Arc Science
Technical Report II

Antigen Conditioning, Guided Sequential Monte Carlo Sampling, Maturation, and Mathematical Supplement

Part II of the Omniverso Specification (Conditioned Dynamics and Supplementary Section S1)

Danil Kotelnikov

Abstract. Target-conditioned antibody generation must preserve coordinate symmetries while steering sampling towards affinity, structure, and developability goals. This volume specifies antigen conditioning, Interface-JEPA scoring, guided sequential Monte Carlo sampling, and affinity maturation through unbalanced optimal transport. Supplementary Section S1 gives the full symbol dictionary and all 62 mathematical definitions, assumptions, propositions, lemmas, theorems, corollaries, and proofs. Components beyond the unconditional base model are specified as a design, with the gates that test it and the proofs of its mathematical properties.

Contents

1 Antigen conditioning and Interface-JEPA **3**

1.1 Interface-JEPA 6

2 Guidance and sequential Monte Carlo **7**

2.1 Sequential Monte Carlo 10

3 Maturation and editing **11**

4 Mathematical supplement **14**

4.1 Notation and conventions 15

4.2 Geometry and gauges 16

4.3 Scores, diffusion reversal and flow likelihood 19

4.4 Transport and particle limits 23

4.5 Decision statistics and calibration 30

4.6 Representation floor and predictive information 35

4.7 Decoder pullback metric and common descent 36

4.8 Numerical error budgets 39

4.9 Sensitivity and identifiability 42

4.10 Standing of the results 43

5 Scope, Ethics, Licence, and Declarations **43**

5.1 Scope, Dual Use, and Licensing 43

5.2 Declarations 44

1. Antigen conditioning and Interface-JEPA

The full system adds an antigen channel and a predictive critic, Interface-JEPA, to the base model. The base model takes no antigen, hotspot or format token as input. Where a value comes from a published configuration it serves as a reference value for a model trained from scratch, and no released weight or code is loaded. Figure 2 shows log-domain unbalanced optimal transport on PDB 4KRL beside the Interface-JEPA architecture and eigenvalue floor (panels a and b).

Conditioning contract. The antigen channel tells the generator where an antibody may touch the antigen and which chemistry the touching residues offer. Let A be the antigen structure, E the requested epitope (a set of antigen residues), F the fixed antibody context and X the generated antibody state. Let C_{pred} be a predicted soft contact plan, with weights in $[0, 1]$ between movable antibody residues and epitope residues (loop residues with a fixed framework, the whole variable domain with a free one). Let $\chi_E(A)$ be the chemical features of the epitope residues. The model generates through the plan, in the context-anchored coordinate profile of Technical Report I (Section "State and geometry"). Only the movable antibody residues are transported, and the rest of the antibody and the antigen are conditioning data:

$$p_\theta(X \mid A, E, F) = \int p_\theta\big(X \mid C_{\text{pred}},\, C_{\text{pred}} \odot \chi_E(A),\, F\big)\, p_\theta(C_{\text{pred}} \mid A, E, F)\, \mathrm{d}C_{\text{pred}}.$$

The symbol $\odot$ stands for a plan-gated message, $g_i = \sum_{j \in E} C_{\text{pred},ij}\, \chi_j(A)$ for antibody position i. A contact that the plan proposes passes the charge, polarity and hydrogen-bonding geometry of the epitope residue to the decoder. A global antigen embedding without this gate is excluded, because the identity of the target could then bypass the plan.

Four antigen schemas are specified: protein epitope, peptide epitope, glycoprotein epitope (usable after gate F-GLY passes) and small molecule (which needs an atom-and-bond schema and a ligand-bearing validation set). An epitope record carries one provenance type: experimental contact, experimental mapping, computational pose, predicted patch or user-declared patch. The types stay separate in training and evaluation. A report finds that baseline CDR generators return nearly identical CDRs whatever the target [1], so gate F-CB compares generation under the true antigen with generation under a decoy, with the antigen removed and with a renumbered copy. A separate complex predictor scores each output against the true antigen.

Residue encoder and cross-attention. Each epitope residue becomes one token c_j that a rigid motion of the antigen leaves unchanged. The token omits deposited residue numbers, which shift under harmless renumbering and can encode target identity. It is built from amino-acid identity, chain and signed within-chain offset. For each of the atom slots N, C, O and C-beta it adds a presence mask, a radial-basis encoding of the distance to C-alpha (in units of the coordinate scale ℓ_0) and the unit direction to C-alpha in the local residue frame. A masked sum over the slots and a linear map give c_j. The local frame uses N, C-alpha and C, with fallbacks at peptide termini and for glycine, and flags for the fallbacks are embedded in the token.

Antibody residues read the tokens by cross-attention in an epitope-centred frame whose origin is the mean C-alpha position of E. Let $\bar{x}_i$ and r_j be the scaled coordinates of antibody residue i and epitope residue j, q_i a query, k_j a key, d_h the head width and ϵ_C a small constant. The weight is

$$\alpha_{ij} = \mathrm{softmax}_{j \in E}\Big(\tfrac{q_i^\top k_j}{\sqrt{d_h}} + b_{\mathrm{geo}}(\|\bar{x}_i - r_j\|_2) + \log(C_{\mathrm{pred},ij} + \epsilon_C) \Big).$$

The first term compares content, the second is a bias that depends on distance, and the third lets the plan raise or lower a pair. Pair distances are unchanged by a joint rotation and translation, and gate F-CB checks this together with invariance to renumbering. The geometric bias is ablated by master-clock interval (Technical Report I (Section "Generative model")), because it may carry little signal near noise.

Contact-plan objective. A solver may refine the learned output C_{pred}, starting from $\Gamma^{(0)}$ and returning an accepted plan $\Gamma^\star$. The plan $\Gamma^\star$ minimises an unbalanced transport objective. Let μ_R and μ_E be normalised exposure weights on movable antibody positions and epitope residues, and let $C = C^{\mathrm{geom}} + \beta_{\mathrm{chem}} C^{\mathrm{chem}}$ be the dimensionless ground cost, with a geometric and a chemical term. Then

$$\Gamma^\star = \arg\min_{\Gamma \geq 0} \langle \Gamma, C \rangle + \rho_{\mathrm{OT}} \,\mathrm{KL}(\Gamma \mathbf{1} \mid \mu_R) + \rho_{\mathrm{OT}} \,\mathrm{KL}(\Gamma^\top \mathbf{1} \mid \mu_E) + \varepsilon_{\mathrm{OT}} \,\mathrm{KL}(\Gamma \mid \mu_R \otimes \mu_E).$$

The first term charges transport cost. The two middle terms charge mass created or destroyed at the margins, so part of a declared patch may stay unmatched. The last term smooths the plan and makes the minimiser unique. KL is the generalised divergence for positive vectors of unequal mass. The solver works in the log domain [2, 3]. For this unbalanced family, Sinkhorn divergences remove the entropic bias when two measures are compared [4].

For $\varepsilon_{\mathrm{OT}} > 0$, $\rho_{\mathrm{OT}} > 0$, finite C and strictly positive marginals, the minimiser exists and is unique. The smoothing changes the optimal value by at most $\varepsilon_{\mathrm{OT}} \,\mathrm{KL}(\Gamma_0^\star \mid \mu_R \otimes \mu_E)$, where $\Gamma_0^\star$ minimises the objective at $\varepsilon_{\mathrm{OT}} = 0$ (Theorem 4.2 and Proposition 4.5 of Supplementary Section S1).

Fidelity controls. Two training controls address a shortcut in which the generator copies framework context and ignores the antigen. Framework dropout zeroes the input features of a framework residue with probability p_{drop}, except at positions that the format profile protects. For a VHH, the protected set may include the FR2 positions that formerly met the light chain. Framework masks use the retained ANARCII numbering map described in Technical Report I (Section "Numbering, regions and the CDR3 prior"). The probability is chosen by a sweep that includes 0.3.

Same-antigen patch contrast adds a retrieval loss. It asks the model to tell the requested patch from non-overlapping patches of matched area and accessibility on the same antigen:

$$\mathcal{L}_{\mathrm{patch}} = -\log \frac{\exp(r^\top e_+ / \tau)}{\sum_{k=1}^{K} \exp(r^\top e_k / \tau)}.$$

Here r summarises the continuous decoder logits, e_k embeds patch k among K patches that include the requested patch e_+, and τ is a temperature. Gate F-CYC reports retrieval accuracy by antigen-fold cluster, patch area, accessibility, CDR-H3 length and format, and the loss is active once the gate accepts.

The six design inputs. Six design inputs act through generation, steering and selection: the antigen, its topology, its physicochemistry, its surroundings, the shape of the paratope and the regime of the self-proteome panel. The first four enter the generator as one antigen description. The antigen is given as atom37 coordinates, amino-acid identity and a hotspot indicator, as in the published recipe [5]. Topology enters through pair features and the frame features of the

residue token, and physicochemistry through amino-acid identity, atom37 features and the plan-gated chemistry. The surroundings are the antigen residues inside the crop, and hotspot tokens mark the epitope. The crop holds at most 500 residues in total and at least 50 antigen residues, with candidates within 15 Å of the antibody taken in contiguous stretches [5]. For each example, hotspots are drawn uniformly from 3 to 12 interface residues (heavy-atom contact within 4.5 Å), or all of them with probability 0.2.

The paratope shape is set by the fixed framework clamped as context, or by a free framework, together with a CDR length prior. The prior comes from the empirical nanobody distribution in SAbDab by IMGT numbering, from the 5th to the 95th percentile with add-one smoothing [6]. The self-proteome regime is the declared panel. It acts through selection and through negative potentials in steering, and the generator has no tensor for it. Global translation noise with a spread of 0.2 nm is added to the antibody coordinates in training [5]. This equals the spread of the shared translation in the base-model noise (0.2 nm, 2 Å). A format token selects the format profile. VHH de novo is the first admitted format, and the paired Fv format follows once its gates pass.

Epitope registration. Registration measures whether a decoded pose occupies the requested patch. The contacted antigen residues $\mathcal{C}$ are those with a heavy atom within $d_c = 4.5$ Å of a heavy atom of the full antibody variable domain R_V, framework included. Precision, coverage and their harmonic mean are

$$P_E = \frac{|\mathcal{C} \cap E|}{|\mathcal{C}|}, \qquad R_E = \frac{|\mathcal{C} \cap E|}{|E|}, \qquad \text{EpiF1} = \frac{2P_E R_E}{P_E + R_E},$$

and each is zero when its denominator is zero. The scope is the full domain because framework residues often contact antigen in single-domain antibodies. In 345 non-redundant single-domain complexes, IMGT position 66 lies in 50.4 percent of interaction-defined paratopes, and positions 52, 55 and 42 in 31.6, 27.2 and 24.1 percent [7]. A loop-only footprint can therefore hide a miss. Take an epitope of ten residues that loop contacts cover completely, with framework contacts on ten further antigen residues outside it. The loop-only footprint scores $P_E = R_E = \text{EpiF1} = 1$, and the full-domain footprint scores $P_E = 0.5$ and $\text{EpiF1} = 2/3$. The loop-only value is a diagnostic, and the full-domain value is the admission criterion.

Training labels ($Y_{ij} = 1$ when antibody residue i and antigen residue j lie within d_c), the contact target of Interface-JEPA and gate F-EPI share one residue scope, one atom mask and one cutoff. The metrics describe a structure proposal and become native-pose metrics when E and the reference geometry come from an experimental complex. Gate F-EPI reports each provenance stratum separately.

Glycan-aware accessibility. A glycan can hide an epitope. A declared set $\mathcal{H}$ of glycosylation scenarios fixes the protein conformer, occupied sites, glycoforms, glycan library and a seed. The accessibility $\sigma_j^{(h)}$ of epitope residue j in scenario h is the fraction of grafted glycan conformers through which a probe still reaches the residue. Grafting follows GlycoSHIELD [8]. The envelope over scenarios is $\underline{\sigma}_j = \min_{h \in \mathcal{H}} \sigma_j^{(h)}$ and $\overline{\sigma}_j = \max_{h \in \mathcal{H}} \sigma_j^{(h)}$.

The antigen token carries a central value, the interval width and a support mask, and the gate reads the lower envelope over required residues. Scenarios on both sides of a floor leave the accessibility unresolved. Until gate F-GLY accepts, the feature is descriptive. The scenarios cover antigen glycosylation only, so a binder with an N-glycan sequon is unresolved for glycan effects (Technical Report III (Section “Liability, immunogenicity and glycan channels”)).

1.1. Interface-JEPA

Interface-JEPA (joint-embedding predictive architecture) is a predictive representation layer for antibody-antigen interfaces. It reconstructs masked interface targets in latent space and exposes four typed heads. It acts as critic and conditioner and generates no molecules. Following I-JEPA [9], an online context encoder and a predictor see the visible context. A target encoder, an exponential moving average of the online encoder, sees the masked native target and passes no gradient. A structure-aware joint-embedding predictive architecture has been combined with diffusion to reconstruct antibody CDR loops [10].

Targets. The contact target holds, for antibody residue i and antigen residue j, the contact indicator y_{ij}, the clipped minimum distance over ℓ_0, a distance-bin vector, orientation features and the exposure of each residue. Here ℓ_0 is the length scale of the full system, which differs from the base-model factor of 0.1 nm per ångström. The set V_B holds the antibody residues of the labelled complex, taken over the full variable domain R_V, loops and framework, and V_A holds its antigen residues. The target has shape $|V_B| \times |V_A| \times (4 + B_d + B_o)$ for B_d distance bins and B_o orientation channels. A complex of 120 antibody and 180 antigen residues gives 21,600 pairs. The complex target adds both residue tensors and the relative pose after antigen-frame anchoring.

Objective. The target parameters $\bar{\theta}$ follow the online parameters θ by $\bar{\theta}_{k+1} = \mu_k \bar{\theta}_k + (1 - \mu_k)\theta_k$, where k counts successful optimiser updates. A step skipped for a non-finite gradient advances neither k nor $\bar{\theta}$, and all ranks share that decision. The target starts as a copy of the online network, projector included. The momentum μ_k rises linearly from 0.996 to 1.0 over the training steps, the schedule reported for I-JEPA [9]. For masked units $r \in \mathcal{M}$ with predictions $\hat{h}_r$ and target latents h_r, the main loss is

$$\mathcal{L}_{\text{pred}} = \frac{1}{|\mathcal{M}|} \sum_{r \in \mathcal{M}} \left\| \operatorname{norm}(\hat{h}_r) - \operatorname{stopgrad}(\operatorname{norm}(h_r)) \right\|_2^2,$$

where $\operatorname{norm}(x) = x / \max(\|x\|_2, \epsilon_n)$ with $\epsilon_n = 10^{-6}$. Anchor terms keep the information that the heads need: a clipped binary cross-entropy on contacts, a Huber loss on distances and a residue term, with weights λ_c, λ_d and λ_r. The clip uses $\epsilon_b = 10^{-6}$, because the excess-risk bound below needs a bounded loss. The full objective is $\mathcal{L}_{\text{JEPA}} = \mathcal{L}_{\text{pred}} + \mathcal{L}_{\text{anchor}} + \lambda_v \mathcal{L}_{\text{var}} + \lambda_{\text{cov}} \mathcal{L}_{\text{cov}}$, with the last two terms defined below.

Heads. The four heads are contact, complex, comparative maturation and generic polyspecificity. The contact head predicts the contact indicator, distance bins and orientation, and is calibrated pairwise by antigen cluster and label provenance. The complex head scores a proposed complex and is judged by ranking within an antigen and by cognate discrimination. The comparative maturation head orders two variants from one assay context by $\Pr(a \succ b) = \sigma(r_\eta(a) - r_\eta(b))$, with logistic σ and scalar score r_η [11]. The generic polyspecificity head predicts the assay response to a polyspecificity reagent, and an isotonic regression fitted after freezing calibrates it. Binding probability and affinity belong to separate assay heads outside this layer. Each head has its own target schema, calibration cohort and gate.

Data are split into P_{hist} for training and selection, P_{cal} for calibration, and P_{gen}, an untouched cohort for a frozen generator. The calibration population is divided into a map-fitting subset and a risk-calibration subset (Technical Report III (Section "Thresholds and conformal risk control")). A head is admitted when its gates hold. Gates F-J1a, F-J1b and F-J1c compare one cognate antigen with ten nested decoys on at least 200 independent designs, and all three must hold.

Anti-collapse terms. A latent predictor can collapse to a constant. Two terms act on a batch of B raw projection outputs $z \in \mathbb{R}^d$ in FP32, before normalisation, because the trace of normalised vectors is at most $B/(B-1)$ and cannot reach the covariance margin. The variance term $\mathcal{L}_{\text{var}}$ is a hinge that pushes the standard deviation of each dimension, computed with a small constant δ under the square root, to at least τ. The covariance term $\mathcal{L}_{\text{cov}}$ penalises the squared off-diagonal entries of the batch covariance. Both follow VICReg [12], whose target standard deviation is 1.

Two properties are stated for these terms. First (Theorem 6.1 of Supplementary Section S1), if $\mathcal{L}_{\text{var}} \le \eta$, $\mathcal{L}_{\text{cov}} \le \kappa$ and $B \ge d+1$, the smallest eigenvalue of the batch covariance $\hat{\Sigma}$ satisfies $\lambda_{\min}(\hat{\Sigma}) \ge (\tau - d\eta)_+^2 - \delta - \sqrt{(d-1)\kappa}$, where $(x)_+ = \max(x, 0)$. The floor constrains the minimum per-dimension variance, so gate F-J3 reads the minimum. Second (Theorem 6.3), for a loss bounded by M, the optimal risk of a predictor of a label Y that sees only the representation h of an input X exceeds the optimal risk given X by at most $M\sqrt{2\, I(X; Y \mid h)}$, with I the conditional mutual information. A small prediction loss does not bound this information (Example 6.4), so the anchor terms provide the empirical handle on it.

The design values are in Table 1. The weights follow the VICReg defaults of 25, 25 and 1 once the prediction weight is set to 1. The vector counts equal $1 + (d-1)^2/0.80982$ at $\tau = 1$, $\delta = 10^{-4}$ and margin 0.1. At $d = 32$ the gate requires $(1 - 32\eta)_+^2 - 10^{-4} - \sqrt{31\kappa} \ge 0.1$. The two ceilings in Table 1 are one-at-a-time limits, each with the other penalty at zero. A pair of values is checked against the joint requirement. Gate F-J3 evaluates the bounds on a fixed random projection of the head input to 32 dimensions, on at least 200 untouched complexes, with one vector per interface residue, equal weight per complex and a cluster bootstrap over complexes and antigen families.

Table 1: Design values of Interface-JEPA: widths in numbers of channels, loss weights and variance floor dimensionless, vector counts in numbers of independent unit-variance Gaussian vectors, ceilings dimensionless. The anchor weights are chosen on P_{hist}.

Item	Value
Representation width read by the heads	256
Width d of the regularised projection	32
Variance floor τ; constant δ	1; 10^{-4}
Weights of $\mathcal{L}_{\text{var}}$ and $\mathcal{L}_{\text{cov}}$ relative to $\mathcal{L}_{\text{pred}}$	1 and 0.04
Anchor weights $\lambda_c, \lambda_d, \lambda_r$	each from 0.25, 0.5, 1, 2
Momentum schedule	0.996 to 1.0, linear
Vectors needed at $d = 16, 32, 64, 256$	279; 1,188; 4,903; 80,297
One-at-a-time ceilings at $d = 32$, margin 0.1	$\eta^{\max} = (1 - \sqrt{0.1001})/32 = 0.02136$; $\kappa^{\max} = 0.8999^2/31 = 0.02612$

2. Guidance and sequential Monte Carlo

The full system steers the sampler towards a declared utility in two ways. Guidance moves each trajectory along a preconditioned utility direction, and sequential Monte Carlo (SMC) weights and resamples a population of trajectories. The base model includes neither guidance nor a generation sampler. Figure 3 illustrates Pareto MGDA descent in the pullback metric on PDB 4KRL (panel a) alongside Feynman–Kac particle steering and the Lean 4 formal proof ledger (panel b).

Stochastic interpolants and the score. Steering needs a stochastic sampler and a score. A stochastic interpolant adds a Gaussian bridge term to a deterministic path from a base state $\bar{q}_0$ to a data state $\bar{q}_1$:

$$\bar{q}_u = \mathcal{I}(u, \bar{q}_0, \bar{q}_1) + \gamma(u)\,\zeta, \qquad \zeta \sim \mathcal{N}(0, I_D).$$

Here $\mathcal{I}$ is the deterministic interpolation, $\gamma(u)$ is the noise width, zero at both ends and positive between them, and D is the dimension of the packed state of Technical Report I (Section "State and geometry"). The score of the density p_u is the average hidden noise divided by the width, with its sign reversed:

$$\nabla_q \log p_u(q) = -\frac{1}{\gamma(u)}\,\mathbb{E}[\zeta \mid \bar{q}_u = q]\,, \qquad 0 < u < 1.$$

For any diffusion coefficient $\epsilon(u) \geq 0$, adding $\epsilon\,\nabla \log p_u$ to the velocity and injecting noise $\sqrt{2\epsilon}\,\mathrm{d}W_u$ keeps the same one-time marginals [13]. Velocity and score must belong to the same interpolant. Under independent pairing of base and data states the score follows from the velocity heads, so the design needs no separate score head. The published La-Proteina sampler integrates this equation by the Euler-Maruyama method in 400 steps, with a noise scale of 0.1 for both blocks [14]. The design takes these values as starting settings and registers one bridge-noise profile and one diffusion profile per block, because one scalar does not serve both clocks.

