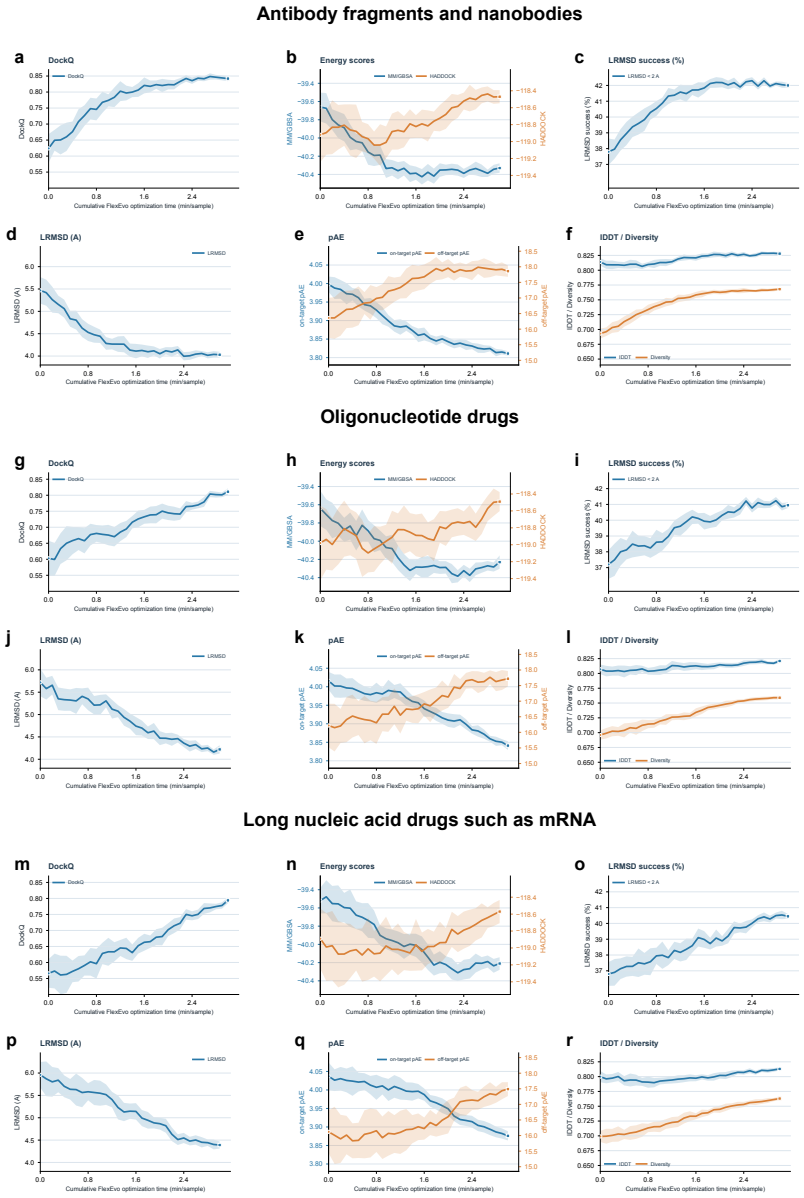

Figure 30: Optimization trajectories across antibody fragments and nanobodies, oligonucleotide drugs, and long nucleic acid drugs such as mRNA. Panels a–f, g–l, and m–r correspond to the three categories, respectively. Within each category, the six panels report DockQ, MM/GBSA and HADDOCK scores, LRMSD success rate, LRMSD, on-target and off-target pAE, and IDDT and diversity, in that order. The horizontal axes indicate cumulative FlexEvo optimization time per sample. The recorded trajectories show higher DockQ and LRMSD success rates, lower LRMSD and on-target pAE, and increased diversity over optimization, together with local fluctuations and category-dependent convergence profiles. Energy and pAE panels use separate left and right axes, with colors identifying the corresponding metrics.

D DETERMINISTIC SCORE BOUNDS AND CONDITIONAL ROBUSTNESS OF FLEXEVO

This appendix analyzes the coordinate-level scores and selection rules defined in Appendix A. In particular, the contact and clash kernels reproduce Eq. equation 17, and the objective coefficients reproduce Eqs. equation 23 and equation 25. The results concern a fixed candidate pose and its geometric contact and clash descriptors. Adaptation uses a single observed conformation P^0 and synthetic perturbation probes derived from it. An unseen target enters only through an explicit proximity hypothesis, or through radii measured after that target is supplied. Held-out natural conformations are not arguments of the objectives, and are reserved for evaluation.

Throughout, $R \geq 1$ is the number of represented target residues, the genome support satisfies $\mathcal{R}_G \subseteq \{1, \dots, R\}$, and consequently $0 \leq \Omega(G) := |\mathcal{R}_G|/R \leq 1$. Scales satisfy $0 < z_{\min} \leq z_r \leq z_{\max} < \infty$, and every role extent b_ℓ is positive and finite. Set $b_{\max} = \max_\ell b_\ell$ and write $[v]_+ = \max\{v, 0\}$, $C(v) = \min\{1, \max\{0, v\}\}$. Every candidate used in a score formula has $n = n(x) \geq 1$ finite coordinates and a finite edit history. Candidate records include that history; equal coordinates alone need not imply equal objective values. Every target atom set is finite and nonempty. The contact, clash, and objective formulas below are not applied to empty candidate records. If no adapted candidate is available, the documented genome-level evaluator returns $(0, 0)$; this return does not define any candidate-level expression containing $1/n$. We use $0 \leq \rho, \lambda \leq 1$, $0 < \tau < 1$, and $\varepsilon_q > 0$ throughout. The analysis uses exact real arithmetic: finite-precision implementations must handle threshold endpoints and non-finite values consistently. Distances below are numerical values in angstroms. Throughout every target comparison, the candidate coordinates and their number are fixed. The bounds for q and χ do not transfer the entire genome fitness or the output of a newly run adaptation procedure. All sums over candidate coordinates run from 1 to n, and all probe extrema and sums run from 0 to $K - 1$.

The reference configuration uses

$$\begin{aligned} &(z_{\min}, z_{\max}) = (0.55, 1.75), \qquad (b_{\text{stable}}, b_{\text{flexible}}, b_{\text{forbidden}}) = (3.0, 4.5, 2.9), \\ &(\rho, \tau, \varepsilon_q) = (0.60, 0.75, 10^{-9}), \quad (\lambda, \eta) = (0.70, 0.10), \\ &(N_G, g_{\max}, n_o) = (8, 4, 2). \end{aligned} \tag{32}$$

The profile bounds below hold for every $\rho \in [0, 1]$, and the aggregation inequality holds for every $\lambda \in [0, 1]$. Numerical objective ranges use the stated reference coefficients.

D.1 GEOMETRY OF THE SPATIAL PRIOR

For a covered residue r, let $L_r = b_{\ell_r} z_r > 0$ and $Q_r^\top Q_r = I_3$. Its FlexBox is

$$\mathcal{B}_r(G) = \{y \in \mathbb{R}^3 : \|Q_r^\top (y - c_r)\|_\infty \leq L_r/2\}. \tag{33}$$

Proposition D.1 (Box geometry and support)**.** *Each box is nonempty, compact, and convex, with volume L_r^3, Euclidean diameter $\sqrt{3}\, L_r$, and enclosing radius $\sqrt{3}\, L_r/2$ about c_r. For a common center and frame, boxes nest in the order of the products $b_{\ell_r} z_r$. Moreover, with $\Omega(G) = |\mathcal{R}_G|/R$,*

$$\operatorname{Vol}\Big(\bigcup_{r \in \mathcal{R}_G} \mathcal{B}_r(G) \Big) \leq |\mathcal{R}_G|\, (b_{\max} z_{\max})^3 = R\, \Omega(G)\, (b_{\max} z_{\max})^3. \tag{34}$$

Proof. The box is the image of $[-L_r/2, L_r/2]^3$ under the affine isometry $u \mapsto c_r + Q_r u$, which proves nonemptiness, compactness, convexity, volume, and diameter. Every local coordinate satisfies $\|u\|_2 \leq \sqrt{3}\, \|u\|_\infty \leq \sqrt{3}\, L_r/2$. Comparing ∞-norm bounds proves nesting. Subadditivity of volume and $L_r \leq b_{\max} z_{\max}$ give equation 34. □

Remark D.2 (Editing prior, not a hard pose constraint). Role names alone do not order boxes when scales differ; the products $b_{\ell_r} z_r$ do. The volume bound controls the edit prior and ignores overlaps, so it is generally loose. A repaired candidate coordinate need not lie in any box.

D.2 FIXED-POSE SCORES AND FINITE-PROBE BOUNDS

Fix a genome G and probes $(\widetilde{P}_G^m)_{m=0}^{K-1}$ with $\widetilde{P}_G^0 = P^0$ and integer $K = M + 1 \geq 1$. The reference construction has $K = 7$; distinct indices may share the same geometry. Candidate coordinates remain fixed across probes: the profile measures the response of one candidate, not a separately re-optimized pose for each probe.

For a finite nonempty target atom set $\mathcal{Y}(P)$ in one fixed scoring frame, set $d_i(P) = \min_{y\in\mathcal{Y}(P)} \|p_i(x) - y\|_2$ and

$$\begin{aligned} h_q(d) &= \mathbf{1}_{\{2\leq d\leq 4.5\}} + 0.35\,\mathbf{1}_{\{1.2\leq d<2\}} - 1.5\,\mathbf{1}_{\{d<1.15\}}, \\ h_\chi(d) &= \mathbf{1}_{\{d<1.15\}} + 0.35\,\mathbf{1}_{\{1.15\leq d<1.45\}}, \\ q(x;P) &= C\Big(n^{-1}\sum_i h_q\big(d_i(P)\big)\Big), \qquad \chi(x;P) = n^{-1}\sum_i h_\chi\big(d_i(P)\big). \end{aligned} \tag{35}$$

The indicators inside each kernel are pairwise disjoint, so $h_q \in \{-1.5, 0, 0.35, 1\}$ and $h_\chi \in \{0, 0.35, 1\}$. In particular the clash mid-piece is half-open: $h_\chi(1.45) = 0$, while $h_\chi(d) = 0.35$ for every $d \in [1.15, 1.45)$. Contact clipping is applied after averaging; clash is not clipped, but its average still lies in $[0, 1]$.

Write $q_m = q(x; \widetilde{P}_G^m)$, $\chi_m = \chi(x; \widetilde{P}_G^m)$, and

$$\begin{aligned} &q_{\min} = \min_m q_m, && q_{\max} = \max_m q_m, \quad \bar{q} = K^{-1}\sum_m q_m, \quad \Delta_q = q_{\max} - q_{\min}, \\ &q_{\mathrm{rob}} = \rho q_{\min} + (1-\rho)\bar{q}, && \chi_{\max} = \max_m \chi_m, \\ &\kappa_q = \begin{cases} K^{-1}\sum_m \mathbf{1}_{\{q_m \geq \tau q_0\}}, & q_0 > \varepsilon_q, \\ 0, & q_0 \leq \varepsilon_q. \end{cases} \end{aligned} \tag{36}$$

Proposition D.3 (Finite-probe contact bound)**.** *One has* $0 \leq q_{\min} \leq q_{\mathrm{rob}} \leq \bar{q} \leq q_{\max} \leq 1$ *and* $0 \leq \chi_m \leq \chi_{\max} \leq 1$. *Define*

$$\begin{aligned} L_\kappa &= \begin{cases} \tau q_0, & q_0 > \varepsilon_q \text{ and } \kappa_q = 1, \\ 0, & \text{otherwise}, \end{cases} \\ L_{\mathrm{prof}} &= \max\Big\{0,\ q_0 - \Delta_q,\ L_\kappa,\ q_{\mathrm{rob}} - (1-\rho)\tfrac{K-1}{K}\Delta_q\Big\}. \end{aligned} \tag{37}$$

Then every probe satisfies $q_m \geq q_{\min} \geq L_{\mathrm{prof}}$. *The bound involving* q_{rob} *and* Δ_q *is sharp over arbitrary tuples in* $[0, 1]^K$.