Clean-endpoint utilities. A property model needs a nearly finished molecule, and a state in mid-path is far from one. Every utility therefore acts on an estimate of the clean endpoint. For the linear block paths of the base model the estimate adds the rest of the velocity:

$$\widehat{x}_1 = x_u + [1 - t_x(u)]\,v_\theta^x, \qquad \widehat{z}_1 = z_u + [1 - t_z(u)]\,v_\theta^z.$$

Here x_u and z_u are the coordinate and latent blocks at master-clock progress u, $t_x(u)$ and $t_z(u)$ are the schedules of Technical Report I (Section "Block loss and independent block clocks"), and v_θ^x, v_θ^z are the velocity heads. A frozen decoder maps $\widehat{q}_1 = (\widehat{x}_1, \widehat{z}_1)$ to soft decoded features $h_\phi(\widehat{q}_1, c_{\mathrm{gen}})$, and a utility head $\widetilde{U}_j$ scores them:

$$U_j = \widetilde{U}_j\big(h_\phi(\widehat{q}_1, c_{\mathrm{gen}}),\, c_{\mathrm{guide}}\big).$$

The condition c_{gen} is the generation condition, and c_{guide} names the admitted heads, gains, negative-design panel members, feature weights and trust limits. Each head is trained on such estimates at the corruption times of the generator and takes the time as input, so its gradient exists at intermediate states. The gradient is taken with the denoiser output held fixed, so a step needs no backward pass through the denoiser.

Each head declares an endpoint type, for example a contact prediction, a pose score or a generic polyspecificity response. Scores combine only within one type, unit, assay mapping and target condition. A named off-target needs its own calibrated pair oracle, and without one the term is unavailable. Hard chemical, scaffold and assembly rules stay outside the optimiser as projections or rejections. The pose and polyspecificity heads of Interface-JEPA (Section 1.1) are candidate utilities. The guided field adds a preconditioned utility direction to the learned master-time field f_θ:

$$\frac{\mathrm{d}\bar{q}}{\mathrm{d}u} = f_\theta(\bar{q}, u, c_{\mathrm{gen}}) + \kappa(u; c_{\mathrm{guide}})\, G_\phi(\bar{q})^{-1}\, \nabla_{\bar{q}} U.$$

The gain $\kappa(u)$ and the utility U are dimensionless, larger U is preferred, and G_ϕ is the decoder metric defined next.

Decoder-feature metric and trust region. Equal latent steps can change the decoded molecule by unequal amounts, so the design measures a step by its effect on decoded features (Figure 3a and Figure 1b). A soft-token adapter $h_\phi : \mathbb{R}^D \to \mathbb{R}^{D_h}$ replaces each designable residue by the probability-weighted mixture of the 20 amino acids and runs the decoder once. For n_m assembled residues it returns $D_h = 131\, n_m$ features: per residue 3 for the C_α position, 108 for the other atom means and 20 for the residue probabilities. Let $J_h = \partial h_\phi / \partial \bar{q}$ be its Jacobian, W_h a symmetric non-negative feature weight matrix set by the steering task, and $\tau_{\mathrm{ridge}} > 0$ a ridge. The metric is

$$G_\phi = J_h^\top W_h J_h + \tau_{\mathrm{ridge}}\, I_D.$$

A smooth decoder induces the pullback $J_h^\top J_h$ as a Riemannian metric on the latent space [15]. The ridge gives $v^\top G_\phi v \geq \tau_{\mathrm{ridge}} \|v\|^2 > 0$, so G_ϕ is positive definite and the solve $G_\phi d = \nabla U$ is defined.

The cap bounds the guided increment in the local decoder-feature metric. For solver step h, gain κ and ascent direction $d = G_\phi^{-1} \nabla U$, the guided increment is $\Delta_{\mathrm{guide}} = h\kappa\alpha d$ with

$$\alpha = \min\left(1,\ \frac{\Delta_G}{h\kappa\sqrt{d^\top G_\phi d} + \epsilon}\right),$$

where ϵ is a small constant. The cap holds the metric norm of the increment at or below the trust radius Δ_G, so the increment stays small in that metric even when the inverse is ill-conditioned. Acceptance additionally uses checks on the hard-decoded candidate, and a failed molecular check rejects the step.

Several objectives. A fixed weighted sum can push against a conflicting objective. The multiple-gradient descent algorithm (MGDA) takes the weights from the gradients themselves [16, 17]. Guided diffusion samplers have used the same Pareto idea for multi-objective generation [18]. Write each criterion as a scaled loss $\widetilde{f}_i = (f_i - b_i)/r_i$ with registered baseline b_i and scale $r_i > 0$, where a utility enters as its negative. Let $g_i = \nabla \widetilde{f}_i$ and $a_i = G_\phi^{-1} g_i$. The weights solve a small quadratic programme over the simplex Δ_{n_f} of n_f non-negative weights that sum to one:

$$\lambda^\star = \arg\min_{\lambda \in \Delta_{n_f}} \Big\| \sum_i \lambda_i a_i \Big\|_{G_\phi}^2, \qquad w = \sum_i \lambda_i^\star a_i.$$

This projects the origin onto the convex hull of the preconditioned gradients in the metric G_ϕ. For every i, $g_i^\top w \geq \|w\|_{G_\phi}^2$, so $-w$ lowers every scaled loss to first order when $w \neq 0$, and $w = 0$ marks a state that is Pareto-stationary to first order [16]. As a check of this property, $G = I$, $g_1 = (1, 0)$ and $g_2 = (0, 2)$ give the weights $(0.8, 0.2)$ and $w = (0.8, 0.4)$, and both inner products equal $\|w\|^2 = 0.8$ (Figure 3a).

A finite step is accepted when every scaled loss falls by at least $c_{\mathrm{A}} h \|w\|_{G_\phi}^2$ for a constant $0 < c_{\mathrm{A}} < 1$ (Armijo condition) and $h\|w\|_{G_\phi} \leq \Delta_G$. Backtracking shortens h until both hold. The weights change with the state, so the combined field is a controlled vector field that is non-conservative, with no single potential behind it. Gate F-MN compares min-norm steering with a fixed-weight sum at matched seeds and compute before activation.

2.1. Sequential Monte Carlo

Guidance moves particles without correcting the law that results. SMC runs N_p trajectories together, weights each move by how far the proposal differs from the base process, and resamples when a few weights dominate [19]. The design uses SMC for the stochastic profile only, because two deterministic updates are point masses on different states and have no density ratio. The proofs do not cover the deterministic guided field (the controlled-ODE pushforward).

Choose checkpoints $0 = u_0 < \cdots < u_M = 1$. Let K_j be the base transition kernel, R_j the proposal kernel and $\Psi_j \geq 0$ a potential that carries the preference. The target is the base path law reweighted by the product of the potentials, with normaliser Z_M. The twisted construction follows Wu and colleagues [20]. The incremental weight is

$$\omega_j(q_{j-1}, q_j) = \Psi_j(q_{j-1}, q_j)\, \frac{\mathrm{d}\mathsf{K}_j(q_{j-1}, \cdot)}{\mathrm{d}\mathsf{R}_j(q_{j-1}, \cdot)}(q_j).$$

The ratio exists when $\mathsf{K}_j(q, \cdot)$ is absolutely continuous with respect to $\mathsf{R}_j(q, \cdot)$. Every move the base kernel can make must be open to the proposal, because no finite weight restores missing support (Figure 3b).

First configuration. The base kernel serves as proposal, so $\omega_j = \Psi_j$ for $j \geq 1$, and $\Psi_0 \equiv 1$. Rewards are evaluated at steps 100, 200, 300 and 400 of 400, so $M = 4$. Set $R_0 = 0$ and use $\Psi_j = \exp\{\beta(R_j - R_{j-1})\}$, where R_j is the interface confidence of the complex-prediction judge (Technical Report III (Section "Judging and specificity")) at check j, computed on the clean-endpoint estimate of the state, and β is an inverse temperature. The product telescopes to $\exp\{\beta(R_M - R_0)\} = \exp\{\beta R_M\}$, a tilt by the final reward, as in recent Feynman-Kac and particle-steered generative samplers [21–25] and in Proteina-Complexa, whose reward is an interface score from a structure predictor [5]. The settings are $N_p \in \{8, 16, 32\}$ and $\beta \in \{0.5, 1, 2\}$, chosen on a validation set before any evaluation set is read. Guided proposals can be added later, and the support condition is then checked for each configuration. The normaliser Z_M belongs to this fixed, discrete, reduced-noise model.

Weights and resampling. Weights are held as logarithms and normalised by a shifted log-sum-exp, which stays finite where the exponentials underflow. Missing support, all log weights equal to $-\infty$ and a NaN or $+\infty$ are failures, whereas finite log weights never abort a run. The effective sample size $\mathrm{ESS} = 1/\sum_k (\bar{w}^{(k)})^2$ lies between 1 and N_p (Figure 3b). Resampling is stratified, runs when ESS falls below $N_p/2$ and is skipped at the last step. The ancestor indices, the weights before and after resampling, the ESS and the log normaliser are stored.

Estimator. The normaliser is estimated by the product, over the steps, of the mean incremental weight:

$$\widehat{Z}_M = \prod_{j=1}^{M} \bar{\ell}_j, \qquad \bar{\ell}_j = \sum_{k=1}^{N_p} \widetilde{w}_{j-1}^{(k)}\, \omega_j\big(q_{j-1}^{(a_k)}, q_j^{(k)}\big).$$

Here $\widetilde{w}_{j-1}^{(k)}$ is the normalised weight with which particle k enters step j ($\widetilde{w}_{j-1}^{(k)} = 1/N_p$ after resampling and $\widetilde{w}_{j-1}^{(k)} = \bar{w}_{j-1}^{(k)}$ when resampling is skipped), and a_k is its ancestor. The first particles come from the initial law with equal weights. The untwisted configuration has no step-0 factor, because $\Psi_0 \equiv 1$ and the initial proposal is the initial law. Just after resampling the weights are equal and $\bar{\ell}_j$ is the plain mean. When resampling is skipped, the carried weights are generally

unequal, and replacing them by equal weights can bias the estimate. If some $\bar{\ell}_j = 0$, then $\widehat{Z}_M = 0$. The normalised estimate of an expectation is $\widehat{\eta}_M(\varphi) = \sum_k \bar{w}_M^{(k)} \varphi(q_M^{(k)})$.

A twist $h_j(q)$ is the expected product of the remaining potentials from state q at checkpoint j, with $h_M \equiv 1$. Keep K_j as the reference kernel. For positive functions h_j define $\Psi_0^h = h_0$ and $\Psi_j^h(q, q') = \Psi_j(q, q')\, h_j(q')/h_{j-1}(q)$. Compute the weights as $\omega_0 = h_0\, \mathrm{d}p_0/\mathrm{d}\mathsf{R}_0$ and $\omega_j = \Psi_j^h\, \mathrm{d}\mathsf{K}_j/\mathrm{d}\mathsf{R}_j$, and include the initial factor in $\widehat{Z}_M = \prod_{j=0}^{M} \bar{\ell}_j$. For exact backward functions choose the proposal $\mathsf{R}_j = \mathsf{K}_j^h = \mathsf{K}_j \Psi_j h_j / h_{j-1}$ and $\mathsf{R}_0 = p_0 h_0 / Z_M$. Then $\omega_0 = Z_M$ and $\omega_j = 1$ for $j \geq 1$, which gives zero variance (Theorem 4.10 of Supplementary Section S1). Approximate positive twists retain unbiasedness under the stated importance-weighting and resampling assumptions. The first configuration omits twisting.

The statements about this estimator are mathematical results on idealised or finite models, and Supplementary Section S1 gives the assumptions and proofs. Under the support condition, bounded incremental weights and offspring counts whose mean given the past is $N_p \bar{w}$, $\mathbb{E}[\widehat{Z}_M] = Z_M$ for every $N_p \geq 1$ and every resampling schedule that depends only on the past, an ESS trigger included. With the exact twist and that initial proposal, $\widehat{Z}_M = Z_M$ on every realisation, and with an approximate positive twist $\widehat{Z}_M$ stays unbiased. The logarithm $\log \widehat{Z}_M$ is biased low by Jensen's inequality, and $\widehat{\eta}_M$ is a ratio that is biased in general at finite N_p. Consistency of the weighted particles is shown for resampling before every step by a multinomial, residual or stratified scheme [26]. This result excludes systematic resampling on unordered particles, whose offspring counts are not negatively associated, and an adaptive schedule. A bias of order $1/N_p$ has no usable a-priori constant, the proofs stop short of the controlled-ODE pushforward, and the guarantees for the SMC estimator, head calibration and specificity control stay separate.

Replicates and diagnostics. Because no a-priori error constant is available, the design runs $R \geq 32$ independent replicates from independent seeds and reports their mean, the empirical standard error and a 95 percent percentile interval. The floor of 32 is a setting of the design. When the estimator is biased, a percentile interval over replicates does not certify coverage of the true value. ESS is a diagnostic that reads the weights only [27], and it equals N_p right after resampling even when every particle descends from one ancestor, so the stored ancestor indices are inspected too.

3. Maturation and editing

Maturation edits one existing antibody or nanobody, the parent, within a declared mutation budget. The base model has no parent input, no editing mask and no maturation step. The full system adds them by partial noising: a generative model restarted from a noised copy of a sample and run on to the data end returns a variant of that sample, as SDEdit does for images [28]. The restart point lies on the master clock of the flow (Technical Report I (Section "Generative model")). Corruption, guidance, solver, hard decoding, context restoration and rejection together form one proposal kernel K_{edit} (Technical Report III (Table: *Contract of the three modules that act on a finished candidate*)). Published methods steer discrete sequence models with a predictor or classifier [29–31]. The editor here also moves structure, and the design excludes fine-tuning.

Editing contract. Three nested masks decide what an edit may touch (Technical Report I (Section "State and geometry")): $d_{\mathrm{res}} \preceq e_{\mathrm{res}} \preceq m_{\mathrm{res}}$ for design-mutable, transported and assembled

residues. Rows in m_{res} outside e_{res} are fixed context. Rows in e_{res} outside d_{res} move in space while their amino-acid identity stays clamped. The frozen encoder supplies the packed parent state $q^{\text{par}} = (x_e^{\text{par}}, z_e^{\text{par}})$, the C-alpha positions and 8-number latents of the n_e transported residues. One edit draws

$$q' \sim K_{\text{edit}}(\cdot \mid q^{\text{par}}, m_{\text{res}}, e_{\text{res}}, d_{\text{res}}, \mathcal{S}, c_{\text{gen}}, u^*, c_{\text{guide}}, \omega)\,.$$

Here $\mathcal{S}_i$ is the set of amino acids allowed at position i, which holds only the parent residue where $d_{\text{res},i} = 0$. The symbol c_{gen} is the antigen and format context, u^* the re-entry point, c_{guide} the guidance configuration and ω the noise and decoding seeds. Edits use the context-anchored profile (Technical Report I (Section "Rigid motions and coordinate gauges")), so the flow state has $11n_e$ numbers and fixed rows add no dimension, loss or velocity. Positions follow IMGT labels (Technical Report I (Section "Numbering, regions and the CDR3 prior")).

Four format profiles set the masks. `VHH_CDR_EDIT` designs selected VHH CDR positions, transports them with registered seam neighbours and fixes the remaining framework and scaffold, with a VHH scaffold and cysteine table as an additional gate. `FV_CDR_EDIT` does the same for selected VH and VL CDR positions, fixes the frameworks, the chain register and protected interface anchors, and adds the assembled VH/VL pairing and pose as a gate (Technical Report I (Section "Scaffold profiles, cysteines and pairing")). `FRAMEWORK_HUMANIZE` designs listed framework positions, transports them with local geometry neighbours, fixes the parental CDRs, and gates on humanness and retention of the parental binding mode. `LIABILITY_REPAIR` designs the residues of one liability window and gates on liability removal without introduced liabilities.

Partial corruption on the master clock. The flow moves a state from noise at master time $u = 0$ to data at $u = 1$. The schedules $t_x(u)$ and $t_z(u)$ give the share of the data endpoint that coordinates and latents have reached. An edit enters part-way. It picks one re-entry point $u^* \in (0, 1)$, draws base noise $(x_e^0, z_e^0) \in \mathbb{R}^{11n_e}$ and mixes noise and parent on the linear path used in training:

$$x_{e,u^*} = (1 - t_x(u^*))\, x_e^0 + t_x(u^*)\, x_e^{\text{par}}, \qquad z_{e,u^*} = (1 - t_z(u^*))\, z_e^0 + t_z(u^*)\, z_e^{\text{par}}.$$

A larger u^* keeps more of the parent. One master clock keeps the edit on the sampling curve, while the two schedules still corrupt coordinates and latents to different degrees. With the schedules $t_x(u) = 2u - u^2$ and $t_z(u) = u^2$ and $u^* = 0.75$, the edit keeps $t_x = 0.9375$ of the parent coordinates and $t_z = 0.5625$ of the parent latents.

After each solver step the fixed context is written back. This inserts conditioning data and leaves the transported density unchanged. A hard decode then turns the category logits into one amino acid per residue inside $\mathcal{S}_i$, which restores parent identities outside d_{res}. Atom-mask, bond, chirality, assembly and profile checks decide acceptance. A seam check measures bond geometry, clashes and decoder residuals around every mask boundary, and a mask and profile pair that fails it is disabled.

Mutation-budget controller. A user asks for k mutations, but the editor controls u^*, so a controller learns the link. For each u^* on a grid it decodes R unguided proposals per calibration parent and averages the number of design positions that differ from the parent:

$$\widehat{k}(u^*; L, p) = \frac{1}{R} \sum_{r=1}^{R} d_H\Big(s^{(r)}|_{d_{\text{res}}}, s^{\text{par}}|_{d_{\text{res}}}\Big)\,.$$

Here d_H is the Hamming distance, $s^{(r)}$ the r-th decoded sequence, L the editable length and p the format profile. For a requested expected budget k the controller interpolates this curve inside the calibration grid where it is monotone. An exact budget is a separate mode that accepts only decodes with $d_H = k$ and reports the rejection rate. If no u^* yields k, the controller reports the budget as unavailable, with the realised edit-count distribution. Gate F-M4 tests fidelity on held-out parents, by length and format stratum. It requires an upper confidence limit of the deviation within $\tau(k) = \max\{1, 0.2k\}$ mutations in every cell, and monotonicity and invertibility of the curve over each qualified length band. The protocol uses 300 parents per length decile and budgets $k = 1$ to 10.

Guided editing. Guidance follows the endpoint, metric and trust-region contracts of Section 2. Affinity change uses the thermodynamic sign $\Delta\Delta G = \Delta G_{\text{variant}} - \Delta G_{\text{parent}}$, so a negative value predicts stronger binding, and the affinity utility is $U_{\text{aff}} = -\widehat{\Delta\Delta G}$. The utility gradient, preconditioned by the decoder-feature metric G_ϕ, gives the direction $d = G_\phi^{-1}\nabla U_{\text{aff}}$ on the packed e_{res} axes, with coordinate part d^x and latent part d^z. The edit solver adds it to the schedule-scaled base field, with the arguments of the heads left out:

$$\frac{\mathrm{d}}{\mathrm{d}u}\begin{bmatrix} x_e \\ z_e \end{bmatrix} = \begin{bmatrix} t_x'(u)\, v_\theta^x \\ t_z'(u)\, v_\theta^z \end{bmatrix} + \begin{bmatrix} \kappa_x(u)\, I & 0 \\ 0 & \kappa_z(u)\, I \end{bmatrix} \begin{bmatrix} d^x \\ d^z \end{bmatrix}.$$

The heads give velocity per unit block time, and the schedule derivatives convert it to master time once. The gains $\kappa_x(u)$ and $\kappa_z(u)$ are stored with the edit profile. The trust cap of Section 2 limits the metric norm of each guided increment. A property model trained on finished molecules can be unreliable on partly corrupted states. A guidance-depth check therefore compares matched guided and unguided edits on a frozen grid of u^* values for soft-to-hard score agreement, enrichment by an independent ranker, validity and seam failures, pose retention and diversity. A schedule is active only where u clears every check.

Mutation supply. The default editor proposes amino-acid changes at d_{res} positions through the learned latent kernel. Real hypermutation rates depend on neighbouring nucleotides, which the S5F branch supplies when the parent has an observed nucleotide sequence, a species and chain class inside the model scope and a compatible lineage. It uses the 5-mer targeting and substitution models of Yaari et al. [32], whose stated scope is human heavy chains. A change is weighted by how often its site mutates, how often the base becomes the new base, and how much the model prefers the translated result. For nucleotide position i and substitution $b \to b'$,

$$Q_{\text{S5F}}(i, b \to b') \propto \mu_{\text{mut}}(c_5(i))\, \sigma(b \to b' \mid c_5(i))\, \exp\{\beta_{\text{S}}\, U(\text{translate}(s_{\text{nt}}^{i \to b'}))\}.$$

Here $c_5(i)$ is the 5-mer around i, μ_{mut} the relative mutability of that context, σ the substitution probability given it, $s_{\text{nt}}^{i \to b'}$ the nucleotide sequence after the change, U a model utility and β_{S} its strength. Stop codons, frameshifts, forbidden cysteine changes and protected residues get zero weight. Gate F-M5 admits the branch once motif assets, source scope and a fixed-horizon comparison against uniform targeting are fixed, and targeting is uniform until then.

Holding the framework and revising it on purpose. Hard preservation covers chain register, identities outside d_{res}, coordinates outside e_{res}, disulfide topology, required scaffold positions, and atom and chemistry validity. A violation rejects the draw. Soft preservation adds a cost on the

distance from the parent in the decoder-feature metric, so a move that barely changes the decoded molecule costs little: $\delta_G(X, X^{\mathrm{par}}) = \sqrt{\Delta q^\top G_\phi(q^{\mathrm{par}}; c_{\mathrm{gen}})\,\Delta q}$ for the change Δq of the packed state.