Proof. Clipping and the clash image give the score ranges. The average lies between the minimum and maximum, and q_{rob} is a convex combination of the minimum and average. Since $q_0 \leq q_{\max}$, one has $q_{\min} \geq q_0 - \Delta_q$. When $\kappa_q = 1$ and $q_0 > \varepsilon_q$, every consistency indicator equals one, so $q_{\min} \geq \tau q_0$. At least one score equals $q_{\min}$ and all others are at most $q_{\min} + \Delta_q$, whence

$$\bar{q} \leq q_{\min} + \tfrac{K-1}{K}\Delta_q, \qquad q_{\mathrm{rob}} \leq q_{\min} + (1-\rho)\tfrac{K-1}{K}\Delta_q.$$

Rearrangement gives the remaining bound. Equality holds for one score equal to $q_{\min}$ and the other $K - 1$ equal to $q_{\min} + \Delta_q$, whenever these values lie in $[0, 1]$. □

For $K = 7$ and $\rho = 0.60$, the last bound is $q_{\min} \geq q_{\mathrm{rob}} - (12/35)\Delta_q$. This inequality gives a lower bound on aggregate contact quality shared by all seven probes for the fixed candidate. It connects the optimized robust summary and contact variation to worst-probe contact retention. Transfer to natural target geometries is characterized by Theorem D.13 under its stated neighborhood conditions.

Corollary D.4 (Bound from the robust summary alone)**.** *If only* q_{rob}, K, *and* ρ *are retained, then*

$$q_{\min} \geq L_K := \left[\frac{Kq_{\mathrm{rob}} - (1-\rho)(K-1)}{1 + \rho(K-1)}\right]_+. \tag{38}$$

This bound is sharp over arbitrary real-valued tuples in $[0, 1]^K$. *Sharpness here does not assert realizability of an arbitrary tuple by the specified seven geometric probes or by one molecular structure. For a complete profile it is no stronger than* $q_{\mathrm{rob}} - (1-\rho)(K-1)\Delta_q/K$ *after clipping at zero.*

Proof. At least one score equals $q_{\min}$ and the remaining scores are at most one, so $K\bar{q} \le q_{\min} + K - 1$ and

$$Kq_{\mathrm{rob}} \le \big(1 + \rho(K-1)\big)q_{\min} + (1-\rho)(K-1).$$

The denominator of equation 38 is positive because $\rho \ge 0$ and $K \ge 1$. If the numerator is nonnegative, equality is attained by $(L_K, 1, \dots, 1)$. If it is negative, then $K > 1$ and $\rho < 1$; the tuple $(0, t, \dots, t)$ with $t = Kq_{\mathrm{rob}}/[(1-\rho)(K-1)] \in [0, 1)$ attains equality.

For the comparison, write $a_K = (1-\rho)(K-1)/K$. Then $a_K \in [0, 1)$ and

$$1 - a_K = \frac{1 + \rho(K-1)}{K}, \qquad \frac{q_{\mathrm{rob}} - a_K}{1 - a_K} = \frac{Kq_{\mathrm{rob}} - (1-\rho)(K-1)}{1 + \rho(K-1)},$$

so $L_K = [(q_{\mathrm{rob}} - a_K)/(1 - a_K)]_+$. Proposition D.3 gives $q_{\mathrm{rob}} \le q_{\min} + a_K\Delta_q = q_{\max} - (1 - a_K)\Delta_q \le 1 - (1 - a_K)\Delta_q$, hence

$$\big(q_{\mathrm{rob}} - a_K\Delta_q\big) - \frac{q_{\mathrm{rob}} - a_K}{1 - a_K} = \frac{a_K\big(1 - q_{\mathrm{rob}} - (1 - a_K)\Delta_q\big)}{1 - a_K} \ge 0.$$

In particular, whenever $(q_{\mathrm{rob}} - a_K)/(1 - a_K) \ge 0$, one has $q_{\mathrm{rob}} - a_K\Delta_q \ge (q_{\mathrm{rob}} - a_K)/(1 - a_K) \ge 0$. If $(q_{\mathrm{rob}} - a_K)/(1 - a_K) < 0$, then $L_K = 0$, while the profile term is at least zero after clipping. Positive parts therefore preserve the ordering. □

D.3 Computable bounds under target perturbations

All compared targets are expressed in one fixed scoring frame. Any superposition is chosen by a target-based rule independent of the candidate being scored. All targets in a comparison are placed in this one reference frame before the distances below are computed; separate pairwise best-fit distances are not substituted into these formulas. The atom sets are exactly those used in the nearest-target score. Hausdorff distance ignores atom identities, which is sufficient here because these two score kernels also depend only on nearest distances. For finite nonempty sets,

$$d_{\mathrm{H}}(P, P') = \max\left\{ \max_{y \in \mathcal{Y}(P)} \min_{y' \in \mathcal{Y}(P')} \|y - y'\|_2,\ \max_{y' \in \mathcal{Y}(P')} \min_{y \in \mathcal{Y}(P)} \|y - y'\|_2 \right\}. \tag{39}$$

Lemma D.5 (Nearest-target distance stability)**.** *For every represented coordinate,* $|d_i(P') - d_i(P)| \le d_{\mathrm{H}}(P, P')$.

Proof. Let $y \in \mathcal{Y}(P)$ attain $d_i(P)$ and choose $y' \in \mathcal{Y}(P')$ with $\|y - y'\|_2 \le d_{\mathrm{H}}(P, P')$. The triangle inequality gives $d_i(P') \le d_i(P) + d_{\mathrm{H}}(P, P')$. Exchanging the targets proves the claim. □

For $\varepsilon \ge 0$, set $I_i(P, \varepsilon) = [\max\{0, d_i(P) - \varepsilon\},\ d_i(P) + \varepsilon]$. Let $C_q = C$ and $C_\chi(v) = v$. For $s \in \{q, \chi\}$,

$$L_s(x; P, \varepsilon) = C_s\Big(n^{-1} \sum_i \min_{d \in I_i(P,\varepsilon)} h_s(d)\Big), \tag{40}$$

$$U_s(x; P, \varepsilon) = C_s\Big(n^{-1} \sum_i \max_{d \in I_i(P,\varepsilon)} h_s(d)\Big).$$

Each kernel maps a nonempty interval to a nonempty finite set of values, so its minimum and maximum are attained despite its discontinuities.