Revising the framework is a separate act. With protein sequence only, IgBLAST returns the nearest translated germline V reference [33], and an edit toward it is named `aa_reference_reversion`. For a candidate reversion set R the editor chooses at most k_R reversions that raise humanness most while pose, affinity and liabilities stay acceptable:

$$\max_{S\subseteq R,\ |S|\le k_R} U_{\mathrm{human}}(X_S) \quad \text{subject to} \quad \mathrm{PosePass}(X_S)=1,\ \widehat{\Delta\Delta G}(X_S) \le \tau_{\Delta\Delta G},\ \mathrm{LiabilityPass}(X_S)=1,$$

where X_S is the candidate with the reversions in S, U_{human} a humanness score and $\tau_{\Delta\Delta G}$ an affinity-loss limit. Because mutation effects interact, the editor enumerates exactly for small $|R|$, otherwise uses a bounded beam, and evaluates the final subset jointly. Gate F-M6 uses at least 100 held-out parents under one declared affinity assay. It requires a lower parent-cluster-bootstrap limit of 0.05 on the median OASis humanness gain [34], a lower exact-binomial limit of 0.90 on the fraction with non-inferior measured affinity, and no newly introduced structure-confirmed liability.

Liability-aware editing. Every hard-decoded proposal is scanned by the liability channel (Technical Report III (Section "Liability, immunogenicity and glycan channels")). Motifs inherited from the parent stay apart from motifs the edit adds:

$$\mathcal{L}_{\mathrm{new}}(X', X^{\mathrm{par}}) = \mathcal{L}(X') \setminus \mathcal{L}(X^{\mathrm{par}}),$$

where $\mathcal{L}(X)$ is the set of liability motifs in X. A sequence motif only nominates a site for a structural check, because motif matching over-predicts. In 37 therapeutic antibodies it gave false-positive rates of 0.31 for aspartate and 0.41 for asparagine sites, against 0.02 and 0.04 for a structure-based classifier [35]. Structure-confirmed liabilities drive rejection. Gate F-M7 takes 500 accepted variants and requires a one-sided binomial upper limit of at most 0.02 on the fraction that introduce a structure-confirmed CDR liability. It reports its rejection yield, so rejecting every variant cannot pass. Generic polyspecificity is measured again after every round, for example with a sequence-based nanobody predictor [36]. Pair-specific off-targets use their own panel (Technical Report III (Section "Judging and specificity")).

Rounds and acceptance. Rounds, mutation budgets, guidance schedules and judge budgets are fixed before the evaluation that decides release. Gate F-MAT compares matured candidates with the parent and a declared baseline on seven co-primary endpoints, with the parent as statistical unit. They cover affinity change, pose and epitope retention, off-target margin, polyspecificity response, budget fidelity, liability and scaffold compliance, and diversity. AbBiBench supplies affinity-ranking tasks once its data, licence and splits are admitted [37]. Technical Report III (Section "Evaluation design") places these gates in the evaluation design.

4. Mathematical supplement

This supplement holds the proofs and derivations behind the Methods. The results concern mathematical objects: flows, densities, particle systems, transport plans, metrics and decision rules.

Whether a designed molecule binds, whether a model is accurate and which of two models is better are questions for an assay or a benchmark, and no proposition here answers them.

A result marked cited is a theorem from the literature, used as stated. A proof marked sketch gives the idea and the key step. Results are numbered by section, so that Theorem 4.3 is the third numbered result of Section S1.4. Sections S1.2 to S1.9 cover geometry, scores and flow likelihood, transport and particle limits, decision statistics, the representation floor, the decoder metric, numerical error budgets and sensitivity. Section 4.10 states which results are proved in full and which are cited. Worked examples are numerical illustrations of the statement beside them, computed from the stated formulas.

4.1. Notation and conventions

The symbols follow the notation of the Methods. Table 2 lists the symbols that the proofs use and the section where each is used. A few glyphs have a different local meaning in different sections, as the table notes.

Table 2: Symbols of the mathematical supplement.

Symbol	Meaning	Section
n_e, d_x, D	Transported residue count; coordinate dimension; dimension of the packed state, $D = d_x + 8n_e$	S1.2, S1.3
$q \in \mathbb{R}^D$	Dimensionless intrinsic state	S1.2 to S1.8
$u \in [0,1]$, f_u, ψ_u, p_u	Master clock; field; its flow; density of the state at clock u	S1.2, S1.3
$T = (R, \mathbf{t})$	Rigid motion with rotation R and shift $\mathbf{t}$ (this glyph is used only in S1.2)	S1.2
$(R_c, \mathbf{a}_c)$, $(A_g, \mathbf{o}_g)$	Context frame; chart matrix and offset of coordinate profile g	S1.2
$\Xi(u)$, ζ, $\gamma(u)$	Diffusion matrix of an SDE; Gaussian noise vector; width of a noise bridge	S1.3
ξ_{H}	Hutchinson probe vector	S1.3
Γ, C, μ_R, μ_E, ρ_{OT}, $\varepsilon_{\mathrm{OT}}$	Contact plan; cost; reference masses; marginal and entropic weights	S1.4
M, K_j, R_j, Ψ_j, ω_j, N_{p}	Horizon; reference kernel; proposal kernel; potential; incremental weight; particle count	S1.4
h_j	Backward information function (this glyph also names the decoder feature map h_ϕ in S1.7)	S1.4
$\widetilde{w}_j^{(k)}$, $\bar{w}_j^{(k)}$, $\bar{\ell}_j$, $\widehat{Z}_M$	Carried weight; normalised weight; step factor; estimate of the normalising constant Z_M	S1.4
$\widetilde{f}_i$, g_i, a_i, w, n_f	Scaled loss; its gradient; preconditioned gradient; common direction; number of losses	S1.7
G_ϕ, h_ϕ, J_h, W_h, τ_{ridge}, Δ	Decoder metric; feature map; its Jacobian; feature weights; ridge; trust radius (written Δ_G in the Methods)	S1.7
η, π_j, ν_j	Entropic parameter; reference and alternative weights of plausible worlds	S1.5
$\mathcal{F}_t$, e_t, E_t	Filtration; evidence factor; accumulated evidence	S1.5
ϑ, $\mathcal{O}$	Parameter vector of the sensitivity analysis; map from parameters to observables	S1.9

Norms are Euclidean unless a subscript names another. For a symmetric positive definite matrix G, $\|v\|_G^2 = v^\top G v$. The expression x_+ means $\max(x, 0)$. The generalised relative entropy of nonnegative vectors u and v of equal length is

$$\mathrm{KL}(u \mid v) = \sum_i \left[u_i \log \frac{u_i}{v_i} - u_i + v_i \right],$$

with the conventions $0 \log 0 = 0$, $0 \log(0/0) = 0$, and $+\infty$ for a positive u_i against $v_i = 0$. It is nonnegative, convex in both arguments, and zero exactly when $u = v$. The last two terms make it meaningful when the total masses differ. For probability vectors it is the ordinary relative entropy.

4.2. Geometry and gauges

This section shows that a flow with a symmetric field produces symmetric samples (Theorem 2.1). The shift is removed by a linear change of coordinates whose choice alters no density or divergence (Theorem 2.3, Proposition 2.8). A frame attached to the antigen keeps the pose of the antibody relative to the antigen in the state (Proposition 2.4). Removing rotations as well creates points at which the geometry breaks down (Proposition 2.6).

The base-model code subtracts the mean of the present atoms of each complex from the coordinates in ångströms and then multiplies them by 0.1, so that the encoder input and the flow targets are in nanometres. It stores coordinates in the padded ambient layout and computes no density, score or divergence. The charts below define the space on which densities are evaluated, and they are needed wherever a density or a divergence is computed. The anchored chart belongs to the antigen-conditioned system.

Let $\mathbf{x} = (\mathbf{x}_1, \dots, \mathbf{x}_n) \in (\mathbb{R}^3)^n$ hold the Cα coordinates of $n = n_e \geq 3$ transported residues. A rigid motion $T = (R, \mathbf{t})$ with $R \in \mathrm{SO}(3)$ and $\mathbf{t} \in \mathbb{R}^3$ acts on every point by the same rule:

$$T \cdot \mathbf{x} = (R\mathbf{x}_1 + \mathbf{t}, \dots, R\mathbf{x}_n + \mathbf{t}).$$

Distances, bond angles and signed volumes do not change under this action. A reflection keeps distances but reverses the sign of a signed volume, and so turns L-amino acids into D-amino acids. The symmetry group is therefore $\mathrm{SE}(3) = \mathrm{SO}(3) \ltimes \mathbb{R}^3$ and does not include reflections. A velocity is a difference of positions, so a shift does not change it and a rotation turns it by R. Write $\rho_T \mathbf{v} = (R\mathbf{v}_1, \dots, R\mathbf{v}_n)$ for the induced action on velocities. This is an orthogonal map.

A field $f_u(\mathbf{x}, c)$ that depends on a condition c (the antigen coordinates, for instance) is called equivariant if

$$f_u(T \cdot \mathbf{x}, T \cdot c) = \rho_T f_u(\mathbf{x}, c)$$

for all states, clocks u and rigid motions. A scalar U is called invariant if $U(T \cdot \mathbf{x}, T \cdot c) = U(\mathbf{x}, c)$.

Theorem 4.1 (2.1 — equivariance of the flow). *Fix a condition c. Let $f_u(\mathbf{x}, c)$ be continuous in $u \in [0, 1]$, uniformly Lipschitz in $\mathbf{x}$, and equivariant. Let $\psi_{u,c}(\mathbf{x}_0)$ be the solution of $\frac{\mathrm{d}}{\mathrm{d}u}\psi_{u,c}(\mathbf{x}_0) = f_u(\psi_{u,c}(\mathbf{x}_0), c)$ with $\psi_{0,c}(\mathbf{x}_0) = \mathbf{x}_0$. Then for every rigid motion T,*

$$\psi_{u,T\cdot c}(T \cdot \mathbf{x}_0) = T \cdot \psi_{u,c}(\mathbf{x}_0).$$

If in addition the base law satisfies $p_0(\cdot \mid T \cdot c) = T_{\#} p_0(\cdot \mid c)$, then the endpoint laws satisfy $p_1(\cdot \mid T \cdot c) = T_{\#} p_1(\cdot \mid c)$.

Proof. Put $\mathbf{y}(u) = T \cdot \psi_{u,c}(\mathbf{x}_0)$. The action is affine, so $\dot{\mathbf{y}} = R\dot{\psi} = \rho_T f_u(\psi_{u,c}(\mathbf{x}_0), c)$. Equivariance with the state $\psi_{u,c}(\mathbf{x}_0)$ gives $\rho_T f_u(\psi_{u,c}(\mathbf{x}_0), c) = f_u(\mathbf{y}(u), T \cdot c)$. Also $\mathbf{y}(0) = T \cdot \mathbf{x}_0$. So $\mathbf{y}$ solves the equation with condition $T \cdot c$ and start $T \cdot \mathbf{x}_0$. The Lipschitz condition gives uniqueness (Picard–Lindelöf), hence $\mathbf{y}(u) = \psi_{u,T\cdot c}(T \cdot \mathbf{x}_0)$. For the laws, $p_1(\cdot \mid T \cdot c) = (\psi_{1,T\cdot c})_{\#} T_{\#} p_0(\cdot \mid c) = (\psi_{1,T\cdot c} \circ T)_{\#} p_0(\cdot \mid c) = (T \circ \psi_{1,c})_{\#} p_0(\cdot \mid c) = T_{\#} p_1(\cdot \mid c)$. □

The theorem concerns a field that is equivariant by assumption.

Lemma 4.2 (2.2 — centring projector). *Let $P_n = I_n - \mathbf{1}\mathbf{1}^\top/n$. Then $P_n = P_n^\top = P_n^2$, $P_n\mathbf{1} = 0$ and* $\operatorname{rank} P_n = n - 1$. *The matrix $P = P_n \otimes I_3$ is the orthogonal projector onto $\mathcal{X}_n = \{\mathbf{x} : \sum_i \mathbf{x}_i = 0\}$.*

Proof. Since $\mathbf{1}^\top\mathbf{1} = n$, $(\mathbf{1}\mathbf{1}^\top/n)^2 = \mathbf{1}\mathbf{1}^\top/n$, which gives symmetry and idempotence. The null space of P_n is spanned by $\mathbf{1}$ and its range is $\mathbf{1}^\perp$. Tensoring with I_3 applies the same projection to each of the three coordinate columns. □

Theorem 4.3 (2.3 — translation-free chart). *Let the columns of $B \in \mathbb{R}^{n\times(n-1)}$ be an orthonormal basis of $\mathbf{1}^\perp$ and put $A = B^\top \otimes I_3$. Then A maps $\mathcal{X}_n$ isometrically onto $\mathbb{R}^{3n-3}$ with inverse $A^\top$, and every $\mathbf{x} \in (\mathbb{R}^3)^n$ splits uniquely as*

$$\mathbf{x} = A^\top A\mathbf{x} + \mathbf{1} \otimes \bar{\mathbf{x}}, \qquad \bar{\mathbf{x}} = \frac{1}{n}\sum_i \mathbf{x}_i.$$

If $\widetilde{B}$ is another orthonormal basis, then $\widetilde{A} = QA$ for an orthogonal $Q \in \mathrm{O}(3n-3)$, and a density p_A in the A chart becomes $p_{\widetilde{A}}(\widetilde{q}) = p_A(Q^\top\widetilde{q})$, with no log-determinant term.

Proof. The basis identities are $B^\top B = I_{n-1}$ and $BB^\top = P_n$. They give $AA^\top = I_{3n-3}$ and $A^\top A = P$. For centred $\mathbf{x}$, $P\mathbf{x} = \mathbf{x}$, so $\|A\mathbf{x}\|^2 = \mathbf{x}^\top P\mathbf{x} = \|\mathbf{x}\|^2$ and $A^\top A\mathbf{x} = \mathbf{x}$. For any q, $A^\top q$ is centred and $A(A^\top q) = q$, so A is a bijection. Further, $P\mathbf{x} = \mathbf{x} - \mathbf{1}\otimes\bar{\mathbf{x}}$ gives the splitting, and its two summands are orthogonal. Two bases span the same space, so $\widetilde{B} = BQ_0$ with Q_0 orthogonal, and $Q = Q_0^\top \otimes I_3$ is orthogonal. A change of variables by an orthogonal map has Jacobian determinant of absolute value 1. □

For $n_e = 130$ transported residues the latent block has $8n_e = 1{,}040$ numbers. The translation-free chart then has $3\cdot 129 = 387$ coordinate numbers and $D = 1{,}427$. The context-anchored chart below keeps the relative shift, has 390 coordinate numbers, and $D = 1{,}430$. In general $D = 11n_e - 3$ for the free chart and $D = 11n_e$ for the anchored chart.

A Gaussian vector $\ell_0 A^\top\xi$ with $\xi \sim \mathcal{N}(0, I_{3n-3})$ has ambient covariance $\ell_0^2 P$. This matrix is singular in $\mathbb{R}^{3n}$, so the density of such a vector exists only on the centred subspace. Any density in this supplement is therefore evaluated in the declared chart.

Proposition 4.4 (2.4 — anchored coordinates). *Let the condition fix an origin $\mathbf{a}_c \in \mathbb{R}^3$ and a rotation $R_c \in \mathrm{SO}(3)$, and define $\mathbf{y}_i = R_c^\top(\mathbf{x}_i - \mathbf{a}_c)$. A rigid motion $T = (R, \mathbf{t})$ acts on the frame by $\mathbf{a}_{T\cdot c} = R\mathbf{a}_c + \mathbf{t}$ and $R_{T\cdot c} = RR_c$. Then $\mathbf{y}_i(T \cdot \mathbf{x}, T \cdot c) = \mathbf{y}_i(\mathbf{x}, c)$, and the map $\mathbf{x} \mapsto \mathbf{y}$ keeps all $3n$ numbers of the antibody.*

Proof. Substitute: $(RR_c)^\top\big((R\mathbf{x}_i + \mathbf{t}) - (R\mathbf{a}_c + \mathbf{t})\big) = R_c^\top R^\top R(\mathbf{x}_i - \mathbf{a}_c) = \mathbf{y}_i(\mathbf{x}, c)$. The map is affine and invertible, with inverse $\mathbf{x}_i = \mathbf{a}_c + R_c\mathbf{y}_i$. □

Proposition 4.5 (2.5 — error in an estimated frame). *For rotations $R, \widehat{R}$ and origins $\mathbf{a}, \widehat{\mathbf{a}}$,*

$$\|\widehat{R}^\top(\mathbf{x}-\widehat{\mathbf{a}}) - R^\top(\mathbf{x}-\mathbf{a})\| \le \|\widehat{R}-R\|_2\,\|\mathbf{x}-\mathbf{a}\| + \|\widehat{\mathbf{a}}-\mathbf{a}\|.$$

To first order, with $\widehat{R} = R\exp([\delta\boldsymbol{\omega}]_\times)$ and $\widehat{\mathbf{a}} = \mathbf{a} + \delta\mathbf{a}$, the anchored coordinates change by $\delta\mathbf{y} = -[\delta\boldsymbol{\omega}]_\times\mathbf{y} - R^\top\delta\mathbf{a}$.

Proof. Write the difference as $\widehat{R}^\top\big((\mathbf{x}-\widehat{\mathbf{a}}) - (\mathbf{x}-\mathbf{a})\big) + (\widehat{R}-R)^\top(\mathbf{x}-\mathbf{a})$ and use $\|\widehat{R}\|_2 = 1$. For the expansion, $\widehat{R}^\top(\mathbf{x}-\widehat{\mathbf{a}}) = \exp(-[\delta\boldsymbol{\omega}]_\times)R^\top(\mathbf{x}-\mathbf{a}-\delta\mathbf{a})$, and $\exp(-[\delta\boldsymbol{\omega}]_\times) = I - [\delta\boldsymbol{\omega}]_\times + O(\|\delta\boldsymbol{\omega}\|^2)$. □

A rotation error of the frame therefore grows with the distance from the origin, and a shift error moves every point equally.

The flow of the Methods works in the flat charts above, and shape space is used only to compare conformations.

Proposition 4.6 (2.6 — orbit tangent, inertia tensor and horizontal projector). *For $\boldsymbol{\omega} \in \mathbb{R}^3$ let $(K_{\mathbf{x}}\boldsymbol{\omega})_i = \boldsymbol{\omega}\times\mathbf{x}_i$. This map spans the vertical space, the tangent space of the rotation orbit through $\mathbf{x}$. Let $I_{\mathbf{x}} = \sum_i(\|\mathbf{x}_i\|^2 I_3 - \mathbf{x}_i\mathbf{x}_i^\top)$. Then:*

1. The adjoint of $K_{\mathbf{x}}$ is $K_{\mathbf{x}}^*\mathbf{v} = \sum_i \mathbf{x}_i\times\mathbf{v}_i$ (the angular momentum) and $K_{\mathbf{x}}^*K_{\mathbf{x}} = I_{\mathbf{x}}$.
2. $\boldsymbol{\omega}^\top I_{\mathbf{x}}\boldsymbol{\omega} = \sum_i \|\boldsymbol{\omega}\times\mathbf{x}_i\|^2$. Hence $I_{\mathbf{x}} \succeq 0$, and $I_{\mathbf{x}}$ is singular exactly when some nonzero axis is parallel to every nonzero $\mathbf{x}_i$. For a centred configuration this means that all points lie on one line through the origin.
3. If $I_{\mathbf{x}} \succ 0$, the map $P_{\mathbf{x}}^{\mathrm{H}} = I - K_{\mathbf{x}}I_{\mathbf{x}}^{-1}K_{\mathbf{x}}^*$ is the orthogonal projector onto the horizontal space, the set of velocities with zero angular momentum. It satisfies $P_{R\mathbf{x}}^{\mathrm{H}}(R\mathbf{v}) = R\,P_{\mathbf{x}}^{\mathrm{H}}\mathbf{v}$.
4. If the right-hand side $\mathbf{b}$ of $I_{\mathbf{x}}\boldsymbol{\omega} = \mathbf{b}$ is perturbed by $\delta\mathbf{b}$, the solution changes by at most $\|\delta\mathbf{b}\|/\lambda_{\min}(I_{\mathbf{x}})$.

Proof. (1) The scalar triple product gives $\langle K_{\mathbf{x}}\boldsymbol{\omega}, \mathbf{v}\rangle = \sum_i(\boldsymbol{\omega}\times\mathbf{x}_i)^\top\mathbf{v}_i = \boldsymbol{\omega}^\top\sum_i\mathbf{x}_i\times\mathbf{v}_i$. Also $\mathbf{x}\times(\boldsymbol{\omega}\times\mathbf{x}) = (\|\mathbf{x}\|^2 I - \mathbf{x}\mathbf{x}^\top)\boldsymbol{\omega}$, and summing over i gives $K_{\mathbf{x}}^*K_{\mathbf{x}} = I_{\mathbf{x}}$. (2) The quadratic form is $\|K_{\mathbf{x}}\boldsymbol{\omega}\|^2$, which vanishes exactly when every $\boldsymbol{\omega}\times\mathbf{x}_i = 0$. (3) When $I_{\mathbf{x}} = K_{\mathbf{x}}^*K_{\mathbf{x}}$ is invertible, $K_{\mathbf{x}}I_{\mathbf{x}}^{-1}K_{\mathbf{x}}^*$ is the orthogonal projector onto range $K_{\mathbf{x}}$. Therefore $K_{\mathbf{x}}^*P_{\mathbf{x}}^{\mathrm{H}}\mathbf{v} = 0$, the removed part lies in the vertical space, and the two parts are orthogonal. Cross products obey $(R\mathbf{x})\times(R\mathbf{v}) = R(\mathbf{x}\times\mathbf{v})$ and $I_{R\mathbf{x}} = RI_{\mathbf{x}}R^\top$, which gives equivariance. (4) The two solutions differ by $I_{\mathbf{x}}^{-1}\delta\mathbf{b}$, and $\|I_{\mathbf{x}}^{-1}\|_2 = 1/\lambda_{\min}(I_{\mathbf{x}})$. □

By item 2, the orbit through a collinear configuration has dimension two, which is why the quotient by rotations is not a smooth manifold there, and by item 4 the horizontal solve becomes unstable as $\lambda_{\min}(I_{\mathbf{x}}) \to 0$.