Theorem D.6 (Coordinatewise interval enclosure)**.** *If* $d_{\mathrm{H}}(P, P') \le \varepsilon$*, then*

$$L_s(x; P, \varepsilon) \le s(x; P') \le U_s(x; P, \varepsilon), \qquad s \in \{q, \chi\}. \tag{41}$$

The same enclosure holds for the reference score $s(x; P)$*, because* $d_i(P) \in I_i(P, \varepsilon)$*. The endpoints are exact extrema over the product relaxation* $(d'_1, \dots, d'_n) \in \prod_i I_i(P, \varepsilon)$*. Either endpoint may fail to be attainable by a single target geometry inside the Hausdorff neighborhood. The enclosure does not assert attainability of every intermediate score, or simultaneous attainability of contact and clash endpoints.*

Proof. Lemma D.5 places each $d_i(P')$ in $I_i(P,\varepsilon)$. The reference distances already lie in those intervals. Averaging and using that each C_s is nondecreasing yields equation 41 for both P' and P. In the product relaxation, each coordinate can attain its kernel minimum or maximum independently, so the envelope endpoints are attained there. A realizable target couples the nearest-target distances, so attainment need not extend to the geometric neighborhood. □

Lemma D.7 (Kernel oscillations). *Let* $\mathcal{T}_q = \{1.15, 1.20, 2.00, 4.50\}$ *and* $\mathcal{T}_\chi = \{1.15, 1.45\}$*, and define*

$$\omega_s(\varepsilon) = \sup\big\{|h_s(d) - h_s(d')| : d, d' \geq 0,\ |d - d'| \leq \varepsilon\big\}.$$

Then $\omega_s(0) = 0$*, and for* $\varepsilon > 0$

$$\omega_q(\varepsilon) = \begin{cases} 1.50, & 0 < \varepsilon \leq 0.05, \\ 1.85, & 0.05 < \varepsilon \leq 0.85, \\ 2.50, & \varepsilon > 0.85, \end{cases} \qquad \omega_\chi(\varepsilon) = \begin{cases} 0.65, & 0 < \varepsilon \leq 0.30, \\ 1.00, & \varepsilon > 0.30. \end{cases} \tag{42}$$

Proof. The ordered contact piece values are $-1.5, 0, 0.35, 1, 0$. The clash piece values are $1, 0.35, 0$.

For any $\varepsilon > 0$, choose $t = \min\{\varepsilon/2, 0.01\} > 0$. The pair $(1.15 - t, 1.15)$ consists of nonnegative distances separated by at most ε. It gives contact difference 1.50 and clash difference 0.65.

A contact difference greater than 1.50 must use -1.5 together with 0.35 or 1. The former requires $d < 1.15$ and $d' \geq 1.20$, hence $|d - d'| > 0.05$. For $\varepsilon > 0.05$, take $t = \min\{(\varepsilon - 0.05)/2, 0.01\}$. Then $(1.15 - t, 1.20)$ attains difference 1.85 within the budget. The difference 2.50 requires $d < 1.15$ and $d' \geq 2$, hence separation strictly greater than 0.85. For $\varepsilon > 0.85$, take $t = \min\{(\varepsilon - 0.85)/2, 0.01\}$. The pair $(1.15 - t, 2)$ attains 2.50 within the budget. No larger difference occurs among the contact kernel values.

A clash difference greater than 0.65 must be 1. It requires $d < 1.15$ and $d' \geq 1.45$, hence separation strictly greater than 0.30. For $\varepsilon > 0.30$, take $t = \min\{(\varepsilon - 0.30)/2, 0.01\}$. The pair $(1.15 - t, 1.45)$ attains difference 1. These strict separations also establish the three endpoint conventions in equation 42. When $\varepsilon = 0$, the two arguments coincide and both oscillations are zero. □

Proposition D.8 (Centered modulus enclosure). *Define the sensitive fraction*

$$\nu_s(x; P, \varepsilon) = n^{-1}\big|\{i : \operatorname{dist}(d_i(P), \mathcal{T}_s) \leq \varepsilon\}\big| \tag{43}$$

and $\delta_s = \min\{1, \omega_s(\varepsilon)\, \nu_s(x; P, \varepsilon)\}$*. Then*

$$\big[L_s(x; P, \varepsilon), U_s(x; P, \varepsilon)\big] \subseteq \big[\max\{0, s(x; P) - \delta_s\},\ \min\{1, s(x; P) + \delta_s\}\big]. \tag{44}$$

In particular, $\nu_s = 0$ *implies exact score invariance throughout the declared neighborhood.*

Proof. If a coordinate is not counted by ν_s, its interval contains no threshold and its kernel contribution is constant. For every other coordinate, any value on $I_i(P,\varepsilon)$ differs from the value at $d_i(P)$ by at most $\omega_s(\varepsilon)$, because every point of that interval is at most ε from $d_i(P)$. Averaging bounds the change of the unclipped mean by $\omega_s(\varepsilon)\nu_s$. Both C_q and C_χ are 1-Lipschitz, and the resulting scores lie in $[0,1]$. Hence every score arising from the product relaxation—including the envelope endpoints—lies in the centered interval of half-width δ_s about $s(x;P)$. □

Corollary D.9 (Realized change, centered deviation, and envelope width). *Write* $s_0 = s(x;P)$ *and suppress* $(x;P,\varepsilon)$ *in* L_s, U_s*. Define the centered deviation* $D_s^{\mathrm{ctr}} = \max\{s_0 - L_s,\ U_s - s_0\}$*. If* $d_{\mathrm{H}}(P, P') \leq \varepsilon$*, then*

$$|s(x; P') - s_0| \leq D_s^{\mathrm{ctr}} \leq \min\{U_s - L_s,\ \delta_s\} = \min\big\{U_s - L_s,\ \min\{1, \omega_s(\varepsilon)\, \nu_s(x; P, \varepsilon)\}\big\}. \tag{45}$$

Also $D_s^{\mathrm{ctr}} \leq U_s - L_s \leq 2D_s^{\mathrm{ctr}}$*. The full envelope width satisfies*

$$U_s - L_s \leq \min\Big\{1,\ n^{-1}\sum_i (M_i - m_i)\Big\} \leq \min\{1, \omega_s(2\varepsilon)\, \nu_s(x; P, \varepsilon)\}, \tag{46}$$

where $m_i = \min_{I_i} h_s$ *and* $M_i = \max_{I_i} h_s$*. The factor* 2ε *in equation 46 cannot in general be replaced by* ε*. Because* $s_0 \in [L_s, U_s]$*, one always has* $D_s^{\mathrm{ctr}} \leq U_s - L_s$*. The quantities that are incomparable as upper bounds on* $|s(x;P') - s_0|$ *are the envelope width* $U_s - L_s$ *and the modulus* δ_s*: either may be strictly smaller, depending on* (x, P, ε)*.*

Proof. Theorem D.6 places both s_0 and $s(x;P')$ in $[L_s, U_s]$, so the absolute deviation from s_0 is at most D_s^{ctr}, which cannot exceed the width $U_s - L_s$. Proposition D.8 gives $D_s^{\mathrm{ctr}} \le \delta_s$. Both envelope endpoints are attained in the product relaxation, so D_s^{ctr} is the exact maximum absolute deviation from s_0 on that domain.