Rotations themselves are handled by the matrix exponential. For $\Omega^\top = -\Omega$ and $R \in \mathrm{SO}(3)$ the update $R^+ = R\exp(h\Omega)$ stays in $\mathrm{SO}(3)$ for every real h, because $\exp(h\Omega)^\top = \exp(h\Omega)^{-1}$ and $\det\exp(h\Omega) = \exp(h\,\mathrm{tr}\,\Omega) = 1$. By Rodrigues' formula, $\exp([\boldsymbol{\omega}]_\times) = I + \frac{\sin\theta}{\theta}[\boldsymbol{\omega}]_\times + \frac{1-\cos\theta}{\theta^2}[\boldsymbol{\omega}]_\times^2$ with $\theta = \|\boldsymbol{\omega}\|$. At $\theta = \pi$ the sine term vanishes, so a half-turn about $\mathbf{n}$ and about $-\mathbf{n}$ coincide. The logarithm of a half-turn is therefore two-valued, and a velocity target built from it is discontinuous. The flow of the Methods transports Cα coordinates in Euclidean charts and decodes local frames afterwards, so it never uses a logarithm of a rotation.

Lemma 4.7 (2.7 — gradients of invariant objects). *Let ρ be an orthogonal representation acting on a chart and let $U(\rho_T q, T \cdot c) = U(q, c)$. Then $\nabla_q U(\rho_T q, T \cdot c) = \rho_T \nabla_q U(q, c)$. In particular, if a density $p(\cdot \mid c)$ is invariant, with $p(\rho_T q \mid T \cdot c) = p(q \mid c)$ and a volume that rigid motions preserve, its score $\nabla_q \log p$ is equivariant. If $G(\rho_T q, T \cdot c) = \rho_T G(q, c) \rho_T^\top$, then $G^{-1}\nabla U$ is equivariant.*

Proof. For every perturbation $\mathbf{h}$, invariance gives $\nabla U(\rho_T q, T \cdot c)^\top \rho_T \mathbf{h} = \nabla U(q, c)^\top \mathbf{h}$. Hence $\rho_T^\top \nabla U(\rho_T q, T \cdot c) = \nabla U(q, c)$, and multiplication by ρ_T gives the claim. The statement for $\log p$ follows because $|\det R| = 1$ keeps the volume, and the statement for $G^{-1}\nabla U$ because $(\rho_T G \rho_T^\top)^{-1} = \rho_T G^{-1} \rho_T^\top$. □

Proposition 4.8 (2.8 — divergence under a change of chart). *Let $Q \in \mathrm{O}(d)$, $\mathbf{y} = Q\mathbf{x}$ and $\widetilde{f}(\mathbf{y}) = Q f(Q^\top \mathbf{y})$. Then $\operatorname{div}_{\mathbf{y}} \widetilde{f}(\mathbf{y}) = \operatorname{div}_{\mathbf{x}} f(\mathbf{x})$. Consequently the rate $\frac{\mathrm{d}}{\mathrm{d}u} \log p_u$ along a flow and the integrated log-likelihood of Theorem 3.7 do not depend on the orthonormal basis B chosen in Theorem 2.3.*

Proof. The Jacobian of $\widetilde{f}$ is $Q D f Q^\top$, and the trace is invariant under cyclic permutation: $\operatorname{tr}(Q D f Q^\top) = \operatorname{tr}(D f)$. □

4.3. Scores, diffusion reversal and flow likelihood

This section shows that the regression target of the flow loss is the conditional mean of the velocity (Theorem 3.2), also with independent block times (Corollary 3.3). It derives the score (Proposition 3.4) and links stochastic and deterministic samplers (Theorems 3.5 and 3.6). It gives the density of a generated sample with the errors of its numerical evaluation (Theorem 3.7, Propositions 3.8 and 3.9). Figure 1 illustrates the zero-COM Sherman–Morrison subspace projection, chiral SE(3) handedness invariance, and controlled flow transport within a Riemannian trust ball toward PDB 1MEL.

Theorem 3.2 and Corollary 3.3 are the basis of the flow loss of the base model: the code draws independent uniform times for the coordinate block and the latent block and regresses the velocity heads on the difference of data and noise. The results after Corollary 3.3 support guidance, sampling with a score and likelihood evaluation, which the base-model code does not contain.

Lemma 4.9 (3.1 — conditional-mean projection). *Let Y be square-integrable, let X be any random element, and let $h(X)$ be square-integrable. Then*

$$\mathbb{E}\|Y - h(X)\|^2 = \mathbb{E}\|Y - \mathbb{E}(Y \mid X)\|^2 + \mathbb{E}\|\mathbb{E}(Y \mid X) - h(X)\|^2.$$

Hence the conditional mean is the almost surely unique minimiser.

Proof. Put $m = \mathbb{E}(Y \mid X)$ and write $Y - h = (Y - m) + (m - h)$. The cross term has expectation $\mathbb{E}\big[\mathbb{E}(Y - m \mid X)^\top (m - h)\big] = 0$. □

Theorem 4.10 (3.2 — flow-matching target). *Let $q_1 \sim p_{\mathrm{data}}(\cdot \mid c)$ and let q_u be drawn from a conditional path $p_u(\cdot \mid q_1, c)$ with conditional velocity $b_u(q \mid q_1, c)$. Put $B = b_u(q_u \mid q_1, c)$ and assume that (q_u, B) is square-integrable. Then:*

1. The minimiser over measurable fields v of $\mathbb{E}\|v(q_u, u, c) - B\|^2$ is $b_u(q, c) = \mathbb{E}[B \mid q_u = q]$. The conditional loss $\mathcal{L}_{\mathrm{CFM}}(\theta)$ equals the marginal loss $\mathcal{L}_{\mathrm{FM}}(\theta) = \mathbb{E}\|v_\theta - b_u(q_u, c)\|^2$ plus a constant that does not depend on θ. Hence the two losses have the same gradient.
2. If every conditional path satisfies the continuity equation with velocity $b_u(\cdot \mid q_1, c)$ in weak form, then the marginal path $p_u(q \mid c) = \int p_u(q \mid q_1, c)\, p_{\mathrm{data}}(q_1 \mid c)\, \mathrm{d}q_1$ satisfies the continuity equation with the field $b_u(q, c)$.

Proof. (1) Apply Lemma 3.1 with $Y = B$ and $X = q_u$ (at fixed u and c). The second term, $\mathbb{E}\|m - B\|^2$, does not depend on v. For $v = v_\theta$ it is the constant in the claim, and the first term is $\mathcal{L}_{\mathrm{FM}}$. (2) For a smooth compactly supported test function φ, the weak form for the conditional paths gives $\frac{\mathrm{d}}{\mathrm{d}u}\mathbb{E}\varphi(q_u) = \mathbb{E}[\nabla\varphi(q_u)^\top B] = \mathbb{E}[\nabla\varphi(q_u)^\top m(q_u)]$ by the tower property. This is the weak continuity equation for the marginal path with velocity m. Differentiation under the expectation needs the stated integrability and a dominating function for the derivative. □

Corollary 4.11 (3.3 — independent block times and the master clock). *Let the state have blocks $b \in \{x, z\}$ with paired linear paths $q_b = (1 - t_b)q_{b,0} + t_b q_{b,1}$, where $q_0 \sim p_0$ is independent of q_1 and the times $t_x, t_z \in [0, 1]$ are independent of the pair. Let $v_b(q, t_x, t_z, c) = \mathbb{E}[q_{b,1} - q_{b,0} \mid q_{t_x,t_z} = q]$. Then:*

1. For each block, the minimiser of $\mathbb{E}\|v - (q_{b,1} - q_{b,0})\|^2$ over fields that see (q, t_x, t_z, c) is v_b.
2. For any absolutely continuous nondecreasing schedule $u \mapsto (t_x(u), t_z(u))$ with $t_x(0) = t_z(0) = 0$ and $t_x(1) = t_z(1) = 1$, the laws $p_{t_x(u),t_z(u)}$ satisfy the continuity equation with the master-clock field $f_u = \big(t_x'(u)\, v_x,\ t_z'(u)\, v_z\big)$ evaluated at $(q, t_x(u), t_z(u), c)$.

Proof. (1) is Theorem 3.2 with the conditioning variable enlarged to (q, t_x, t_z, c). (2) The state q_u is a function of $(q_0, q_1, t_x(u), t_z(u))$ with $\partial q_u/\partial t_b = q_{b,1} - q_{b,0}$. By the chain rule, $\mathrm{d}q_u/\mathrm{d}u = \big(t_x'(q_{x,1} - q_{x,0}),\ t_z'(q_{z,1} - q_{z,0})\big)$. For a test function φ, the tower property gives $\frac{\mathrm{d}}{\mathrm{d}u}\mathbb{E}\varphi(q_u) = \mathbb{E}[\nabla\varphi(q_u)^\top \mathrm{d}q_u/\mathrm{d}u] = \mathbb{E}[\nabla\varphi(q_u)^\top f_u(q_u)]$. □

Proposition 4.12 (3.4 — score from noise prediction). *Let $Q = \alpha Y + LZ$, where $Z \sim \mathcal{N}(0, I_D)$ is independent of Y and the matrix L is invertible. Then:*

1. $\nabla_q \log p_Q(q) = -L^{-\top}\, \mathbb{E}[Z \mid Q = q]$.
2. For a predictor $\widehat{Z}(q)$ of the noise, the induced score $\widehat{s} = -L^{-\top}\widehat{Z}$ satisfies $\|\widehat{s} - \nabla \log p_Q\| \le \|L^{-\top}\|_2\, \|\widehat{Z} - \mathbb{E}(Z \mid Q)\|$.
3. The denoising score loss $\mathbb{E}\|s_\theta(Q) + L^{-\top}Z\|^2$ equals $\mathbb{E}\|s_\theta(Q) - \nabla \log p_Q(Q)\|^2$ plus a constant.
4. If α is invertible, then $\mathbb{E}[Y \mid Q = q] = \alpha^{-1}\big(q + LL^\top \nabla \log p_Q(q)\big)$.

For an isotropic block such as the latent coordinates, $\alpha = t_z I$ and $L_z = (1 - t_z)I$, so $\nabla_{q_z} \log p = \dfrac{t_z v_z - q_z}{1 - t_z}$ with $v_z = \mathbb{E}[q_{z,1} - q_{z,0} \mid q]$. Under the shared translation noise of the coordinate block, $q_{x,0} \sim \mathcal{N}(0, \Sigma_0)$ with $\Sigma_0 = (I_n + c_d^2 \mathbf{1}_n \mathbf{1}_n^\top) \otimes I_3$ and $L_x L_x^\top = (1 - t_x)^2 \Sigma_0$, so $\nabla_{q_x} \log p = (1 - t_x)^{-1}\Sigma_0^{-1}(t_x v_x - q_x)$ with $\Sigma_0^{-1} = \big(I_n - \frac{c_d^2}{1 + n c_d^2}\mathbf{1}_n \mathbf{1}_n^\top\big) \otimes I_3$, while $\lambda_{\min}(\Sigma_0) = 1$ still gives $\|L_x^{-\top}\|_2 = 1/(1 - t_x)$. For the interpolant $q = \mathcal{I} + \gamma(u)\zeta$ one gets $\nabla \log p_u = -\mathbb{E}[\zeta \mid q]/\gamma(u)$.

Proof. (1) Given y, the kernel is $k(q \mid y) = \mathcal{N}(q; \alpha y, LL^\top)$, and $\nabla_q \log k(q \mid y) = -(LL^\top)^{-1}(q - \alpha y) = -L^{-\top} z$ with $z = L^{-1}(q - \alpha y)$. Since $p_Q = \int k(\cdot \mid y)\, p_Y(\mathrm{d}y)$, differentiation under the

integral (justified by dominated convergence) gives $\nabla \log p_Q(q) = \mathbb{E}[\nabla_q \log k(q \mid Y) \mid Q = q]$, which is the claim. (2) Subtract the two expressions for the score. (3) The target $T = -L^{-\top} Z$ has $\mathbb{E}[T \mid Q] = \nabla \log p_Q(Q)$ by (1), so Lemma 3.1 splits the loss into the score error and a term free of s_θ. (4) From (1), $\mathbb{E}[LZ \mid q] = -LL^\top \nabla \log p_Q(q)$, and $\mathbb{E}[LZ \mid q] = q - \alpha \mathbb{E}[Y \mid q]$. For the paired path, $\mathbb{E}[Z_b \mid q] = q_b - t_b v_b$ because $q_b = Z_b + t_b(Y_b - Z_b)$. □

The norm $\|L^{-\top}\|_2 = 1/(1-t_b)$ tends to infinity as $t_b \to 1$. A bounded error in the noise prediction does not bound the score error at the data end. Training can regress the noise without a singular target. A sampler then evaluates the score only for $t_b \le 1 - \delta$ and records separately whether it integrates an analytic limit at the end or targets an endpoint smoothed at scale δ. Optimal-transport couplings between noise and data change the Gaussian kernel used in the proof, so item 1 applies to independent coupling only.

Theorem 4.13 (3.5 — shared marginals of an SDE and an ODE). *Let* $\mathrm{d}Q_u = b(Q_u, u)\,\mathrm{d}u + \Lambda(u)\,\mathrm{d}W_u$ *with* W *a* D*-dimensional Brownian motion and* $\Xi(u) = \Lambda\Lambda^\top$ *positive definite and independent of the state. Let the density* $p_u > 0$ *be twice continuously differentiable, let* $s_u = \nabla \log p_u$*, and assume that the Fokker–Planck equation has a unique solution in the class containing* p_u*. Then:*

1. With $v = b - \frac{1}{2}\Xi s_u$, the density obeys $\partial_u p_u = -\nabla \cdot (v p_u)$. If v is locally Lipschitz in the state with linear growth, the continuity equation has a unique solution, and the ODE $\dot q = v(q, u)$ started from p_0 has the marginals p_u.
2. Conversely, let a field v transport the density path p_u. For every diffusion matrix $\Xi(u) \succeq 0$ independent of the state, with a dispersion $\Lambda(u)$ such that $\Lambda\Lambda^\top = \Xi$, the SDE with drift $v + \frac{1}{2}\Xi s_u$ has the marginals p_u. For $\Xi = 2\epsilon(u) I$ this is the SDE with drift $v + \epsilon \nabla \log p_u$ and noise $\sqrt{2\epsilon}\,\mathrm{d}W$.

Proof. The Fokker–Planck equation is $\partial_u p = -\nabla \cdot (bp) + \frac{1}{2}\sum_{ij} \Xi_{ij} \partial_{ij} p$. Since Ξ does not depend on the state, $\frac{1}{2}\sum_{ij} \Xi_{ij}\partial_{ij} p = \nabla \cdot (\frac{1}{2}\Xi \nabla p) = \nabla \cdot (\frac{1}{2}\Xi s\, p)$, using $\nabla p = sp$. So $\partial_u p = -\nabla \cdot ((b - \frac{1}{2}\Xi s)p)$, which is (1). For (2), write the Fokker–Planck equation of the SDE for a candidate p': $\partial_u p' = -\nabla \cdot ((v + \frac{1}{2}\Xi s_u)p') + \nabla \cdot (\frac{1}{2}\Xi \nabla p')$. For $p' = p_u$ the two terms with Ξ cancel because $s_u p_u = \nabla p_u$, and the equation reduces to $\partial_u p_u = -\nabla \cdot (v p_u)$, which holds. Uniqueness gives $p' = p_u$. □

The marginals agree and the path laws do not. The SDE has positive quadratic variation and explores across barriers. The ODE has bounded variation and defines a deterministic map that can be inverted and replayed.

Theorem 4.14 (3.6 — time reversal; the Markov property is cited). *Under the conditions of Theorem 3.5, assume in addition that* b *is continuous in* u *and locally Lipschitz with linear growth, that* Ξ *is continuously differentiable and uniformly positive definite, and that* $p_u|b|$ *and* $\Xi \nabla p_u$ *vanish at infinity. Then* $Y_r = Q_{1-r}$ *is a Markov diffusion with dispersion* $\Lambda(1-r)$ *and drift*

$$\bar b(y, r) = -b(y, 1-r) + \Xi(1-r)\,\nabla \log p_{1-r}(y).$$

Equivalently, run backward in u, $\mathrm{d}Q_u = \big[b(Q_u, u) - \Xi(u)\nabla \log p_u(Q_u)\big]\mathrm{d}u + \Lambda(u)\,\mathrm{d}\widetilde{W}_u$.

Derivation of the drift. Write the Fokker–Planck equation as a conservation law, $\partial_u p = -\nabla \cdot j$, with probability current $j = pb - \frac{1}{2}\Xi \nabla p = p(b - \frac{1}{2}\Xi s)$. For $\bar p_r = p_{1-r}$ one has $\partial_r \bar p_r = \nabla \cdot j_{1-r}$. A diffusion with dispersion $\Lambda(1-r)$ and drift $\bar b$ has current $\bar j = \bar p \bar b - \frac{1}{2}\Xi(1-r)\nabla \bar p$. The reversed process carries

the opposite current, $\bar{j}_r = -j_{1-r}$. Dividing by the positive density gives $\bar{b} - \frac{1}{2}\Xi s = -b + \frac{1}{2}\Xi s$, hence $\bar{b} = -b + \Xi s$. That the time-reversed process is a Markov diffusion with these coefficients is Anderson's theorem [38], which is used as stated (cited).

Theorem 4.15 (3.7 — density of a continuous flow). *Let $f_u(q)$ be continuous in u and continuously differentiable in q, and let its flow ψ_u be a diffeomorphism on the region of interest. Then*

$$\det D\psi_u(q_0) = \exp\left\{\int_0^u \nabla \cdot f_s(\psi_s(q_0))\,\mathrm{d}s\right\} > 0.$$

If f pushes p_0 to p_u, then $\log p_1(\psi_1(q_0)) = \log p_0(q_0) - \int_0^1 \nabla \cdot f_u(\psi_u(q_0))\,\mathrm{d}u$.

Proof. The Jacobian $J_u = D\psi_u(q_0)$ solves $\dot{J} = Df\,J$ with $J_0 = I$. Jacobi's formula gives $\frac{\mathrm{d}}{\mathrm{d}u}\log|\det J| = \operatorname{tr}(J^{-1}\dot{J}) = \operatorname{tr} Df = \nabla \cdot f$. The determinant cannot cross zero while ψ_u is a diffeomorphism, so it stays positive. Integration gives the first claim, and the change of variables $p_1(\psi_1(q_0))\det J_1 = p_0(q_0)$ gives the second. □

Three remarks complete the theorem. First, the packed field is $f_u = (t'_x v^x,\, t'_z v^z)$, and its divergence is $t'_x \operatorname{tr}(\partial v^x/\partial y_x) + t'_z \operatorname{tr}(\partial v^z/\partial z_e)$, because the cross blocks lie off the diagonal and add nothing to the trace. Second, the value is a density in the dimensionless chart. Rescaling the coordinates by ℓ_0 and each latent channel by $\sigma_{z,r}$ multiplies the density by a constant Jacobian, so that

$$\log p^{\mathrm{phys}}(y_x, z_e) = \log p(q) - d_x \log \ell_0 - n_e \sum_{r=1}^{8} \log \sigma_{z,r}.$$

Third, a chart that changes with the clock, $y = \Phi_u(q)$, has the velocity $D\Phi_u\, f + \partial_u \Phi_u$ and a time-dependent Jacobian. The Methods fix the chart during one solve, so that no term $\partial_u \Phi_u$ is omitted.

Proposition 4.16 (3.8 — variance of trace probes). *Let $J \in \mathbb{R}^{D\times D}$, $S = \frac{1}{2}(J + J^\top)$, and let ξ_{H} have mean zero and identity covariance. Then $\mathbb{E}[\xi_{\mathrm{H}}^\top J \xi_{\mathrm{H}}] = \operatorname{tr} J$. For a standard Gaussian probe, $\operatorname{Var}(\xi_{\mathrm{H}}^\top J \xi_{\mathrm{H}}) = 2\|S\|_{\mathrm{F}}^2$. For a Rademacher probe, $\operatorname{Var}(\xi_{\mathrm{H}}^\top J \xi_{\mathrm{H}}) = 2\sum_{i\neq j} S_{ij}^2$. The mean of K independent probes has $1/K$ of these variances.*

Proof. The expectation is $\operatorname{tr}(J\,\mathbb{E}[\xi_{\mathrm{H}}\xi_{\mathrm{H}}^\top]) = \operatorname{tr} J$. Since $\xi^\top(J - J^\top)\xi = 0$, the quadratic form equals $\xi^\top S \xi$. For a Gaussian probe, Isserlis' identity gives $\mathbb{E}(\xi^\top S\xi)^2 = (\operatorname{tr} S)^2 + 2\operatorname{tr}(S^2)$, and subtracting $(\operatorname{tr} S)^2$ leaves $2\|S\|_{\mathrm{F}}^2$. For a Rademacher probe, $\xi_i^2 = 1$, so the diagonal terms are constant. The off-diagonal terms $\sum_{i\neq j} S_{ij}\xi_i\xi_j$ are pairwise uncorrelated with variance $2\sum_{i\neq j} S_{ij}^2$. Independence divides the variance by K. □

For a diagonal J one Rademacher probe returns $\operatorname{tr} J$ exactly, while a Gaussian probe has variance $2\sum_i \lambda_i^2$.

Proposition 4.17 (3.9 — error of a computed log density). *Let f be L_f-Lipschitz in the state and let $d = \nabla \cdot f$ be L_d-Lipschitz, both uniformly in u. Let a numerical trajectory $\widehat{q}(u)$ with $\widehat{q}(0) = q_0$ satisfy $\|\dot{\widehat{q}} - f(\widehat{q}, u)\| \le \varepsilon_f$, and let a divergence estimate satisfy $|\widehat{d} - d| \le \varepsilon_d$. Then*

$$\left| \int_0^1 \widehat{d}(\widehat{q}(u), u)\,\mathrm{d}u - \int_0^1 d(q(u), u)\,\mathrm{d}u \right| \le \varepsilon_d + L_d\,\varepsilon_f\,\Phi(L_f),$$

where $\Phi(L_f) = (e^{L_f} - 1 - L_f)/L_f^2$ for $L_f > 0$ and $\Phi(0) = \frac{1}{2}$.

Proof. Let $e = \widehat{q} - q$ and $r = \dot{\widehat{q}} - f(\widehat{q}, u)$. Then $\|r\| \le \varepsilon_f$ and $D^+\|e\| \le L_f\|e\| + \varepsilon_f$, so Grönwall's inequality gives $\|e(u)\| \le \varepsilon_f(e^{L_f u} - 1)/L_f$. For each u, $|\widehat{d}(\widehat{q}) - d(q)| \le |\widehat{d}(\widehat{q}) - d(\widehat{q})| + |d(\widehat{q}) - d(q)| \le \varepsilon_d + L_d\|e\|$. Integrate over $[0, 1]$, using $\int_0^1 (e^{L_f u} - 1)\,\mathrm{d}u = (e^{L_f} - 1 - L_f)/L_f$. The case $L_f = 0$ integrates u to $\frac{1}{2}$. □

A guided field $\widetilde{f} = f_\theta + \kappa\, G_\phi^{-1}\nabla U$ defines its own flow, and Theorem 3.7 applies to it with the divergence of the complete assembled field, including the guidance term. The endpoint law is the pushforward of p_0 under that flow. It does not equal the base law tilted by e^U unless the guidance term is derived from an exact Doob transform of a stated path measure. Every statement of this supplement concerns the pushforward law. Raw log density values of guided and unguided runs therefore need the same reference measure before they are compared.

4.4. Transport and particle limits

This section treats two tools. The first is a static soft assignment of contact mass between antibody and epitope residues, a convex problem with a unique solution (Theorems 4.2 and 4.3, Propositions 4.4 to 4.6). The second is a particle method that steers a stochastic sampler while it keeps an account of the weights (Theorems 4.7, 4.10 and 4.14 and the results between them). A contact plan is not a generative law, and a high effective sample size does not show that the contact optimisation has converged. Figure 2 and Figure 3 illustrate the contact plan and the particle weights.

Contact plans from unbalanced transport. Let $\mu_R \in \mathbb{R}^{n_R}_{>0}$ and $\mu_E \in \mathbb{R}^{n_E}_{>0}$ be the reference masses of the antibody residues and the epitope residues, and let C be a finite cost matrix. Write $\rho = \rho_{\mathrm{OT}}$ and $\varepsilon = \varepsilon_{\mathrm{OT}}$, both positive. For a plan $\Gamma \ge 0$ with marginals $\Gamma\mathbf{1}$ and $\Gamma^\top\mathbf{1}$ the objective is

$$\mathcal{F}(\Gamma) = \langle C, \Gamma\rangle + \rho\,\mathrm{KL}(\Gamma\mathbf{1} \mid \mu_R) + \rho\,\mathrm{KL}(\Gamma^\top\mathbf{1} \mid \mu_E) + \varepsilon\,\mathrm{KL}(\Gamma \mid \mu_R \otimes \mu_E),$$

and $\mathrm{OT}_{\varepsilon,\rho} = \min_{\Gamma \ge 0} \mathcal{F}(\Gamma)$. The marginal terms let a residue that has no good partner send out less mass than its reference, at a price set by ρ. The last term spreads mass over near-equal poses and makes the problem strictly convex.

Debiased divergence. For two measures α and β on one space, with one shared cost, the debiased divergence is

$$S_{\varepsilon,\rho}(\alpha, \beta) = \mathrm{OT}_{\varepsilon,\rho}(\alpha, \beta) - \tfrac{1}{2}\mathrm{OT}_{\varepsilon,\rho}(\alpha, \alpha) - \tfrac{1}{2}\mathrm{OT}_{\varepsilon,\rho}(\beta, \beta) + \tfrac{\varepsilon}{2}\big(m(\alpha) - m(\beta)\big)^2,$$

where m is the total mass [4]. The last term vanishes when the two masses are equal.

Lemma 4.18 (4.1 — two elementary facts). *(a) For $u \ge 0$ with sum s and $v > 0$ with sum m, $\mathrm{KL}(u \mid v) \ge s\log(s/m) - s + m$. (b) For $y > 0$, $\lambda > 0$ and $\phi_y(x) = x\log(x/y) - x + y$ on $x \ge 0$, $\inf_{x \ge 0}\{\lambda\phi_y(x) - \sigma x\} = -\lambda y\,(e^{\sigma/\lambda} - 1)$, with the unique minimiser $x^\star = ye^{\sigma/\lambda}$.*

Proof. (a) Jensen's inequality for the convex function $x \log x$ with weights v_i/m gives $\sum_i v_i\, \phi(u_i/v_i) \geq m\, \phi(s/m)$ with $\phi(x) = x \log x$. Rearranging gives the claim. (b) The function is strictly convex with derivative $\lambda \log(x/y) - \sigma$, which tends to $-\infty$ at $x = 0$. So the minimiser is interior and solves $\lambda \log(x/y) = \sigma$. Substitution gives $\lambda(y - x^\star) = -\lambda y(e^{\sigma/\lambda} - 1)$. □

Theorem 4.19 (4.2 — existence, uniqueness and positivity). *The problem* $\min_{\Gamma \geq 0} \mathcal{F}(\Gamma)$ *has a unique minimiser* $\Gamma^\star$*, and every entry of* $\Gamma^\star$ *is strictly positive.*

Proof. The objective is continuous on the closed orthant under the convention $0 \log 0 = 0$. It is strictly convex, because the entropic term is strictly convex when every reference entry $\mu_{R,i}\mu_{E,j}$ is positive, and the other terms are convex. For coercivity put $s = \sum_{ij} \Gamma_{ij}$ and $m_R = \sum_i \mu_{R,i}$. By Lemma 4.1(a) applied to the row marginal, $\mathcal{F}(\Gamma) \geq (\min_{ij} C_{ij})\, s + \rho\, [s \log(s/m_R) - s + m_R]$, which tends to $+\infty$ with s because the term $\rho s \log s$ beats any linear growth. Hence a minimiser exists, and strict convexity makes it unique.

For positivity, suppose $\Gamma_{ij} = 0$ and raise that entry by $t > 0$. Let r_i and c_j be its row and column masses. The change of $\mathcal{F}$ is $[\varepsilon + \rho\, \mathbf{1}\{r_i = 0\} + \rho\, \mathbf{1}\{c_j = 0\}]\, t \log t + O(t)$ as $t \downarrow 0$. The bracket is positive and $t \log t < 0$, so the change is negative for small t. This contradicts minimality. □

Strict positivity says that the optimal plan is dense. It does not say that every antibody residue touches every epitope residue. A discrete contact map is extracted by a threshold or a matching, and a contact claim must report how it depends on the threshold and on the retained mass $\sum_{ij} \Gamma^\star_{ij}$.

Theorem 4.20 (4.3 — dual problem and scaling equations). *With* $\tau_{\mathrm{OT}} = \rho/(\rho + \varepsilon)$*, strong duality holds:*

$$\mathrm{OT}_{\varepsilon,\rho} = \max_{\mathbf{a},\mathbf{b}} \Big\{ -\rho \sum_i \mu_{R,i} \big(e^{-a_i/\rho} - 1\big) - \rho \sum_j \mu_{E,j} \big(e^{-b_j/\rho} - 1\big) - \varepsilon \sum_{ij} \mu_{R,i}\mu_{E,j} \big(e^{(a_i + b_j - C_{ij})/\varepsilon} - 1\big) \Big\}.$$

The dual objective is strictly concave, and at its maximiser $(\mathbf{a}^\star, \mathbf{b}^\star)$ the primal plan is $\Gamma^\star_{ij} = \mu_{R,i}\mu_{E,j} \exp\{(a^\star_i + b^\star_j - C_{ij})/\varepsilon\}$. The maximiser satisfies

$$a_i = -\tau_{\mathrm{OT}} \varepsilon \log \sum_j \mu_{E,j} \exp\Big\{ \frac{b_j - C_{ij}}{\varepsilon} \Big\}, \qquad b_j = -\tau_{\mathrm{OT}} \varepsilon \log \sum_i \mu_{R,i} \exp\Big\{ \frac{a_i - C_{ij}}{\varepsilon} \Big\}.$$

Proof. Introduce free marginal variables $r = \Gamma \mathbf{1}$ and $c = \Gamma^\top \mathbf{1}$ with multipliers $\mathbf{a}$ and $\mathbf{b}$, and form the Lagrangian $\mathcal{F}$ with the terms $\mathbf{a}^\top(r - \Gamma\mathbf{1}) + \mathbf{b}^\top(c - \Gamma^\top \mathbf{1})$ added. Minimise over each r_i, each c_j and each Γ_{ij} separately, using Lemma 4.1(b) with $(\lambda, \sigma) = (\rho, -a_i)$, $(\rho, -b_j)$ and $(\varepsilon, a_i + b_j - C_{ij})$. This gives the dual objective, and the minimiser over Γ_{ij} gives the plan. The primal problem is convex with finite value and a feasible point in the interior of its domain, so there is no duality gap. Each dual term is strictly concave in its own variable, so the objective is strictly concave. Setting $\partial/\partial a_i$ to zero gives $\mu_{R,i} e^{-a_i/\rho} = \sum_j \mu_{R,i}\mu_{E,j} e^{(a_i + b_j - C_{ij})/\varepsilon}$. Take logarithms and use $1/\rho + 1/\varepsilon = 1/(\tau_{\mathrm{OT}}\varepsilon)$. The equation for b_j is identical. □

The two equations are the log-domain form of generalised Sinkhorn scaling. The sums are evaluated with the shifted log-sum-exp of Section 4.8, and convergence is certified by the gap between primal and dual values. For one site, $n_R = n_E = 1$, stationarity reads $C + \rho \log(\gamma/\mu_R) + \rho \log(\gamma/\mu_E) +$

$\varepsilon \log\{\gamma/(\mu_R\mu_E)\} = 0$, so $\gamma^\star = \exp\{[(\rho+\varepsilon)(\log\mu_R + \log\mu_E) - C]/(2\rho+\varepsilon)\}$. For $\mu_R = \mu_E = 1$ the optimal value is $(2\rho+\varepsilon)(1-\gamma^\star)$.

Proposition 4.21 (4.4 — envelope gradient). *Let $C(\vartheta)$, $\mu_R(\vartheta)$ and $\mu_E(\vartheta)$ be differentiable, with the assumptions of Theorem 4.2 on a neighbourhood. Let $r = \Gamma^\star\mathbf{1}$, $c = \Gamma^{\star\top}\mathbf{1}$, $m_R = \sum_i \mu_{R,i}$ and $m_E = \sum_j \mu_{E,j}$. Then $T(\vartheta) = \mathrm{OT}_{\varepsilon,\rho}$ is differentiable and*

$$\nabla_\vartheta T = \langle \Gamma^\star, \nabla_\vartheta C\rangle + \sum_i \Big[\rho\Big(1 - \frac{r_i}{\mu_{R,i}}\Big) + \varepsilon\Big(m_E - \frac{r_i}{\mu_{R,i}}\Big)\Big]\nabla_\vartheta \mu_{R,i} + \sum_j \Big[\rho\Big(1 - \frac{c_j}{\mu_{E,j}}\Big) + \varepsilon\Big(m_R - \frac{c_j}{\mu_{E,j}}\Big)\Big]\nabla_\vartheta \mu_{E,j}.$$

For constant reference masses, $\nabla_\vartheta T = \langle \Gamma^\star, \nabla_\vartheta C\rangle$. Note that $r_i/\mu_{R,i} = e^{-a_i^\star/\rho}$.

Proof. The value is $T(\vartheta) = \min_\Gamma \mathcal{F}(\Gamma;\vartheta)$. The minimiser is unique and, by strict convexity and local uniform coercivity, depends continuously on ϑ. Danskin's theorem therefore gives $\nabla_\vartheta T = \partial_\vartheta \mathcal{F}(\Gamma^\star;\vartheta)$. For fixed Γ: $\partial\mathcal{F}/\partial C_{ij} = \Gamma_{ij}$; $\partial\,\mathrm{KL}(r \mid \mu_R)/\partial\mu_{R,i} = 1 - r_i/\mu_{R,i}$; and $\partial\,\mathrm{KL}(\Gamma \mid \mu_R\otimes\mu_E)/\partial\mu_{R,i} = \sum_j(\mu_{E,j} - \Gamma_{ij}/\mu_{R,i}) = m_E - r_i/\mu_{R,i}$. The same holds for μ_E, and the second marginal term does not involve μ_R. □

Proposition 4.22 (4.5 — regularisation gap). *Let $H(\Gamma) = \mathrm{KL}(\Gamma \mid \mu_R\otimes\mu_E)$, let Γ_ε minimise $\mathcal{F}$, and let Γ_0 be any minimiser of $\mathcal{F} - \varepsilon H$ (it exists by the coercivity argument of Theorem 4.2). Then*

$$0 \le \varepsilon H(\Gamma_\varepsilon) \le \mathrm{OT}_{\varepsilon,\rho} - \mathrm{OT}_{0,\rho} \le \varepsilon H(\Gamma_0).$$

Proof. Put $\mathcal{F}_0 = \mathcal{F} - \varepsilon H$, so that $\mathcal{F} = \mathcal{F}_0 + \varepsilon H$ and $H \ge 0$. Optimality of Γ_0 gives $\mathcal{F}_0(\Gamma_\varepsilon) \ge \mathrm{OT}_{0,\rho}$, so $\mathrm{OT}_{\varepsilon,\rho} = \mathcal{F}_0(\Gamma_\varepsilon) + \varepsilon H(\Gamma_\varepsilon) \ge \mathrm{OT}_{0,\rho} + \varepsilon H(\Gamma_\varepsilon)$. Optimality of Γ_ε gives $\mathrm{OT}_{\varepsilon,\rho} \le \mathcal{F}(\Gamma_0) = \mathrm{OT}_{0,\rho} + \varepsilon H(\Gamma_0)$. $H(\Gamma_0)$ is finite because Γ_0 is bounded and the reference is positive. □

The bound grows with the number of reference pairs, so ε must shrink as the interface grows to hold a given gap.

Proposition 4.23 (4.6 — soft plans and integer contacts). *(a) A binary contact map with given integer valences for each residue and exactly k contacts that minimises a linear cost exists whenever one is feasible, because it is a minimum-cost flow problem, and flow problems with integer capacities have integer optima (standard). This controls the linear part of the objective. (b) No bound on the distance between $\Gamma^\star$ and a binary map follows from ε, ρ and $\|C\|_\infty$ alone: take $n_R = n_E = n$, $\mu_R = \mu_E = \mathbf{1}/n$ and $C = 0$. Then $\Gamma^\star_{ij} = 1/n^2$ for all $\rho, \varepsilon > 0$, and for every $S \in \{0,1\}^{n\times n}$ with $|S|$ unit entries,*

$$\|\Gamma^\star - S\|_1 = |S| + 1 - \frac{2|S|}{n^2} \;\ge\; 1 = \|\Gamma^\star\|_1.$$

A permutation matrix gives $n + 1 - 2/n$. A permutation scaled to the plan's mass, $\frac{1}{n}P$, gives $2(1-1/n)$, which tends to the largest possible value 2.

Proof. (b) The first-order condition of $\mathcal{F}$ is $C_{ij} + \varepsilon\log\frac{\Gamma_{ij}}{\mu_{R,i}\mu_{E,j}} + \rho\log\frac{r_i}{\mu_{R,i}} + \rho\log\frac{c_j}{\mu_{E,j}} = 0$. At $\Gamma_{ij} = 1/n^2$ every ratio equals 1, so all logarithms vanish and $C = 0$ holds. By uniqueness this is the minimiser. Entries with $S_{ij} = 1$ contribute $1 - 1/n^2$ each, and the other $n^2 - |S|$ entries

contribute $1/n^2$ each, which gives the formula. The bound $|S|(1-2/n^2) \geq 0$ holds for $n \geq 2$. For $\frac{1}{n}P$ the n unit positions contribute $(n-1)/n^2$ each and the other n^2-n entries $1/n^2$ each. $\square$

Particle systems for a guided stochastic sampler. Fix a horizon M and a state space $\mathcal{X}$. Let p_0 be the initial law and let $\mathsf{K}_j(q_{j-1}, \mathrm{d}q_j)$ be the reference transition kernels, $j = 1, \ldots, M$. Let $\Psi_0 : \mathcal{X} \to (0, C_\Psi]$ and $\Psi_j : \mathcal{X}^2 \to (0, C_\Psi]$ be bounded positive potentials. The unnormalised path measure, its total mass and its normalised form are

$$\gamma_M(\mathrm{d}q_{0:M}) = p_0(\mathrm{d}q_0)\Psi_0(q_0)\prod_{j=1}^{M}\mathsf{K}_j(q_{j-1}, \mathrm{d}q_j)\Psi_j(q_{j-1}, q_j), \qquad Z_M = \gamma_M(1), \qquad \eta_M = \frac{\gamma_M}{Z_M}.$$

The default of the Methods is $\Psi_0 \equiv 1$, which this supplement keeps as a special case. A proposal law R_0 and proposal kernels R_j replace the reference laws in the simulation. They must dominate: $p_0 \ll \mathsf{R}_0$ and $\mathsf{K}_j(q, \cdot) \ll \mathsf{R}_j(q, \cdot)$. The incremental weights are

$$\omega_0 = \Psi_0 \frac{\mathrm{d}p_0}{\mathrm{d}\mathsf{R}_0}(q_0), \qquad \omega_j = \Psi_j(q_{j-1}, q_j)\frac{\mathrm{d}\mathsf{K}_j(q_{j-1}, \cdot)}{\mathrm{d}\mathsf{R}_j(q_{j-1}, \cdot)}(q_j).$$

If a proposal assigns zero probability to a measurable set with positive reference probability, the domination condition fails and importance weighting cannot recover the missing mass. Proposal draws outside the reference support receive zero weight.

The system has N_p particles. Initial particles $q_0^{(k)}$ are drawn independently from R_0 with carried weights $\widetilde{w}_{-1}^{(k)} = 1/N_\mathrm{p}$. Before each step $j \geq 1$ the system either resamples or holds, and this choice may depend on everything observed up to step $j-1$. After a resampling step the ancestors a_k have offspring counts N_i with $\mathbb{E}[N_i \mid \mathcal{G}_{j-1}] = N_\mathrm{p}\bar{w}_{j-1}^{(i)}$ (conditional unbiasedness), and the carried weights are $\widetilde{w}_{j-1}^{(k)} = 1/N_\mathrm{p}$. After a held step, $a_k = k$ and $\widetilde{w}_{j-1}^{(k)} = \bar{w}_{j-1}^{(k)}$, the normalised weight. Then $q_j^{(k)} \sim \mathsf{R}_j(q_{j-1}^{(a_k)}, \cdot)$ and

$$w_j^{(k)} = \widetilde{w}_{j-1}^{(k)}\,\omega_j(q_{j-1}^{(a_k)}, q_j^{(k)}), \qquad \bar{\ell}_j = \sum_{k=1}^{N_\mathrm{p}} w_j^{(k)}, \qquad \bar{w}_j^{(k)} = \frac{w_j^{(k)}}{\bar{\ell}_j}, \qquad \widehat{Z}_M = \prod_{j=0}^{M}\bar{\ell}_j.$$

For $j = 0$ there is no ancestor and ω_0 depends on $q_0^{(k)}$ only. If some $\bar{\ell}_j = 0$, put $\widehat{Z}_M = 0$. The factor $\bar{\ell}_j$ carries the previous weights. After a resampling step it reduces to the plain mean $N_\mathrm{p}^{-1}\sum_k \omega_j$.

One proposal step followed by one resampling step leaves the expected weighted measure unchanged, and the proof is an induction on this fact.

Theorem 4.24 (4.7 — unbiasedness). *Under the assumptions above, for every $N_\mathrm{p} \geq 1$, every adaptive resampling schedule and every bounded measurable path function φ,*

$$\mathbb{E}\Big[\widehat{Z}_M \sum_{k=1}^{N_\mathrm{p}} \bar{w}_M^{(k)}\,\varphi\big(q_{0:M}^{(k)}\big)\Big] = \gamma_M(\varphi),$$

and in particular,

$$\mathbb{E}\big[\widehat{Z}_M\big] = Z_M.$$

Here $q_{0:M}^{(k)}$ is the ancestral path of particle k. Negative association of the offspring counts is not needed.

Proof. Put $\widehat{\gamma}_j(\varphi) = \widehat{Z}_j \sum_k \bar{w}_j^{(k)} \varphi(q_{0:j}^{(k)})$ and take $\varphi \geq 0$ first. For $j = 0$, $\mathbb{E}\widehat{\gamma}_0(\varphi) = \int \mathsf{R}_0(\mathrm{d}q_0)\, \omega_0\, \varphi = \int p_0(\mathrm{d}q_0)\Psi_0\varphi = \gamma_0(\varphi)$ by the Radon–Nikodym theorem. For $j \geq 1$ note that $\widehat{Z}_j \bar{w}_j^{(k)} = \widehat{Z}_{j-1} w_j^{(k)}$, so

$$\widehat{\gamma}_j(\varphi) = \widehat{Z}_{j-1} \sum_k \widetilde{w}_{j-1}^{(k)}\, \omega_j\big(q_{j-1}^{(a_k)}, q_j^{(k)}\big)\, \varphi\big(q_{0:j-1}^{(a_k)}, q_j^{(k)}\big).$$

Given the history and the ancestors, $q_j^{(k)} \sim \mathsf{R}_j$, and the change of measure gives $\int \mathsf{R}_j(q, \mathrm{d}q')\, \omega_j\, \varphi = \int \mathsf{K}_j(q, \mathrm{d}q')\Psi_j(q, q')\varphi =: (\mathsf{K}_j^{\Psi}\varphi)(q)$. So the conditional expectation is $\widehat{Z}_{j-1} \sum_k \widetilde{w}_{j-1}^{(k)} (\mathsf{K}_j^{\Psi}\varphi)(q_{0:j-1}^{(a_k)})$. If the step is held, this equals $\widehat{\gamma}_{j-1}(\mathsf{K}_j^{\Psi}\varphi)$. If the step resamples, the average over the offspring counts is $N_{\mathrm{p}}^{-1} \sum_i \mathbb{E}[N_i]\, F(q^{(i)}) = \sum_i \bar{w}_{j-1}^{(i)} F(q^{(i)})$ with $F = \mathsf{K}_j^{\Psi}\varphi$, which is again $\widehat{\gamma}_{j-1}(\mathsf{K}_j^{\Psi}\varphi)$.