Since $U_s - L_s = (U_s - s_0) + (s_0 - L_s)$, the width is at most $2D_s^{\mathrm{ctr}}$. Both C_s are nondecreasing and 1-Lipschitz. The attained envelope endpoints lie in $[0,1]$, including for the identity map C_χ, because all clash kernel values lie there. This proves the first comparison in equation 46. Any two points of I_i are at most 2ε apart, so $M_i - m_i \le \omega_s(2\varepsilon)$, and the difference vanishes when the interval contains no threshold. Averaging yields the second comparison.

Why 2ε is necessary. Place one candidate coordinate at the origin and one target atom at distance 1.30, and take $\varepsilon = 0.16$. The clash interval $I = [1.14, 1.46]$ meets all three constancy pieces of h_χ, so $L_\chi = 0$, $U_\chi = 1$, and $U_\chi - L_\chi = 1$, whereas $\omega_\chi(0.16)\nu_\chi = 0.65 = \delta_\chi$. Here $s_0 = 0.35$ and $D_\chi^{\mathrm{ctr}} = 0.65$. Thus $U - L = 1 > \delta = D^{\mathrm{ctr}}$: the modulus is tighter than the envelope width as a bound on $|s' - s_0|$.

A threshold endpoint. With $s = \chi$, $d_1(P) = 1.15$, and $\varepsilon = 0.30$, the closed interval is $[0.85, 1.45]$. Since $h_\chi(1.45) = 0$, its enclosure is $[0,1]$, whereas $\omega_\chi(0.30) = 0.65$. Here too the width and the centered deviation are different quantities.

Envelope tighter than the modulus. Place $d_1(P) = 3.25$ and $\varepsilon = 1.25$ with $s = q$. Then $I_1 = [2.00, 4.50]$, so $h_q \equiv 1$, $U_q - L_q = 0$, while $\omega_q(1.25) = 2.5$ and $\nu_q = 1$, hence $\delta_q = 1$. Thus $U - L = 0 < \delta = 1$: the envelope width is tighter. □

Proposition D.10 (No global Lipschitz bound)**.** *There is no finite constant L such that $|q(x;P') - q(x;P)| \le L\, d_{\mathrm{H}}(P,P')$ for every one-coordinate candidate and every pair of nonempty finite targets. The same statement holds for χ.*

Proof. Place the single coordinate at the origin and all target atoms on the positive first coordinate axis. Singleton targets at distances 4.5 and $4.5 + t$ ($t > 0$) have Hausdorff distance t and contact scores 1 and 0. Singleton targets at distances $1.15 - t$ and 1.15 ($0 < t < 1.15$) give clash difference 0.65. In both cases the ratio of score difference to t is unbounded as $t \downarrow 0$. □

Remark D.11 (Probe displacements and useful radii). With a unit radial direction and a tangential direction that is either zero or a perpendicular unit vector, the flexible amplitude is 1.65. The radial, unit-tangential, and unit-tangential diagonal lengths are 1.65, 1.2375, and $0.9075\sqrt{2}$, respectively. With zero tangential direction, the last two lengths become 0 and 0.9075. The non-flexible fallback amplitude is 0.879. Thus each defined mode displaces each target atom by at most 1.65. Matching each atom to its displaced copy in both directions gives $d_{\mathrm{H}}(P^0, \widetilde{P}_G^m) \le 1.65$. This bound concerns the generated probes and establishes no coverage of natural target conformations.

At a large radius, an enclosure may reduce to the codomain $[0,1]$. An interval enclosure can nevertheless improve on the centered modulus bound. For example, if every $d_i(P) = 2.85$ and $\varepsilon = 1.65$, then every distance interval is $[1.20, 4.50]$, giving $[L_q, U_q] = [0.35, 1]$. Here $\omega_q(\varepsilon)\nu_q = 2.50$, so the centered modulus bound alone gives $[0,1]$.

D.4 Transfer from probe neighborhoods

Definition D.12 (Declared probe neighborhoods)**.** *For a fixed genome and candidate, finite radii $\varepsilon_m \ge 0$ are prescribed. Over the class of finite nonempty target atom sets in the fixed scoring frame, define*

$$\mathcal{U}_m = \{P : d_{\mathrm{H}}(P, \widetilde{P}_G^m) \le \varepsilon_m\}, \qquad L_s^m = L_s(x; \widetilde{P}_G^m, \varepsilon_m), \quad U_s^m = U_s(x; \widetilde{P}_G^m, \varepsilon_m).$$

Theorem D.13 (Union and intersection enclosures)**.** *For the neighborhoods in Definition D.12 and $s \in \{q, \chi\}$:*

1. *if $P^\star \in \bigcup_m \mathcal{U}_m$, then*

$$\min_m L_s^m \le s(x;P^\star) \le \max_m U_s^m; \tag{47}$$

2. *if* $P^\star \in \bigcap_{m\in J} \mathcal{U}_m$ *for* $\emptyset \neq J \subseteq \{0,\dots,K-1\}$*, then*

$$\max_{m\in J} L_s^m \le s(x;P^\star) \le \min_{m\in J} U_s^m. \tag{48}$$

Membership in an unspecified single neighborhood supports only the union statement. If the active index set $J_\star = \{m : P^\star \in \mathcal{U}_m\}$ *is known and nonempty, item 2 with* $J = J_\star$ *is valid and at least as tight as item 1. A union bound restricted to the known active indices is also valid, but is generally weaker than their intersection bound.*

Proof. Theorem D.6 places the score in every interval whose neighborhood contains $P^\star$. A point of one interval lies between the global minimum lower endpoint and the global maximum upper endpoint. A point of every interval indexed by J lies in their intersection. □

Corollary D.14 (Profile form of a union lower bound)**.** *Let* $E_q = \max_m(q_m - L_q^m)$ *and* $E_\chi = \max_m(U_\chi^m - \chi_m)$*. Both are nonnegative. For every* $P^\star \in \bigcup_m \mathcal{U}_m$*,*

$$q(x;P^\star) \ge [L_{\mathrm{prof}} - E_q]_+, \qquad \chi(x;P^\star) \le \min\{1,\, \chi_{\max} + E_\chi\}. \tag{49}$$

The direct bounds in equation 47 are at least as tight.