The schedule is a function of the history, so the two cases combine. The induction hypothesis gives $\mathbb{E}\widehat{\gamma}_{j-1}(\mathsf{K}_j^{\Psi}\varphi) = \gamma_{j-1}(\mathsf{K}_j^{\Psi}\varphi) = \gamma_j(\varphi)$. For $\varphi \geq 0$, Tonelli's theorem justifies every exchange of sum and integral even before finiteness is known, and $\gamma_j(1) \leq C_{\Psi}^{j+1}$ is finite. A signed bounded φ is split into positive and negative parts. The choice $\varphi = 1$ gives the last claim, with the convention $\widehat{Z}_M = 0$ whenever a step factor vanishes. □

The logarithm of $\widehat{Z}_M$ is biased low, by Jensen's inequality, and a ratio such as $\widehat{\eta}_M(\varphi)$ is biased at finite N_{p}. The next two examples show how large the effects can be and why the carried weights matter.

Example 4.8 (carried weights). Let p_0 be uniform on $\{0, 1\}$, let the kernels and proposals be the identity map, let $\Psi_0 \equiv 1$ and $\Psi_1 = \Psi_2 = (1, 2)$, and let $N_{\mathrm{p}} = 2$ with no resampling. The four equally likely initial clouds give products of plain means 1, 2.25, 2.25, 4, with mean 2.375. The target is $\mathbb{E}[\Psi_1\Psi_2] = 2.5$. The carried-weight estimator returns 1, 2.5, 2.5, 4 with mean 2.5.

Example 4.9 (exact finite-particle bias). Let $X \in \{0, 1\}$ be equally likely, with potential $\Psi(0) = 1$, $\Psi(1) = 3$ and the reference kernel as proposal. With n independent particles and $K \sim \mathrm{Binomial}(n, \frac{1}{2})$ the number with $X = 1$, the estimates are $\widehat{Z} = 1 + 2K/n$ and $\widehat{\pi}(1) = 3K/(n + 2K)$. The truth is $Z = 2$ and $\pi(1) = 3/4$. Exact summation over the $n + 1$ outcomes gives Table 3. The estimate $\widehat{Z}$ is unbiased, while its logarithm and the ratio are biased low, and both biases shrink about like $1/n$. Averaging independent runs leaves these biases unchanged, so a particle budget cannot be split into small runs without cost.

Table 3: Exact bias of the normalising-constant estimate, of its logarithm and of the ratio estimate in Example 4.9. The values come from exact summation over the outcomes.

n	$\mathbb{E}\widehat{Z}$	$\log 2 - \mathbb{E}\log\widehat{Z}$	$\frac{3}{4} - \mathbb{E}\widehat{\pi}(1)$
1	2	0.143841	0.250000
4	2	0.034115	0.056250
16	2	0.007997	0.012287
64	2	0.001965	0.002964

Define the backward information function by $h_M \equiv 1$ and $h_{j-1}(q) = \int \mathsf{K}_j(q, \mathrm{d}q')\Psi_j(q, q')h_j(q')$ for $j = M, \ldots, 1$. Since the potentials are positive and bounded, $0 < h_j \leq C_{\Psi}^{M-j}$. The twisted kernels and potentials are

$$\mathsf{K}_j^h(q,\mathrm{d}q') = \frac{\mathsf{K}_j(q,\mathrm{d}q')\Psi_j(q,q')h_j(q')}{h_{j-1}(q)}, \qquad \Psi_0^h = \Psi_0 h_0, \qquad \Psi_j^h(q,q') = \Psi_j(q,q')\frac{h_j(q')}{h_{j-1}(q)}.$$

Theorem 4.25 (4.10 — zero variance under perfect twisting). *With these definitions,* $Z_M = \int p_0(\mathrm{d}q_0)\Psi_0(q_0)h_0(q_0)$, *and* $\Psi_0^h \prod_{j=1}^M \Psi_j^h = \Psi_0 \prod_{j=1}^M \Psi_j$ *along every path, so* γ_M *and* Z_M *are unchanged. Run the sampler for the potentials* Ψ^h *with the initial proposal* $\mathsf{R}_0(\mathrm{d}q_0) = p_0(\mathrm{d}q_0)\Psi_0(q_0)h_0(q_0)/Z_M$ *and the proposals* $\mathsf{R}_j = \mathsf{K}_j^h$. *The weights are constant:*

$$\omega_0(q_0^{(k)}) \equiv Z_M, \qquad \omega_j(q_{j-1}^{(k)}, q_j^{(k)}) \equiv 1 \qquad (j = 1, \dots, M).$$

Consequently, for every $N_\mathrm{p} \geq 1$, every resampling rule and every schedule, $\widehat{Z}_M = Z_M$ on every realisation, and $\mathrm{Var}(\widehat{Z}_M) = 0$. All normalised weights equal $1/N_\mathrm{p}$. The equality holds for the explicit versions of h_j and of the Radon–Nikodym derivatives; for other versions it holds almost surely.

Proof. The recursion for h_j and positivity give finite positive ratios h_j/h_{j-1}. *Normalisation.* By the recursion, $\int \mathsf{K}_j(q,\mathrm{d}q')\Psi_j(q,q')h_j(q') = h_{j-1}(q)$, so $\mathsf{K}_j^h(q,\cdot)$ has mass 1. Integrating γ_M over $q_M, q_{M-1}, \dots, q_1$ in turn with the same recursion gives $Z_M = \int p_0\Psi_0 h_0$, so R_0 is a probability law. *Telescoping.* The factors h_j cancel in the product of twisted potentials, and $h_M \equiv 1$ removes the last one. *Derivatives.* $\mathsf{K}_j^h(q,\cdot)$ has the positive density $\Psi_j h_j/h_{j-1}$ with respect to $\mathsf{K}_j(q,\cdot)$, and R_0 has the positive density $\Psi_0 h_0/Z_M$ with respect to p_0. So the measures are mutually absolutely continuous, with $\mathrm{d}\mathsf{K}_j/\mathrm{d}\mathsf{K}_j^h = h_{j-1}/(\Psi_j h_j)$ and $\mathrm{d}p_0/\mathrm{d}\mathsf{R}_0 = Z_M/(\Psi_0 h_0)$. *Weights.* $\omega_0 = \Psi_0 h_0 \cdot Z_M/(\Psi_0 h_0) = Z_M$ and $\omega_j = (\Psi_j h_j/h_{j-1}) \cdot h_{j-1}/(\Psi_j h_j) = 1$. The carried weights sum to one, so $\bar{\ell}_0 = Z_M$ and $\bar{\ell}_j = 1$ for $j \geq 1$ whatever the schedule, and $\widehat{Z}_M = Z_M$. □

If the original potentials are used with the same proposals, the weights are not constant. The initial weight is $Z_M/h_0(q_0)$ and the step weights are $h_{j-1}(q_{j-1})/h_j(q_j)$. Their product along any path equals Z_M, so a sampler without resampling has equal cumulative weights. Resampling by a single factor adds randomness, and the variance of $\widehat{Z}_M$ is then positive in general (Example 4.11). Zero variance is a property of $\widehat{Z}_M$ alone. Without resampling the particles are an independent sample from η_M, and $\eta^N(\varphi)$ keeps the variance $\mathrm{Var}_{\eta_M}(\varphi)/N_\mathrm{p}$ for $\varphi \in L^2(\eta_M)$.

In practice h_j is out of reach, and a learned $\widehat{h}_j > 0$ with $\widehat{h}_M \equiv 1$ replaces it. The twisted potentials built from $\widehat{h}$ have the same product along every path, so the target is unchanged. Theorem 4.7 applies to them, and $\widehat{Z}_M$ stays unbiased for any $\widehat{h}$. The weights equal one only if $\widehat{h} = h$ and the proposals are K_j^h. That a closer $\widehat{h}$ gives a smaller variance is a tendency that this supplement leaves unproved.

Example 4.11 (resampling by the untwisted weights). Let $\mathcal{X} = \{a, b\}$, $M = 1$, p_0 uniform, $\Psi_0 \equiv 1$, $\mathsf{K}_1(q,\cdot) = \delta_q$, and $\Psi_1(q,q') = g(q)$ with $g(a) = 1$ and $g(b) = 3$. Then $h_1 \equiv 1$, $h_0 = g$, $Z_1 = 2$, $\mathsf{R}_0 = (\frac{1}{4}, \frac{3}{4})$ and $\mathsf{K}_1^h = \mathsf{K}_1$. The untwisted weights are $\omega_0 = Z_1/g = (2, \frac{2}{3})$ and $\omega_1 = g = (1, 3)$, with product 2 on both paths. Take $N_\mathrm{p} = 2$ and resample by multinomial resampling before step 1. The clouds (a, a) and (b, b) have probabilities $\frac{1}{16}$ and $\frac{9}{16}$, and in both $\widehat{Z} = 2$. The mixed cloud has probability $\frac{6}{16}$, $\bar{\ell}_0 = \frac{4}{3}$ and normalised weights $(\frac{3}{4}, \frac{1}{4})$. Its two offspring are (a, a), (a, b) or (b, b) with probabilities $\frac{9}{16}$, $\frac{6}{16}$, $\frac{1}{16}$, so $\bar{\ell}_1 = 1, 2, 3$ and $\widehat{Z} = \frac{4}{3}, \frac{8}{3}, 4$. Hence $\mathbb{E}\widehat{Z} = 2$, $\mathbb{E}\widehat{Z}^2 = \frac{17}{4}$ and $\mathrm{Var}\,\widehat{Z} = \frac{1}{4}$. With the twisted weights $\omega_0 = 2$ and $\omega_1 = 1$, the estimate equals 2 in every case.

Proposition 4.26 (4.12 — mean-unbiased resampling does not give consistency). *Let $\mathcal{X} = \{0, 1\}$, $M = 2$, p_0 uniform, $\Psi_0 \equiv \Psi_1 \equiv 1$, $\Psi_2(q, q') = g(q)$ with $g(0) = \frac{1}{3}$ and $g(1) = 1$, and let all kernels and proposals be the identity. Before every step the system draws one index I with probabilities $\bar{w}$ and gives all N_p offspring to it. Then, given the past, $\mathbb{E}[N_i] = N_\mathrm{p}\bar{w}^{(i)}$, $\mathbb{E}\widehat{Z}_2 = Z_2 = \frac{2}{3}$, and $\widehat{\eta}_2(\mathrm{id})$ is 0 or 1 with probability $\frac{1}{2}$ each. The target value is $\eta_2(\mathrm{id}) = \frac{3}{4}$, so the bias of the normalised estimate is exactly $-\frac{1}{4}$ for every N_p. The effective sample size equals N_p after every step.*

Proof. The offspring means hold by construction, and the claim on $\widehat{Z}_2$ is Theorem 4.7. After step 1 all particles equal $q_0^{(I)}$, where I is uniform because the weights are equal, so $q_0^{(I)}$ is 0 or 1 with probability $\frac{1}{2}$. At step 2 all particles carry the same weight $g(q_0^{(I)})$, so the normalised weights are uniform and $\widehat{\eta}_2(\mathrm{id}) = q_0^{(I)}$. Hence $\mathbb{E}\widehat{\eta}_2 = \frac{1}{2}$. The target is $\eta_2(\mathrm{id}) = \frac{1}{2} \cdot 1/(\frac{1}{2} \cdot \frac{1}{3} + \frac{1}{2} \cdot 1) = \frac{3}{4}$. All weights are equal at every step, so the effective sample size is N_p throughout. □

The example shows that conditional unbiasedness is too weak to support a claim about the weighted cloud, and that the effective sample size cannot detect such a failure.

Proposition 4.27 (4.13 — systematic resampling is not negatively associated). *Let $N_\mathrm{p} = 4$ and $\bar{w} = (1, 1, 1, 5)/8$. Systematic resampling draws $U \sim \mathcal{U}[0, \frac{1}{4})$ and uses the points $U + (i - 1)/4$. Then $N_1 = N_3 = \mathbf{1}\{U < \frac{1}{8}\}$, $N_2 = \mathbf{1}\{U \geq \frac{1}{8}\}$ and $N_4 = 2 + \mathbf{1}\{U \geq \frac{1}{8}\}$. So $\mathbb{P}(N_1 = 1, N_3 = 1) = \frac{1}{2}$ while $\mathbb{P}(N_1 = 1)\,\mathbb{P}(N_3 = 1) = \frac{1}{4}$, and $\mathrm{Cov}(N_1, N_3) = \frac{1}{4} > 0$.*

Proof. The cumulative weights are $(\frac{1}{8}, \frac{2}{8}, \frac{3}{8}, 1)$. The point U lies in $[0, \frac{1}{8})$ exactly when $U < \frac{1}{8}$ (particle 1) and otherwise in $[\frac{1}{8}, \frac{1}{4})$ (particle 2). The point $U + \frac{1}{4} \in [\frac{1}{4}, \frac{1}{2})$ lies in $[\frac{2}{8}, \frac{3}{8})$ (particle 3) exactly when $U < \frac{1}{8}$ and otherwise in $[\frac{3}{8}, \frac{1}{2})$ (particle 4). The points $U + \frac{1}{2}$ and $U + \frac{3}{4}$ lie in $[\frac{1}{2}, 1)$ (particle 4). □

The expected counts are right ($\mathbb{E}N_1 = \frac{1}{2} = 4\bar{w}_1$), so systematic resampling is conditionally unbiased, but its counts move together, and the consistency theorem below does not cover it for unordered particles. Consistency is recovered when the particles are ordered along a coordinate [26]. The Methods use no such ordering for the packed state, so a run that uses systematic resampling treats its weighted cloud as a diagnostic. The resampling laws with proved conditions are multinomial, residual and stratified resampling.

Theorem 4.28 (4.14 — consistency; sketch, with the resampling step cited). *Let the assumptions of Theorem 4.7 hold, let the Radon–Nikodym derivatives $\mathrm{d}p_0/\mathrm{d}\mathsf{R}_0$ and $\mathrm{d}\mathsf{K}_j/\mathrm{d}\mathsf{R}_j$ be bounded, and let $\mathcal{X}$ be a cubifiable subset of $\mathbb{R}^d$ (the condition on the state space in the cited theorem). Let the system resample before every step with a scheme that meets the conditions of Theorem 1 of [26] (multinomial, residual and stratified resampling do). Then for every bounded measurable φ, $\widehat{\eta}_M(\varphi) = \sum_k \bar{w}_M^{(k)} \varphi(q_M^{(k)}) \to \eta_M(\varphi)$ in probability as $N_\mathrm{p} \to \infty$.*

Sketch. At $j = 0$ the law of large numbers gives convergence of the weighted empirical measure for bounded functions. Assume convergence at $j - 1$. Selection replaces the empirical measure by its weighted version. The map from a measure η to $\eta(G\,\mathsf{K}\varphi)/\eta(G)$ is a ratio whose denominator converges to a positive constant, so the continuous mapping theorem gives convergence of the selected measure. The cited theorem gives that resampling preserves the limit. Given the ancestors, the mutations are conditionally independent with kernels K_j and weights ω_j, and the conditional law of large numbers gives the update. A finite induction reaches step M. The normalising constant

factors as a finite product of integrals under the previous laws, so each factor of $\widehat{Z}_M$ converges and so does the product.

The supplement stops at consistency, and a central limit theorem and an explicit constant for the bias of order $1/N_{\mathrm{p}}$ go beyond it. A bias bound of that order needs a bounded weight, which fails for most guided proposals, and Proposition 4.12 shows that a deviation condition on the offspring cannot be dropped.

Proposition 4.29 (4.15 — effective sample size). *For positive weights w_k with mean $\bar{w}$ and normalised weights $\bar{w}_k = w_k / \sum_l w_l$, let* $\mathrm{ESS} = 1/\sum_k \bar{w}_k^2$ *and* $\mathrm{CV}^2 = N^{-1}\sum_k (w_k - \bar{w})^2/\bar{w}^2$. *Then:*

1. $\mathrm{ESS} = N/(1+\mathrm{CV}^2)$.
2. $1 \le \mathrm{ESS} \le N$ and $1/\mathrm{ESS} \le \max_k \bar{w}_k \le 1/\sqrt{\mathrm{ESS}}$.
3. If $w_1, \dots, w_N$ are independent with $\log w_k \sim \mathcal{N}(\mu, s^2)$, then $\mathrm{ESS}/N \to e^{-s^2}$ in probability as $N \to \infty$. This is an asymptotic statement about large N for an idealised model.
4. Immediately after resampling the weights are all $1/N$ and $\mathrm{ESS} = N$, whatever the number of distinct ancestors. Resampling that copies one particle N times leaves $\mathrm{ESS} = N$ with one ancestor.

Proof. (1) Expanding the variance gives $N^{-1}\sum w_k^2 = \bar{w}^2(1+\mathrm{CV}^2)$, and $\sum_l w_l = N\bar{w}$, so $\sum_k \bar{w}_k^2 = (1+\mathrm{CV}^2)/N$. (2) By Cauchy–Schwarz, $1 = (\sum_k \bar{w}_k)^2 \le N\sum_k \bar{w}_k^2$, so $\mathrm{ESS} \le N$. Also $\sum_k \bar{w}_k^2 \le \max_k \bar{w}_k \sum_k \bar{w}_k = \max_k \bar{w}_k \le 1$, which gives $\mathrm{ESS} \ge 1$ and $1/\mathrm{ESS} \le \max_k \bar{w}_k$. Finally $\max_k \bar{w}_k^2 \le \sum_k \bar{w}_k^2$. (3) The law of large numbers gives $N^{-1}\sum_k w_k \to e^{\mu + s^2/2}$ and $N^{-1}\sum_k w_k^2 \to e^{2\mu + 2s^2}$, and ESS/N is the ratio of the square of the first to the second. (4) is the statement of Proposition 4.12 for the effective sample size. □

Because of item 4, the number of distinct ancestral lineages that survive from step 0 is recorded next to the effective sample size, together with the spread of the decoded structures and the variance of $\log \widehat{Z}_M$ across independent complete runs. These are diagnostics, and no theorem here turns them into an error bound. Independent complete runs give an empirical standard error that does not depend on the genealogy. The Methods plan at least 32 independent complete runs.

A deterministic guided flow is a separate case. Distinct deterministic kernels are point masses on different updates and are generally mutually singular, so the domination condition of Theorem 4.7 fails. The particle results above apply to a stochastic transition profile only.

4.5. Decision statistics and calibration

This section states the guarantees that the Methods use for ambiguity over plausible worlds, coverage, conformal calibration, exact count bounds and sequential evidence. Each guarantee holds under named conditions, which the text gives beside it. The results concern finite sets of numbers, weights and probabilities. Technical Report III (Figure on *KL-ball worst-case scoring, totally unimodular portfolio flow, Mondrian conformal risk control, and 53-gate sequential Ville e-process verification*) and Technical Report III (Figure on *A fixed-horizon test is fully specified before outcomes are read, and sequential evidence stays*) illustrate the cautious score under ambiguity and the exact-bound examples.

Ambiguity over plausible worlds. Let the worlds $j = 1, \dots, M$ be surrogate checkpoints or assay mappings, with reference weights $\pi_j > 0$ that sum to one. Each world gives a dimensionless

utility U_j of one design. The entropic pessimistic utility is $U_{\text{ent}} = -\eta^{-1} \log \sum_j \pi_j e^{-\eta U_j}$ for $\eta > 0$. Write Δ_M for the probability vectors of length M and $\mathrm{KL}(\nu \mid \pi) = \sum_j \nu_j \log(\nu_j/\pi_j)$.

Proposition 4.30 (5.1 — entropic identity). *Put $Z = \sum_j \pi_j e^{-\eta U_j}$ and $\nu_j^\star = \pi_j e^{-\eta U_j}/Z$. For every $\nu \in \Delta_M$,*

$$\sum_j \nu_j U_j + \frac{1}{\eta} \mathrm{KL}(\nu \mid \pi) = U_{\text{ent}} + \frac{1}{\eta} \mathrm{KL}(\nu \mid \nu^\star).$$

Hence $U_{\text{ent}} = \min_{\nu \in \Delta_M} \{\sum_j \nu_j U_j + \eta^{-1} \mathrm{KL}(\nu \mid \pi)\}$, the minimum is attained only at $\nu^\star$, and $\min_j U_j \le U_{\text{ent}} \le \sum_j \pi_j U_j$.

Proof. Expand $\mathrm{KL}(\nu \mid \nu^\star) = \sum_j \nu_j \log(\nu_j/\pi_j) + \eta \sum_j \nu_j U_j + \log Z$ and rearrange, using $U_{\text{ent}} = -\eta^{-1} \log Z$. Since $\mathrm{KL}(\nu \mid \nu^\star) \ge 0$ with equality only at $\nu = \nu^\star$, the minimum follows. The choice $\nu = \pi$ gives the upper bound. Also $Z \le e^{-\eta \min_j U_j}$, which gives the lower bound. □

Numerical illustration: for $\pi = (0.5, 0.5)$, $U = (0.9, 0.3)$ and $\eta = 2$, the formula gives $U_{\text{ent}} = 0.5149$ and $\nu^\star = (0.2315, 0.7685)$. The weighted mean utility is 0.4389 and the penalty term is 0.0760, which add to U_{ent}. The plain mean is 0.6.

Proposition 4.31 (5.2 — hard KL ball). *For $\delta > 0$,*

$$\min_{\nu \in \Delta_M,\ \mathrm{KL}(\nu|\pi) \le \delta} \sum_j \nu_j U_j = \sup_{\lambda \ge 0} \Big\{ -\lambda \log \sum_j \pi_j e^{-U_j/\lambda} - \lambda\delta \Big\},$$

where the term at $\lambda \downarrow 0$ is read as $\min_j U_j$.

Proof. The point $\nu = \pi$ has $\mathrm{KL} = 0 < \delta$, so the convex problem has a strictly feasible point and no duality gap. The Lagrangian is $\sum_j \nu_j U_j + \lambda(\mathrm{KL}(\nu \mid \pi) - \delta)$. For $\lambda > 0$, Proposition 5.1 with $\eta = 1/\lambda$ gives its minimum over ν as $-\lambda \log \sum_j \pi_j e^{-U_j/\lambda} - \lambda\delta$. As $\lambda \downarrow 0$ the first term tends to $\min_j U_j$. □

The penalised form fixes the multiplier at $1/\eta$. The ball form optimises it, so the two describe different families of adversaries. A Wasserstein ball around a nominal law has a duality theorem only under conditions on the cost and the utility: lower semicontinuity, integrability and a growth or compactness bound [39]. That theorem is cited as stated. Each utility must be checked against its conditions before the dual form is used.

Coverage and greedy selection. A portfolio of N designs should cover several targets without spending effort on a target that is already covered. With target weights $w_j \ge 0$ and binding inputs $p_j(X) \ge 0$, the coverage of a set $\mathcal{B}$ is $F_{\text{cov}}(\mathcal{B}) = \sum_j w_j \log(1 + S_j(\mathcal{B}))$, where $S_j(\mathcal{B}) = \sum_{X \in \mathcal{B}} p_j(X)$. The gain from adding X is $\Delta_X F_{\text{cov}}(\mathcal{B}) = F_{\text{cov}}(\mathcal{B} \cup \{X\}) - F_{\text{cov}}(\mathcal{B})$.

Proposition 4.32 (5.3 — diminishing returns). *$F_{\text{cov}}(\varnothing) = 0$, F_{cov} is monotone, and for $\mathcal{A} \subseteq \mathcal{B}$ and $X \notin \mathcal{B}$,*

$$\Delta_X F_{\text{cov}}(\mathcal{A}) \ \ge \ \Delta_X F_{\text{cov}}(\mathcal{B}) \ \ge \ 0.$$

Proof. The gain is $\sum_j w_j \log\big(1 + p_j(X)/(1 + S_j(\mathcal{B}))\big)$, which is nonnegative. For $\mathcal{A} \subseteq \mathcal{B}$ every $S_j(\mathcal{A}) \le S_j(\mathcal{B})$, so each denominator grows and each logarithm falls. □

Proposition 4.33 (5.4 — greedy bound, cardinality only). *Let the pool be finite and the only constraint be $|\mathcal{B}| = N$. Let S_t be the set after t greedy steps, each adding the element of largest gain, and let $S^\star$ be a best set of size N. Then*

$$F_{\mathrm{cov}}(S_N) \;\ge\; \Big[1 - \big(1 - \tfrac{1}{N}\big)^N\Big] F_{\mathrm{cov}}(S^\star) \;\ge\; (1 - e^{-1})\, F_{\mathrm{cov}}(S^\star).$$

Proof. By monotonicity and submodularity, $F(S^\star) \le F(S_t \cup S^\star) \le F(S_t) + \sum_{X \in S^\star} \Delta_X F(S_t) \le F(S_t) + N\,[F(S_{t+1}) - F(S_t)]$, because greedy takes the largest gain. Rearranged, $F(S^\star) - F(S_{t+1}) \le (1 - 1/N)[F(S^\star) - F(S_t)]$. Iterating from $F(S_0) = 0$ gives the first inequality, and $(1 - 1/N)^N \le e^{-1}$ gives the second [40]. □

The proof uses unrestricted choice at each step. Subtracting a nonnegative symmetric zero-diagonal pairwise similarity penalty preserves submodularity while breaking monotonicity, and the collision limit and format constraints of the Methods remove unrestricted choice, so the $(1 - e^{-1})$ bound covers the cardinality-only greedy rule.

Conformal calibration and risk control. A score can be turned into an interval or a threshold with a guarantee that needs no model of the scores. The one assumption is exchangeability: the calibration items and the new item are drawn so that their order carries no information. The guarantee is an average over calibration sets and new items. It does not hold for one calibration set or for one new item.

Theorem 4.34 (5.5 — split-conformal rank bound). *Let $S_1, \dots, S_{n+1}$ be exchangeable real scores. Let $\alpha \in (0, 1)$ and put $k = \lceil (n+1)(1-\alpha) \rceil$. Let $\widehat{q}$ be the k-th smallest of $S_1, \dots, S_n$ if $k \le n$, and $+\infty$ if $k = n + 1$. Then $\mathbb{P}(S_{n+1} \le \widehat{q}) \ge 1 - \alpha$. If the scores are almost surely distinct and $k \le n$, the probability equals $k/(n+1)$, which is at most $1 - \alpha + 1/(n+1)$.*

Proof. If the scores are distinct, the rank of S_{n+1} among all $n+1$ scores is uniform on $\{1, \dots, n+1\}$, by exchangeability. The event $S_{n+1} \le \widehat{q}$ is the event that this rank is at most k, which has probability $k/(n+1) \ge 1-\alpha$. With ties, add independent random tie breakers to obtain a uniform rank. Removing the tie breakers while keeping equality as inclusion can only enlarge the event. The case $k = n + 1$ gives probability one. □

For scores $S_i = |Y_i - \widehat{\mu}(X_i)|/\widehat{\sigma}(X_i)$ the interval $\widehat{\mu}(X_{n+1}) \pm \widehat{q}\,\widehat{\sigma}(X_{n+1})$ therefore has marginal coverage at least $1 - \alpha$. The smallest attainable miscoverage step is $1/(n+1)$. A level $\alpha < 1/(n+1)$ forces the infinite threshold.

Theorem 4.35 (5.6 — conformal risk control). *Let $L_1, \dots, L_{n+1}$ be exchangeable random functions of λ in a compact interval Λ with largest element $\lambda_{\max}$. Assume each L_i is nonincreasing, right-continuous, with values in $[0, B]$, and $L_i(\lambda_{\max}) \le \alpha$ almost surely. Put*

$$\widehat{\lambda} = \inf\Big\{\lambda \in \Lambda :\; \frac{1}{n+1}\Big[\sum_{i=1}^{n} L_i(\lambda) + B\Big] \le \alpha\Big\},$$

with $\widehat{\lambda} = \lambda_{\max}$ when the set is empty. Then $\mathbb{E}\, L_{n+1}(\widehat{\lambda}) \le \alpha$.

Proof. Define the look-ahead threshold $\lambda^* = \inf\{\lambda : (n+1)^{-1} \sum_{i=1}^{n+1} L_i(\lambda) \le \alpha\}$. The set contains $\lambda_{\max}$, and right-continuity makes the infimum attained. Because $L_{n+1} \le B$, every λ feasible for $\widehat{\lambda}$ is feasible for λ^*, so $\widehat{\lambda} \ge \lambda^*$, including the fallback. Monotonicity gives $L_{n+1}(\widehat{\lambda}) \le L_{n+1}(\lambda^*)$. The threshold λ^* is a symmetric function of the $n+1$ losses, so exchangeability gives $\mathbb{E}L_{n+1}(\lambda^*) = \mathbb{E}[(n+1)^{-1} \sum_i L_i(\lambda^*)] \le \alpha$. □

Four conditions decide whether the theorem applies. First, the loss must be monotone. The joint loss "admitted and bad" fits. The share of bad designs among the admitted does not, because numerator and denominator both change. Second, a rule with two thresholds needs a nested scalar family, either one summary score or a monotone curve in the plane of the two thresholds, because independently tuned thresholds can make the empirical loss rise and void the bound. Third, the admitted set is $\{s > \lambda\}$ with strict inequality, so that the loss is right-continuous. Fourth, an empty feasible set returns $\lambda_{\max}$ and admits nothing, which is recorded as a typed outcome.

The guarantee is marginal. It bounds $\mathbb{P}(\text{admitted and bad}) \le \alpha$, the joint probability. The share of bad designs among those admitted is a different quantity. If 1 percent of candidates are admitted and every one is bad, then the bound holds at $\alpha = 0.10$, yet every admitted design is bad. A claim about the admitted set needs a different tool, such as a multiple-testing procedure on conformal p-values. The rule is the theorem of the cited source [41], proved here at a fixed horizon.

Proposition 4.36 (5.7 — transfer under a bounded density ratio). *Let the new observation have a fixed law Q with $Q \ll P$ and $\mathrm{d}Q/\mathrm{d}P \le W$. If a calibration procedure gives $\mathbb{E}_{\mathrm{cal},P} L \le \alpha$ and $0 \le L \le B$, then $\mathbb{E}_{\mathrm{cal},Q} L \le \min(W\alpha, B)$.*

Proof. Given the calibration outcome, change measure: $\mathbb{E}_Q L = \mathbb{E}_P[(\mathrm{d}Q/\mathrm{d}P)L] \le W\,\mathbb{E}_P L$. Average over calibration and use $L \le B$. □

Generated candidates are optimised against a predictor, so their law Q can differ from the calibration law P. The bound degrades by the factor W and is vacuous when $W\alpha \ge B$. A law Q that is chosen after the calibration outcomes are seen is not covered by this proof. Particles from one sequential Monte Carlo run share resampling ancestry and fail exchangeability, so split-conformal calibration and risk control draw at most one candidate or cloud summary per independent seed.

Exact count bounds, power and clustered designs. When a rate is estimated from a count of independent units, the exact binomial distribution gives a bound with no normal approximation. The bound is conservative because the count is discrete. Its units must be independent: replicate wells for one construct measure repeatability and do not supply independent constructs.

Proposition 4.37 (5.8 — one-sided exact bound and the zero-failure rule). *Let $K \sim \mathrm{Bin}(n, p)$ and $\alpha \in (0,1)$. Let $U(k)$ be the $(1-\alpha)$ quantile of $\mathrm{Beta}(k+1, n-k)$ for $k < n$, and $U(n) = 1$. Then $\mathbb{P}_p\{p \le U(K)\} \ge 1-\alpha$ for every p. For $K = 0$, $U(0) = 1 - \alpha^{1/n}$, and $U(0) \le p_0$ if and only if $n \ge \log\alpha / \log(1-p_0)$.*

Proof. The identity $\mathbb{P}_p(K \le k) = \mathbb{P}(\mathrm{Beta}(k+1, n-k) > p)$ shows that $U(k)$ is the value of p at which $\mathbb{P}_p(K \le k) = \alpha$, and that $\mathbb{P}_p(K \le k)$ falls as p grows. For fixed p, let k_p be the largest k with $\mathbb{P}_p(K \le k) \le \alpha$. The bound fails only when $K \le k_p$, which has probability at most α. At $k = 0$ the endpoint solves $(1-U)^n = \alpha$. Both logarithms are negative, which gives the condition on n. □

Numerical illustration: with 18 events in 200 independent units and $\alpha = 0.05$, the upper endpoint is 0.1305, so an observed fraction of 0.09 does not clear an upper-risk limit of 0.10. For zero events, the required n is 14, 29, 59, 149 and 299 for $p_0 = 0.20$, 0.10, 0.05, 0.02 and 0.01. A one-sided exact test of $p \geq 0.8$ with $n = 200$ and size at most 0.05 rejects for $K \leq 150$ (size 0.0494), and its power reaches 0.8 only for alternatives $p \leq 0.7258$. A sample-size label alone therefore does not show power against a smaller drop.

Proposition 4.38 (5.9 — design effect). *Let m observations in one cluster have common variance σ^2 and common pairwise correlation ρ_c. Then $\mathrm{Var}(\bar{Y}) = (\sigma^2/m)\,[1 + (m-1)\rho_c]$. For C clusters, the effective sample size is Cm/D with $D = 1 + (m-1)\rho_c$.*

Proof. $\mathrm{Var}(\bar{Y}) = m^{-2}\{m\sigma^2 + m(m-1)\rho_c\sigma^2\}$. □

Wells within a cluster are not independent units, so the number of clusters sets the large-sample degrees of freedom. A balanced model with C clusters, m constructs per cluster and r technical repeats per construct has grand-mean variance $\sigma_c^2/C + \sigma_v^2/(Cm) + \sigma_e^2/(Cmr)$, because the cluster, construct and repeat effects are independent and their means have these variances. More repeats change only the last term, which is why they cannot replace more clusters.

Exploration weights and sequential evidence. A campaign that picks designs partly at random with known probabilities can weight each outcome by the inverse of its selection probability, and a running product of fair-bet factors gives a test that can be watched at any time.

Proposition 4.39 (5.10 — predictable inverse-probability weights). *Given the past $\mathcal{F}_{j-1}$, let u_j, p_j and a bounded loss ℓ_j be fixed on a finite pool, with $p_j > 0$ wherever $u_j > 0$. For $X_j \sim p_j$, $\mathbb{E}[(u_j/p_j)(X_j)\,\ell_j(X_j) \mid \mathcal{F}_{j-1}] = \sum_x u_j(x)\ell_j(x)$.*

Proof. Expand the conditional expectation as a finite sum and cancel $p_j(x)$. □

The identity needs selection to come before the observed outcome. A final predictor fitted to the sampled outcomes is not fixed in this argument. If $p_j = \epsilon u_j + (1-\epsilon)q_j$ then $u_j/p_j \leq 1/\epsilon$.

Theorem 4.40 (5.11 — Ville's inequality). *Let $(E_t)_{t\geq 0}$ be a nonnegative supermartingale with $\mathbb{E}E_0 \leq 1$. For every $\alpha \in (0,1)$, $\mathbb{P}(\sup_t E_t \geq 1/\alpha) \leq \alpha$.*

Proof. For a horizon T let $\tau = \inf\{t \leq T : E_t \geq 1/\alpha\}$, with $\tau = T$ if there is no crossing. Optional sampling gives $\mathbb{E}E_\tau \leq \mathbb{E}E_0 \leq 1$. On the crossing event $E_\tau \geq 1/\alpha$, so $1 \geq \alpha^{-1}\mathbb{P}(\max_{t\leq T} E_t \geq 1/\alpha)$. Let $T \to \infty$ and use monotone convergence of the events. □

A product $E_t = \prod_{s\leq t} e_s$ is a test supermartingale when each factor is nonnegative, measurable with respect to the past, and has conditional expectation at most one under every null law [42]. A stake chosen from past data is allowed. A stake tuned on the current outcome is not predictable and can destroy the property. Stopping at any data-dependent time is then valid. Repeated fixed-horizon tests are not: they inflate the crossing probability.

4.6. Representation floor and predictive information

This section states what the two penalty terms of the Interface-JEPA design guarantee about the spread of a representation (Theorem 6.1) and why they act on raw vectors (Proposition 6.2). It states what a representation must keep to serve a downstream endpoint (Theorem 6.3) and why a small prediction loss does not show that it does (Example 6.4). The base model has no predictive layer.

Let $H \in \mathbb{R}^{B\times d}$ hold B centred raw representation vectors of width d, and let $\Sigma = H^\top H/(B-1)$ be the batch covariance. With a target spread $\tau > 0$ and a small constant $\delta \geq 0$, the two terms are

$$\mathcal{L}_{\mathrm{var}} = \frac{1}{d}\sum_{i=1}^{d}\left(\tau - \sqrt{\Sigma_{ii}+\delta}\right)_+, \qquad \mathcal{L}_{\mathrm{cov}} = \frac{1}{d}\sum_{i\neq j}\Sigma_{ij}^2.$$

Theorem 4.41 (6.1 — noncollapse floor). *If $\mathcal{L}_{\mathrm{var}} \leq \eta$ and $\mathcal{L}_{\mathrm{cov}} \leq \kappa$, then*

$$\lambda_{\min}(\Sigma) \;\geq\; (\tau - d\eta)_+^2 - \delta - \sqrt{(d-1)\kappa}.$$

The bound is attained: for $d \geq 2$, $s > 0$, $0 < a \leq s^2/(d-1)$, $\tau = s$ and $\delta = \eta = 0$, the matrix $\Sigma = s^2 I - a(\mathbf{1}\mathbf{1}^\top - I)$ has $\mathcal{L}_{\mathrm{var}} = 0$, $\mathcal{L}_{\mathrm{cov}} = (d-1)a^2 = \kappa$ and $\lambda_{\min} = s^2 - (d-1)a = s^2 - \sqrt{(d-1)\kappa}$.

Proof. Each hinge term is nonnegative and their mean is at most η, so each is at most $d\eta$. Hence $\sqrt{\Sigma_{ii}+\delta} \geq \tau - d\eta$ and $\Sigma_{ii} \geq (\tau - d\eta)_+^2 - \delta$. Write $\Sigma = D + O$ with D diagonal and O symmetric with zero diagonal. Weyl's inequality gives $\lambda_{\min}(\Sigma) \geq \min_i D_{ii} - \|O\|_{\mathrm{op}}$.

Let μ be an eigenvalue of O with $|\mu| = \|O\|_{\mathrm{op}}$. Since $\operatorname{tr} O = 0$, the other $d-1$ eigenvalues sum to $-\mu$, and Cauchy–Schwarz gives that their squares sum to at least $\mu^2/(d-1)$. So $\|O\|_{\mathrm{F}}^2 \geq \mu^2 d/(d-1)$. Also $\|O\|_{\mathrm{F}}^2 = \sum_{i\neq j}\Sigma_{ij}^2 = d\,\mathcal{L}_{\mathrm{cov}} \leq d\kappa$, which gives $\|O\|_{\mathrm{op}} \leq \sqrt{(d-1)\kappa}$.

For the example, the eigenvalue along $\mathbf{1}$ is $s^2-(d-1)a$ and every orthogonal direction has eigenvalue s^2+a. For $B \geq d+1$ the example is the covariance of a centred sample $H = \sqrt{B-1}\,T\Sigma^{1/2}$, where $T^\top T = I_d$ and $T^\top \mathbf{1} = 0$. The case $d = 1$ has no off-diagonal term. □

A positive floor needs the joint condition $(\tau - d\eta)_+^2 > \delta + \sqrt{(d-1)\kappa}$, and loose separate limits on η and κ can fail it. For $d = 8$, $\tau = 1$ and $\delta = 0$ the lower bound is positive exactly when $\eta < 1/8$ and $\kappa < (1-8\eta)^4/7$. The rank of Σ is at most $B-1$, so $B \geq d+1$ is necessary for a full-rank covariance. A gate that tests the median variance over dimensions does not enforce this bound, because the bound constrains the smallest diagonal entry.

Proposition 4.42 (6.2 — normalised vectors cannot carry the floor). *If every row of H has unit norm before centring, then $\operatorname{tr}\Sigma \leq B/(B-1)$ and $\lambda_{\min}(\Sigma) \leq B/[d(B-1)]$.*

Proof. $\operatorname{tr}\Sigma = (B-1)^{-1}\sum_b \|h_b - \bar{h}\|^2 \leq (B-1)^{-1}\sum_b \|h_b\|^2 = B/(B-1)$, and the smallest eigenvalue is at most the mean eigenvalue $\operatorname{tr}\Sigma/d$. □

For $B = 4096$ and $d = 256$ the limit is 0.003907, so a floor of 0.1 is impossible on normalised vectors. The variance terms therefore act on raw vectors, and cosine comparisons may use normalised copies.

Theorem 4.43 (6.3 — information retained for an endpoint). *Let X, Y and the actions lie in standard Borel spaces, let $Z = h(X)$ be measurable, and let the loss $\ell(a, Y) \in [0, M]$ be measurable. Assume measurable approximate Bayes rules exist with arbitrarily small excess risk. With $R^\star_X$ and $R^\star_Z$ the optimal risks given X and given Z, and I the conditional mutual information in nats,*

$$0 \;\le\; R^\star_Z - R^\star_X \;\le\; M\sqrt{2\, I(X; Y \mid Z)}.$$

Proof. Replacing $P(Y \mid x)$ by $P(Y \mid h(x))$ changes the expected loss of any fixed action by at most M times their total variation distance. Comparing two optimal actions uses this twice, so the excess risk is at most $2M\,\mathbb{E}\,\mathrm{TV}(P_{Y|X}, P_{Y|Z})$. Pinsker's inequality and Jensen's inequality bound this by $2M\sqrt{\mathbb{E}\,\mathrm{KL}(P_{Y|X} \mid P_{Y|Z})/2} = M\sqrt{2I(X; Y \mid Z)}$. The lower bound holds because every rule that uses Z also uses X. Infima are handled by approximate minimisers. □

The theorem concerns optimal risks under a bounded loss. The error of a trained predictor and an unbounded loss, such as a Huber loss on an unbounded range, fall beyond its hypotheses.

Example 6.4 (zero prediction loss, unbounded information loss). Let $X = (U, Y)$ with U independent of a finite-valued Y. Let $h(X) = U$, and let the prediction target and the predictor both equal U. The prediction loss is zero and U may have a full-rank covariance. Yet $I(X; Y \mid h(X)) = H(Y)$, which grows without bound as the alphabet of Y grows. A representation can predict its chosen target exactly and discard another endpoint entirely.

A small prediction loss therefore does not bound the conditional information. A population cross-entropy for a categorical endpoint is at least $H(Y \mid h(X))$, which bounds $I(X; Y \mid h(X))$ from above, but an empirical cross-entropy needs a generalisation argument for the selected predictor and data.