Proof. Since each radius is nonnegative, its center belongs to its own neighborhood. The enclosure theorem therefore gives $L_q^m \le q_m$ and $\chi_m \le U_\chi^m$, proving nonnegativity. Thus $L_q^m \ge q_m - E_q \ge L_{\mathrm{prof}} - E_q$ and $U_\chi^m \le \chi_m + E_\chi \le \chi_{\max} + E_\chi$. Apply equation 47. □

Remark D.15 (Prospective versus retrospective radii). Prospective bounds prescribe the radii independently of the held-out target; coverage remains a separate hypothesis. Setting $\varepsilon_m = d_{\mathrm{H}}(P^\star, \widetilde{P}_G^m)$ after $P^\star$ is supplied yields an a posteriori intersection enclosure. It is not an input of f_{pres} or f_{rob}. For an unobserved natural target, coverage needs an additional hypothesis or an independently justified model of target variation. For an observed target, neighborhood membership can instead be checked directly.

Theorem D.16 (Combining valid probes before score aggregation)**.** *Let* $\emptyset \neq J \subseteq \{0,\dots,K-1\}$ *index probe neighborhoods that all contain the same target* $P^\star$*. For each candidate coordinate, define*

$$\underline{d}_i^J = \max_{m\in J}[d_i(\widetilde{P}_G^m) - \varepsilon_m]_+, \qquad \overline{d}_i^J = \min_{m\in J}\{d_i(\widetilde{P}_G^m) + \varepsilon_m\}, \tag{50}$$
$$I_i^J = [\underline{d}_i^J, \overline{d}_i^J].$$

Every interval I_i^J *is nonempty. For* $s \in \{q,\chi\}$*, set*

$$L_s^J = C_s\left(\frac{1}{n}\sum_i \min_{d\in I_i^J} h_s(d)\right), \qquad U_s^J = C_s\left(\frac{1}{n}\sum_i \max_{d\in I_i^J} h_s(d)\right). \tag{51}$$

Then

$$\max_{m\in J} L_s^m \le L_s^J \le s(x;P^\star) \le U_s^J \le \min_{m\in J} U_s^m. \tag{52}$$

The two new endpoints are exact over the Cartesian-product relaxation $\prod_i I_i^J$*, but need not be attained by a common target geometry.*

Proof. For each $m \in J$, Lemma D.5 places $d_i(P^\star)$ inside $I_i(\widetilde{P}_G^m, \varepsilon_m)$. It therefore belongs to their intersection, which is exactly I_i^J, so every I_i^J is nonempty. For each coordinate,

$$\min_{d\in I_i^J} h_s(d) \le h_s\big(d_i(P^\star)\big) \le \max_{d\in I_i^J} h_s(d).$$

Averaging and applying the nondecreasing map C_s proves $L_s^J \le s(x;P^\star) \le U_s^J$ directly. This argument does not require the intersected intervals to have a common center or radius. For every $m \in J$, the inclusion $I_i^J \subseteq I_i(\widetilde{P}_G^m, \varepsilon_m)$ gives

$$\min_{I_i^J} h_s \ge \min_{I_i(\widetilde{P}_G^m,\varepsilon_m)} h_s, \qquad \max_{I_i^J} h_s \le \max_{I_i(\widetilde{P}_G^m,\varepsilon_m)} h_s.$$

Average these inequalities and apply the monotone map C_s. Taking the largest per-probe lower bound and smallest per-probe upper bound proves the outside inequalities in equation 52. Independent attainment of the scalar extrema proves exactness on the relaxed domain, as in Theorem D.6. □

This refinement can be strict. Consider two candidate coordinates with probe distance vectors $(2.4, 4.4)$ and $(2.0, 4.0)$. Use radii 0.3 and 0.5, respectively. The separate contact enclosures are $[0.5, 1]$ and $[0.675, 1]$. The intersected coordinate intervals are $[2.1, 2.5]$ and $[4.1, 4.5]$, where both contact contributions equal one. Consequently, the refined contact enclosure is $[1, 1]$. The example is geometrically consistent: on one axis, place candidates at $0, 100$, the two probe atom sets at $\{2.4, 104.4\}$ and $\{2.0, 104.0\}$, and the common target at $\{2.2, 104.2\}$. Its Hausdorff distance from each probe is 0.2. This is an analytical score example, not a claim of chemical validity.

The refinement requires simultaneous membership for every index in J. It cannot replace the union bound when only one unspecified covering probe is known to exist. An empty I_i^J disproves simultaneous membership, whereas nonempty coordinate intervals alone do not prove that any common target geometry exists.

D.5 Objective ranges

Write $\mathcal{S}_G$ and $\mathcal{F}_G$ for the stable and flexible residue sets of G, and $d_r(x) = \min_i \|p_i(x) - c_r\|_2$. For nonempty $\mathcal{S}_G$,

$$a(x; G) = |\mathcal{S}_G|^{-1} \sum_{r \in \mathcal{S}_G} \exp\big(-(d_r(x) - 3.1)^2/4.5\big); \tag{53}$$

otherwise $a = 0$. Hence $a \in [0, 1]$. For $n \geq 2$, the methods appendix sets

$$\begin{aligned} g(x) &= C\Big(1 - 0.16 \sum_i \mathbf{1}_{\{d_i^{\mathrm{self}} < 0.55\}} - 0.03[n - 12]_+\Big), \\ c(x) &= n^{-1} \sum_i \mathbf{1}_{\{\exists j \neq i:\ 0.65 \leq \|p_i - p_j\|_2 \leq 2.70\}}, \end{aligned} \tag{54}$$

with $d_i^{\mathrm{self}} = \min_{j \neq i} \|p_i - p_j\|_2$, and with the conventions $g = c = 1$ for $n = 1$ and $g = c = 0$ for an empty candidate. Thus $g, c \in [0, 1]$. Also $\kappa_q, \Delta_q \in [0, 1]$ by equation 36.

Let $t_{ir} = \|Q_r^\top (p_i - c_r)\|_\infty / (b_{\ell_r} z_r)$ and $w_r^{\mathrm{box}} = \min\{1, \max\{0.1, z_r\}\}$. The methods penalty maps are

$$\begin{aligned} r_{\mathrm{flex}}(x; G) &= \frac{0.85 \sum_{r \in \mathcal{F}_G} [1 - d_r(x)/3]_+ + 0.05 \sum_{r \in \mathcal{F}_G} w_r^{\mathrm{box}} \sum_i \mathbf{1}_{\{t_{ir} \leq 0.55\}}}{\max\{1, |\mathcal{F}_G|\}}, \\ \pi_{\mathrm{forb}}(x; G) &= n^{-1} \sum_i \sum_{\substack{r \in \mathcal{R}_G: \\ \ell_r = \mathrm{forbidden}}} \Big(\mathbf{1}_{\{t_{ir} \leq 0.5\}} + 0.2\, \mathbf{1}_{\{0.5 < t_{ir} < 0.7\}}\Big), \\ e(x) &= \sum_{o \in \mathcal{H}_x} w_o, \qquad \Pi = 0.95 r_{\mathrm{flex}} + 1.25 \pi_{\mathrm{forb}} + 0.70 \chi_{\max} + 0.40 e, \end{aligned} \tag{55}$$

where $\mathcal{H}_x$ is a finite indexed sequence of edit records (repeated operations are counted separately), and every record weight satisfies $0 \leq w_o \leq 0.06$. All summands are nonnegative. The reference candidate objectives are

$$\begin{aligned} f_{\mathrm{pres}} &= 1.50a + q_0 + 0.80 q_{\mathrm{rob}} - 1.40 \chi_{\max}, \\ f_{\mathrm{rob}} &= 0.90 g + 0.80 c + 0.75 \kappa_q - 0.75 \Delta_q - \Pi. \end{aligned} \tag{56}$$

In particular, q_{rob} appears inside f_{pres} by definition: the preservation objective already mixes input-state contact with probe-profile contact.

Proposition D.17 (Range and aggregation bounds)**.** *Every scored candidate with $n(x) \geq 1$ satisfies*

$$-1.40 \leq f_{\mathrm{pres}} \leq 3.30, \qquad -0.75 - \Pi \leq f_{\mathrm{rob}} \leq 2.45.$$

These endpoint inequalities are obtained by independent component extrema and need not be attained jointly by one molecular candidate. For any fixed integers $N \geq 1$ and $H \geq 0$, if $n(x) \leq N$ and the edit history has at most H records, then $\Pi \leq B(N, H, R)$ with

$$B(N, H, R) = 0.95(0.85 + 0.05N) + 1.25R + 0.70 + 0.024H. \tag{57}$$

For a finite nonempty adapted pool $\mathcal{X}_G$, define

$$M_j = \max_{x \in \mathcal{X}_G} f_j(x; G), \qquad \mu_j = |\mathcal{X}_G|^{-1} \sum_{x \in \mathcal{X}_G} f_j(x; G).$$

Whenever the standard aggregation $F_j = \lambda M_j + (1-\lambda)\mu_j - \eta_j \Omega(G)$ is used,

$$\mu_j - \eta_j \Omega(G) \le F_j \le M_j - \eta_j \Omega(G). \tag{58}$$

If every $x \in \mathcal{X}_G$ *satisfies $n(x) \le N$ and $|\mathcal{H}_x| \le H$, then, with $\eta_{\mathrm{pres}} = 0$ and $\eta_{\mathrm{rob}} = 0.10$,*

$$-1.40 \le F_{\mathrm{pres}} \le 3.30, \qquad -0.85 - B(N, H, R) \le F_{\mathrm{rob}} \le 2.45.$$

Proof. Substitute the component ranges into equation 56: $a, q_0, q_{\mathrm{rob}}, g, c, \kappa_q, \Delta_q, \chi_{\max} \in [0, 1]$ and $\Pi \ge 0$.

For the budget, if $\mathcal{F}_G = \emptyset$ then $r_{\mathrm{flex}} = 0$. Otherwise each proximity term satisfies $[1 - d_r/3]_+ \le 1$ and each $w_r^{\mathrm{box}} \le 1$, while $\sum_i \mathbf{1}_{\{t_{ir} \le 0.55\}} \le n \le N$. Hence the numerator of r_{flex} is at most $0.85|\mathcal{F}_G| + 0.05|\mathcal{F}_G|N$, and dividing by $|\mathcal{F}_G|$ yields $r_{\mathrm{flex}} \le 0.85 + 0.05N$. For each coordinate and each forbidden residue, the two occupancy indicators in equation 55 are disjoint, so their contribution is at most one; summing over at most R forbidden residues and averaging over n coordinates gives $\pi_{\mathrm{forb}} \le R$. Finally $\chi_{\max} \le 1$ and $e \le 0.06H$, proving equation 57.

Since $\mu_j \le M_j$, their convex combination lies between them; subtracting $\eta_j \Omega(G)$ gives equation 58. The uniform assumptions on the pool give $f_{\mathrm{rob}}(x; G) \ge -0.75 - B(N, H, R)$ for every candidate. The lower bound on F_{rob} follows together with $\eta_{\mathrm{rob}} \Omega(G) \le 0.10$. □

Remark D.18 (Coordinate and history bounds are separate assumptions). The integers N and H in Proposition D.17 bound coordinates per candidate and records per edit history. They are not the inner or outer population sizes N_x and N_G. In particular, the maximum inner population of twelve does not imply $n(x) \le 12$. A fixed number of adaptation rounds alone also does not specify N or H without bounds on initialization and the records and coordinates produced by each invoked operation. For a realized finite nonempty pool of finite candidate records, one may take $N = \max_x n(x)$ and $H = \max_x |\mathcal{H}_x|$; this gives an a posteriori bound, not a uniform bound for every possible input.

Remark D.19 (Equal-aggregate support comparison). If two standard-aggregation evaluations share identical M_j and μ_j for both objectives, then $F_{\mathrm{rob}}(G) - F_{\mathrm{rob}}(G') = 0.10\big(\Omega(G') - \Omega(G)\big)$. In general, changing support also changes scores and probes, so this identity is not a general preference for smaller support. The coordinatewise maxima M_{pres} and M_{rob} need not be attained by the same candidate. The documented zero-return branch that assigns $(F_{\mathrm{pres}}, F_{\mathrm{rob}}) = (0, 0)$ when boxes or adapted candidates are missing satisfies the numerical ranges but is excluded from equation 58; it is not a feasibility penalty.

D.6 Selection: what is and is not preserved

Encode each genome by residue roles, with an additional symbol for absence. The residue-role distance D_G is the Hamming distance between these encodings: it is a metric on encodings and a pseudometric on full genomes (scales are ignored).

For maximized objective pairs, record u dominates v when both coordinates of u are at least those of v, with at least one strict inequality. Nondominated fronts are formed by repeatedly removing the current nondominated records. The outer selector visits increasing front rank, then decreasing crowding, followed by objective sum and identifier. The counterexamples below have equal first-front sums and specify the identifier order, so they do not depend on the direction of objective-sum tie-breaking. Crowding conventions are assumed to produce a defined total traversal order, including fronts with a constant objective coordinate. The counterexample uses the standard boundary-crowding convention: both members of a two-record front have infinite crowding. It defers a record whose role distance to an already accepted record is at most two, then backfills deferred records.

Proposition D.20 (Outer survivor facts and a dominance failure)**.** *For a finite input population U of uniquely identified records with finite real objective pairs, and integer capacity $N_G \ge 1$, the selected set is a subset of U of size $\min\{N_G, |U|\}$. For nonempty U, at least one selected record lies on the first front. Distinct records accepted before backfill have role distance at least three; this*

separation need not survive backfill. For every integer capacity $N_G = N \geq 2$ and every target size $R \geq 3(N-1)+2$, there exists an input on which the selector retains a dominated record while excluding one of its dominators. Thus dominance preservation is not guaranteed uniformly over target sizes and scored input populations. The displayed condition on R is sufficient for the construction below; it is not asserted to be necessary. At capacity one, the sole survivor, when present, lies on the first front.

Proof. Until capacity is reached, each visited record is accepted or deferred. If the first pass finishes below capacity, every input record has been accepted or deferred, so backfill fills capacity or exhausts the remainder. The first visited record of nonempty U lies on the first front and is accepted. Each first-pass acceptance has integer distance greater than two from all prior acceptances. To see why backfill can violate this separation, take capacity two and two records whose role encodings differ at exactly one residue. Whichever record is visited first is accepted; the other is deferred and then restored by backfill. Both survive with distance one.

Fix integers $N \geq 2$ and $R \geq R_0 := 3(N-1)+2$. For an operator-level counterexample at capacity $N_G = N$, take $N+1$ records with objective pairs

$$C = (2, 0), \qquad A = (1, 1), \qquad B_j = (t_j, 1 - t_j), \quad t_j = \tfrac{1}{4} + \tfrac{j}{2N}, \quad 1 \leq j < N.$$

The first front is $\{C, A\}$; both have boundary crowding and sum two. Let identifiers place C first. The B_j are pairwise nondominated; A dominates each of them, whereas C dominates none. Thus they form the second front.

On the first R_0 residue positions, let the role encoding of C be stable everywhere except at position R_0, which is flexible. Let A change only the first position to flexible. Partition the first $3(N-1)$ positions into disjoint triples J_j, and let B_j change exactly the triple J_j to flexible. At every remaining position $R_0 + 1, \ldots, R$, give all records the same stable role. Assign every record any common admissible scale vector. Every record has both a stable and a flexible role. Then

$$D_G(C, A) = 1, \qquad D_G(C, B_j) = 3, \qquad D_G(B_j, B_k) = 6 \quad (j \neq k).$$

The selector accepts C, defers A, and accepts all B_j, in any within-front order, filling capacity before backfill. It therefore excludes a dominator of every selected B_j. The reference capacity $N_G = 8$ is covered whenever $R \geq 23$. This disproves an unconditional guarantee for the selection operator on arbitrary scored records; it does not assert molecular realizability of the specified pairs and encodings. □

Remark D.21 (Why the target-size qualification matters). For $R \leq 2$, any two role encodings have $D_G \leq 2$. For a nonempty input and positive capacity, the first record is accepted; all subsequent records reached in the first pass are deferred. The final selected set is therefore the first $\min\{N_G, |U|\}$ records of the original traversal; no backfill is needed when the first acceptance already fills capacity. Every strict dominator of a record has a lower front rank and occurs earlier in that traversal. Consequently, on such small targets the selector cannot retain a record while excluding one of its dominators. This observation does not guarantee retention of all nondominated records when their number exceeds capacity. The negative statement in Proposition D.20 is therefore a statement about specified sufficiently large target sizes, not about every fixed $R \geq 1$.

Remark D.22 (Inner selection and main-text wording). For inner capacity $N_x = N \geq 2$, use the same objective pairs and centroids $c_C = (0, 0, 0)$, $c_A = (0.5, 0, 0)$, and $c_{B_j} = (2j, 0, 0)$. With the 1 Å centroid niche, C is accepted, A is deferred, and all B_j are accepted. This includes every capacity allowed by the reference inner rule $N_x = \max\{4, \min\{12, 2|\mathcal{X}_0|\}\}$. Here too the objective pairs and centroids are inputs to the selection operator; their simultaneous realization by the molecular evaluator is not asserted. Nondominated rank is a selection criterion; unconditional preservation of nondominated records is not a property of the documented rule. The documented rule therefore uses nondominated sorting with crowding and niching, without guaranteeing retention of every nondominated record or of every dominator of a retained record.

Remark D.23 (Evaluation count). If initialization evaluates N_G genomes and each of $g_{\max}$ completed generations evaluates n_o offspring per parent in a population of size N_G, there are $N_G(1 + g_{\max} n_o)$ scheduled genome evaluations. The reference values $(N_G, g_{\max}, n_o) = (8, 4, 2)$ give 72. This count assumes the stated schedule is completed without early termination or extra evaluations. It counts repeated genomes, not distinct genomes, low-level score calls, or wall-clock time. Finite

scheduling implies termination only when each invoked subroutine also terminates and all required scoring and feedback operations are defined. For an empty parent pool, the reference implementation uses infinite nearest-coordinate distances and zero edit-use counts, yielding zero feedback without evaluating a minimum over an empty set.

D.7 Established properties and their role in adaptation

The analysis establishes geometric control of the editing prior, deterministic finite-probe score bounds, and target-transfer enclosures under the stated Hausdorff-neighborhood conditions. Proposition D.1 bounds the spatial extent of the FlexBox prior. Proposition D.3 converts contact retention and variation into a lower bound shared by all evaluated probes. For $K = 7$ and $\rho = 0.60$, this includes

$$q_{\min} \geq q_{\mathrm{rob}} - \frac{12}{35}\Delta_q.$$

Theorem D.13 extends the geometric score enclosures to target conformations satisfying its explicit neighborhood conditions.

These results complement the molecular structural safeguards in Appendix A.3. Permitted rigid and annotated torsional transformations preserve the supplied molecular graph and covalent bond lengths, while the acceptance gate enforces coordinate and bond-length consistency before survivor selection. The geometric descriptors quantify spatial support and target contacts, and the category-specific validation specification addresses chemical and physical structure checks. Together, these components give an explicit account of structural preservation, perturbation scoring, and conditional cross-state score control within the implemented adaptation procedure.