4.7. Decoder pullback metric and common descent

A step in the latent space can change the decoded molecule by unequal amounts in different directions, so the guidance of the full system measures a step by its effect on decoded features. This section proves that the resulting metric is positive definite (Proposition 7.1) and that the preconditioned gradient is the steepest direction (Theorem 7.2). It bounds the error of the linear solve and of the linear model (Propositions 7.3 and 7.4), covers rotation (Proposition 7.5) and states what a common descent direction guarantees (Proposition 7.7, Theorem 7.8). Figure 1 and Figure 3 illustrate the metric and the common direction.

Let $h_\phi : \mathbb{R}^D \to \mathbb{R}^P$ collect differentiable dimensionless decoded features, with Jacobian J_h. Let $W_h \succeq 0$ hold fixed feature weights and $\tau_{\text{ridge}} > 0$ be a ridge. Put $A = W_h^{1/2} J_h$. The metric and its squared length are

$$G_\phi = J_h^\top W_h J_h + \tau_{\text{ridge}} I_D, \qquad \|v\|^2_{G_\phi} = \|Av\|_2^2 + \tau_{\text{ridge}}\|v\|_2^2.$$

The first term is the linearised displacement of the weighted features. The ridge covers directions that the features do not see.

Proposition 4.44 (7.1 — spectral bounds). *For every v, $\tau_{\text{ridge}}\|v\|^2 \le v^\top G_\phi v \le (\tau_{\text{ridge}} + \|A\|_2^2)\|v\|^2$. Hence $G_\phi \succ 0$ and $\|G_\phi^{-1}\|_2 \le 1/\tau_{\text{ridge}}$.*

Proof. The squared length gives the lower bound by dropping its first nonnegative term, and the operator-norm inequality gives the upper bound. The inverse norm is the reciprocal of the smallest eigenvalue. □

A very small ridge amplifies gradient error in directions the features do not see, and a large ridge removes the preconditioning. The smallest and largest eigenvalue estimates are therefore recorded on validation states.

Theorem 4.45 (7.2 — steepest direction). *For $G \succ 0$ and a differentiable utility U, $\max_{\|d\|_G \le 1} \nabla U^\top d = \sqrt{\nabla U^\top G^{-1} \nabla U}$. When $\nabla U \neq 0$ the maximum is attained at $d^\star = G^{-1}\nabla U / \sqrt{\nabla U^\top G^{-1} \nabla U}$.*

Proof. Put $z = G^{1/2} d$, so $\|z\|_2 \le 1$ and $\nabla U^\top d = (G^{-1/2}\nabla U)^\top z$. Cauchy–Schwarz gives the maximum, attained when z is parallel to $G^{-1/2}\nabla U$. □

The direction $G^{-1}\nabla U$ is a natural gradient of a statistical family only if G is the Fisher information of that family. The pullback metric measures declared feature displacement, and calling it Fisher would need a probabilistic derivation that this supplement does not give.

Proposition 4.46 (7.3 — inverse products and residuals). *(a) $(A^\top A + \tau_{\text{ridge}} I)^{-1} g = \tau_{\text{ridge}}^{-1} g - \tau_{\text{ridge}}^{-1} A^\top (AA^\top + \tau_{\text{ridge}} I_P)^{-1} A g$. (b) If $\widehat{d}$ is any approximate solution of $G_\phi d = g$ with residual $r = g - G_\phi \widehat{d}$, then $\|\widehat{d} - G_\phi^{-1} g\|_2 \le \|r\|_2 / \tau_{\text{ridge}}$.*

Proof. (a) is the Woodbury identity. (b) $\widehat{d} - G_\phi^{-1} g = -G_\phi^{-1} r$, and Proposition 7.1 bounds the norm. □

The bound in (b) concerns the linear solve. The validity of the decoded molecule is checked after the step.

Proposition 4.47 (7.4 — linearisation remainder and trust radius). *If J_h is L_J-Lipschitz in operator norm on the segment from q to $q + \delta$, then*

$$\|h_\phi(q+\delta) - h_\phi(q) - J_h(q)\delta\| \le \tfrac{1}{2} L_J \|\delta\|^2, \qquad \|W_h^{1/2}\{h_\phi(q+\delta) - h_\phi(q)\}\| \le \|A\delta\| + \tfrac{1}{2}\|W_h^{1/2}\|\, L_J \|\delta\|^2.$$

If $\|\delta\|_{G_\phi} \le \Delta$, then $\|\delta\|_2 \le \Delta / \sqrt{\tau_{\text{ridge}}}$ and $\|A\delta\| \le \Delta$, so the weighted displacement is at most $\Delta + \|W_h^{1/2}\| L_J \Delta^2 / (2\tau_{\text{ridge}})$.

Proof. Write $h(q+\delta) - h(q) = \int_0^1 J_h(q + t\delta)\delta \, \mathrm{d}t$, subtract $J_h(q)\delta$, use $\|J_h(q+t\delta) - J_h(q)\| \le t L_J \|\delta\|$ and integrate. Multiply by $W_h^{1/2}$ and apply the triangle inequality. The trust statements follow from the squared length. □

The quadratic term shows why a step that satisfies its linearised metric limit can leave the intended feature ball. The decoded displacement is therefore evaluated after each step, and the step is shortened when it is too large.

Proposition 4.48 (7.5 — equivariance and padding). *Let the state transform by an orthogonal representation ρ_g, and the features by an orthogonal representation σ_g with $h_\phi(\rho_g q, gc) = \sigma_g h_\phi(q, c)$. If $\sigma_g^\top W_h \sigma_g = W_h$, then $G_\phi(\rho_g q, gc) = \rho_g G_\phi(q, c)\rho_g^\top$, and $G_\phi^{-1}\nabla U$ is equivariant for every invariant differentiable U. A feature map must also factor through the active state, $h_{\mathrm{pad}}(q_{\mathrm{pad}}) = h_e(S_e q_{\mathrm{pad}})$ with S_e the selector of active coordinates. A ridge added in padded storage would give padding an artificial geometry, so the metric is built in the packed state and only the resulting vector is scattered back.*

Proof. Differentiating the feature relation gives $J_h(\rho_g q, gc) = \sigma_g J_h(q, c)\rho_g^\top$. Substituting in the definition of G_ϕ and using $\sigma_g^\top W_h \sigma_g = W_h$ and $\rho_g^\top \rho_g = I$ gives the first claim. Lemma 2.7 gives the second. □

Proposition 4.49 (7.6 — density conventions). *If a model declares the Riemannian volume $\mathrm{dvol}_G = \sqrt{\det G_\phi}\,\mathrm{d}q$, a Lebesgue density p_L and a Riemannian density p_G satisfy $p_L = p_G\sqrt{\det G_\phi}$, so $\log p_G = \log p_L - \frac{1}{2}\log\det G_\phi$. The Riemannian divergence is $\mathrm{div}_G\, v = (\det G)^{-1/2}\partial_i\big((\det G)^{1/2} v^i\big)$.*

Proof. The two densities describe the same probability on the same set, so $p_L\,\mathrm{d}q = p_G\,\mathrm{dvol}_G$. □

Using div_G in a likelihood formula while the base density is Lebesgue mixes two reference measures. The Methods use G_ϕ as a preconditioner and keep the Lebesgue density of the intrinsic chart.

A common direction is a step that improves all utilities to first order. It is the point of smallest length in the convex hull of the preconditioned gradients, and the origin in that hull means that no such step exists at this state.

For n_f scaled losses $\widetilde{f}_i = (f_i - b_i)/r_i$ with offsets b_i and scales $r_i > 0$, let $g_i = \nabla\widetilde{f}_i$, $a_i = G_\phi^{-1} g_i$ and $H_{ij} = g_i^\top G_\phi^{-1} g_j$. Let $\lambda^\star$ minimise $\frac{1}{2}\lambda^\top H\lambda$ over the simplex Δ_{n_f}, and put $w = \sum_i \lambda_i^\star a_i$.

Proposition 4.50 (7.7 — projection inequality). *For every i, $g_i^\top w = \langle a_i, w\rangle_{G_\phi} \geq \|w\|_{G_\phi}^2$, with equality when $\lambda_i^\star > 0$.*

Proof. The conditions for the simplex problem are $H\lambda - \nu\mathbf{1} - \mu = 0$, $\mathbf{1}^\top\lambda = 1$, $\lambda, \mu \geq 0$ and $\lambda_i\mu_i = 0$. Row i gives $(H\lambda)_i = \nu + \mu_i \geq \nu$. Multiplying by $\lambda^\top$ and using complementarity gives $\lambda^\top H\lambda = \nu$. Now $(H\lambda)_i = \langle a_i, w\rangle_{G_\phi}$ and $\lambda^\top H\lambda = \|w\|_{G_\phi}^2$. Equality holds where $\mu_i = 0$. □

Theorem 4.51 (7.8 — one step lowers every loss). *Let each $\widetilde{f}_i$ have an L_i-Lipschitz gradient near q and let $w \neq 0$ satisfy Proposition 7.7. Then $\widetilde{f}_i(q - hw) \leq \widetilde{f}_i(q) - h\|w\|_{G_\phi}^2 + \frac{1}{2}L_i h^2\|w\|_2^2$. For any $c \in (0, 1)$, the common Armijo condition $\widetilde{f}_i(q - hw) \leq \widetilde{f}_i(q) - ch\|w\|_{G_\phi}^2$ holds for every i whenever $0 < h \leq \min_i 2(1-c)\|w\|_{G_\phi}^2/(L_i\|w\|_2^2)$ and the step stays in the neighbourhood.*

Proof. The descent lemma gives $\widetilde{f}_i(q - hw) \leq \widetilde{f}_i(q) - h g_i^\top w + \frac{1}{2}L_i h^2\|w\|_2^2$. Use $g_i^\top w \geq \|w\|_{G_\phi}^2$ and rearrange for the fraction c. □

Four limits and one related fact apply. If $w = 0$, a convex combination of the gradients vanishes, which is first-order Pareto stationarity and not global Pareto optimality. An approximate linear solve, an approximate simplex solution or a later projection requires the inequalities of Proposition 7.7 and Theorem 7.8 to be checked on the realised correction. The weights depend on the state, so the combined field need not be the gradient of one potential. With smooth constraints, a local weak

Pareto point satisfies the Karush–Kuhn–Tucker condition $\sum_i \lambda_i \nabla f_i + \sum_j \mu_j \nabla c_j + \sum_k \nu_k \nabla h_k = 0$ with $\lambda \in \Delta$ and $\mu_j \geq 0$ under a constraint qualification. The condition is necessary and, without convexity, not sufficient. The related fact concerns fixed weights. A linear scalarisation with all weights positive returns a Pareto-optimal point, because a dominating point would lower the weighted sum. A nonconvex frontier has Pareto-optimal points that no weighting reaches (standard, by the supporting-hyperplane argument), and the min-norm direction avoids that limit.

4.8. Numerical error budgets

This section bounds the error of each approximation behind a reported number: the integrator step, the stopping rule of a solver, the finite number of probes or particles, and the rounding of floating-point sums. It shows how the bounds add up for a log density and which refinement test measures each part. The base model relies only on the bound on floating-point sums (Proposition 8.6); the other results support the full system, because the base model integrates no flow and runs no particle method.

Coupled state and log density. Let $v(q,t)$ be the field and $d = \nabla \cdot v$ its divergence. The log density ℓ_t along a trajectory solves

$$\frac{\mathrm{d}}{\mathrm{d}t}\begin{bmatrix} q_t \\ \ell_t \end{bmatrix} = \begin{bmatrix} v(q_t,t) \\ -d(q_t,t) \end{bmatrix}, \qquad \ell_0 = \log p_0(q_0),$$

and ℓ_T equals the model log density only for the same trajectory, field and divergence. The production solver integrates both components in one system, or records the interpolation and quadrature error that a separate pass adds. An adaptive solver measures its error in a weighted norm,

$$\|e\|_{\mathrm{tol}}^2 = \frac{1}{D+1}\Big[\sum_{j=1}^{D}\Big(\frac{e_{q,j}}{a_{q,j} + r_q|q_j|}\Big)^2 + \Big(\frac{e_\ell}{a_\ell + r_\ell|\ell|}\Big)^2\Big],$$

where a and r are absolute and relative tolerances. If ℓ shares one absolute tolerance with a state component, its share of the sum is negligible for large D. The solver profile therefore declares separate tolerances for state and log density.

Theorem 4.52 (8.1 — discrete Grönwall bound). *Let a one-step method Ψ_h on a grid $t_k = kh$ satisfy $\|\Psi_h(x,t) - \Psi_h(y,t)\| \leq (1+Lh)\|x-y\|$, and let its local defect along the exact solution satisfy $\|q(t_{k+1}) - \Psi_h(q(t_k),t_k)\| \leq Ch^{p+1}$. After $K = T/h$ steps,*

$$\|q(T) - q_K\| \leq \begin{cases} (C/L)(e^{LT}-1)\,h^p, & L > 0, \\ CT\,h^p, & L = 0. \end{cases}$$

Proof. Let $e_k = \|q(t_k) - q_k\|$. Adding and subtracting $\Psi_h(q(t_k),t_k)$ gives $e_{k+1} \leq (1+Lh)e_k + Ch^{p+1}$. From $e_0 = 0$, $e_K \leq Ch^{p+1}\sum_{j<K}(1+Lh)^j = Ch^p\{(1+Lh)^K - 1\}/L$, and $(1+Lh)^K \leq e^{LT}$. For $L = 0$ the sum is K equal defects. □

The bound is for a fixed-step method with a known defect constant C, and the embedded error estimate of an adaptive solver is not that constant. A convergence study halves the tolerance or

the largest step and checks the endpoint, the integrated divergence and every statistic that a gate reads. A plateau can mean rounding or trace noise and does not by itself show convergence.

Proposition 4.53 (8.2 — projections and events). *If the exact solution stays in a closed convex set $\mathcal{Q}$ and $\Pi_{\mathcal{Q}}$ is the Euclidean projection, then $q_{k+1} = \Pi_{\mathcal{Q}}\{\Psi_h(q_k, t_k)\}$ has no larger error than the unprojected step. If integration stops at a root of $g(q_t, t) = 0$ where $\partial_t g + \nabla g^\top v \neq 0$, the crossing is transversal and local interpolation can locate it.*

Proof. The projection is nonexpansive, and $q(t_{k+1}) \in \mathcal{Q}$ is its own projection, so $\|\Pi(x) - q(t_{k+1})\| \leq \|x - q(t_{k+1})\|$. Transversality is the condition of the implicit function theorem for the crossing time. □

Molecular validity sets, with discrete sequence rules and topology changes, are not convex, and their decoder projection can be discontinuous, so formal order does not transfer across such a projection. Projected generation is therefore a separate solver profile whose endpoint distribution is validated directly. A gate on a discontinuous decoded predicate is an admission test after the step.

Stable sums and entropic transport. The scaling equations of Section S1.4 and the particle weights are sums of exponentials, and the next statements control their evaluation.

Lemma 4.54 (8.3 — log-sum-exp shift). *For finite $z_1, \dots, z_m$ and any c, $\log \sum_i e^{z_i} = c + \log \sum_i e^{z_i - c}$. With $c = \max_i z_i$ every shifted term $e^{z_i - c}$ lies in $(0, 1]$ and the sum is at least one, so the sum can neither overflow nor underflow to zero.*

The scaling equations of Theorem 4.3 are evaluated in this form, for example $a_i = -\tau_{\mathrm{OT}} \varepsilon \log \sum_j \mu_{E,j} \exp\{(b_j - C_{ij})/\varepsilon\}$, with each log-sum-exp shifted by its maximum. When ε is small, the subtraction of nearly equal primal and dual values can still lose accuracy. The solver therefore reports the absolute primal-dual gap, a scale-normalised gap, the marginal residuals and the plan change, and no one of them replaces the other three.

Proposition 4.55 (8.4 — objective gap and plan distance). *Let $\Gamma^\star$ minimise a differentiable μ-strongly convex F on a convex set and let $\widehat{\Gamma}$ be feasible. Then $F(\widehat{\Gamma}) - F(\Gamma^\star) \geq \frac{\mu}{2} \|\widehat{\Gamma} - \Gamma^\star\|_{\mathrm{F}}^2$. For the contact objective of Section S1.4, on the set $\{0 < \Gamma_{ij} \leq \Gamma_{\max}\}$, one valid constant is $\mu = \varepsilon_{\mathrm{OT}} / \Gamma_{\max}$.*

Proof. Optimality of $\Gamma^\star$ gives $\langle \nabla F(\Gamma^\star), \widehat{\Gamma} - \Gamma^\star \rangle \geq 0$, and strong convexity adds $\frac{\mu}{2} \|\widehat{\Gamma} - \Gamma^\star\|^2$. The Hessian of $\varepsilon_{\mathrm{OT}} \mathrm{KL}(\Gamma \mid \mu_R \otimes \mu_E)$ is $\mathrm{diag}(\varepsilon_{\mathrm{OT}} / \Gamma_{ij}) \succeq (\varepsilon_{\mathrm{OT}} / \Gamma_{\max}) I$, and the other terms are convex. □

The constant depends on an upper bound for the plan entries, and iterates near zero have large curvature. A generic tolerance on plan change cannot be turned into a guarantee on the objective, or the reverse, without such a bound.

Particle weights. Particle algorithms add log weights over steps. Set $\ell_k^i = \log \widetilde{w}_{k-1}^i + \log \omega_k$, where ω_k is the incremental weight of Section S1.4, $\widetilde{w}_{k-1}^i = 1/N_{\mathrm{p}}$ after resampling and $\widetilde{w}_{k-1}^i = \bar{w}_{k-1}^i$ when resampling is skipped. Then $\log \bar{\ell}_k = \mathrm{LSE}(\ell_k^{1:N_{\mathrm{p}}})$ and $\bar{w}_k^i = \exp\{\ell_k^i - \log \bar{\ell}_k\}$. Subtracting the maximum prevents overflow, but values far below the maximum underflow to zero. The loss is harmless only if the discarded mass is below a declared tolerance.

Proposition 4.56 (8.5 — mass lost by truncation). *Let $B \geq 0$, let $m = \max_i \ell_i$ and set to zero every term with $\ell_i - m < -B$. If r of the N terms are removed, their normalised total mass is at most $re^{-B}/(1 + re^{-B}) \leq re^{-B}$.*

Proof. Each removed exponential is below e^{-B} times the largest, and the retained sum contains at least the largest term. □

As a numerical illustration, for $r = 100$ and $B = 40$ the discarded mass is at most 4.2×10^{-16}. A weight that is not finite is a typed numerical failure, and replacing it by zero would hide a support or density-ratio error. A complete update checks finiteness before normalisation, positive total mass after it, and the ancestor indices after resampling. If a check fails, the estimate of the normalising constant is withheld.

Floating-point sums. The classical forward bound for summation gives the accuracy of long reductions.

Proposition 4.57 (8.6 — summation bound). *Let u be the unit roundoff, $\gamma_k = ku/(1 - ku)$ and $ku < 1$. Sequential floating-point summation of $x_1, \dots, x_n$ satisfies $|\mathrm{fl}(\sum_i x_i) - \sum_i x_i| \leq \gamma_{n-1} \sum_i |x_i|$. Pairwise summation has the same form with $\gamma_{\lceil \log_2 n \rceil}$.*

This is the classical forward bound and is cited. Numerical illustration: for $n = 10^6$ terms in FP32 ($u = 2^{-24}$) the sequential bound is $\gamma_{n-1} = 0.0634$ and the pairwise bound is $\gamma_{20} = 1.2 \times 10^{-6}$. In BF16, $u = 2^{-8}$, and the condition $ku < 1$ already fails for $n \geq 257$. Accumulation is therefore kept in FP32, and the reduction order is fixed where bitwise agreement is required. The base model sums gradients across ranks in FP32 for this reason: the bound is informative in FP32 and uninformative in BF16.

Bitwise reproducibility needs a fixed reduction tree, deterministic kernels and fixed random streams. Statistical reproducibility needs independent seeds with stable distributions, and its tolerance is declared before evaluation.

Assembling a budget. For a reported log density, a conservative decomposition by the triangle inequality is

$$|\widehat{\ell} - \ell| \leq E_{\mathrm{state}} + E_{\mathrm{quad}} + E_{\mathrm{trace}} + E_{\mathrm{round}} + E_{\mathrm{chart}},$$

for the trajectory error propagated into the divergence, the time quadrature, the probe noise, the floating-point reductions and the chart constants. Proposition 3.9 bounds the first two together from the solver defect ε_f and the divergence error ε_d. Proposition 3.8 gives the variance behind the third. For a guided endpoint statistic, particle error and selection error are added as separate terms.

Each source has its own refinement test. The ODE state is tested with a tighter tolerance or a smaller largest step, comparing the endpoint state in the intrinsic and decoded metrics. The divergence quadrature is tested with denser or higher-order integration on a fixed trajectory, comparing the integrated log-density correction. The trace estimator is tested with more independent probes and, for a small state, an exact trace, comparing the probe mean, the variance and the exact error.

The contact solve is tested with a tighter dual gap and more iterations, comparing the objective, the marginals and the thresholded contacts. The particle approximation is tested with more particles

and independent complete repeats, comparing the normaliser, the endpoint statistic, the effective sample size and the ancestry. Floating-point error is tested with wider accumulation and a fixed reduction tree, comparing the residual and the final decision.

A tolerance is an instruction to an algorithm, and a refinement study measures the error. Solver tolerances, Sinkhorn stopping thresholds and linear residual targets become evidence only when a refinement study compares the scientific quantities that result. A report therefore gives the configured tolerance and the realised change under refinement.

4.9. Sensitivity and identifiability

This section asks which parameters of a model with a finite parameter vector ϑ and declared observables $\mathcal{O}(\vartheta) = \mathbb{E}_\vartheta T(X)$ can be told apart from data, and how far an output moves when an input moves. The rank of the Jacobian of $\mathcal{O}$ gives a sufficient identifiability test (Theorem 9.1), and the same matrix gives the covariance of a linearised inverse (Proposition 9.2). A selection step is discontinuous and has an exact stability radius (Proposition 9.3). Bounded and Lipschitz observables are insensitive to small changes of a law (Proposition 9.4).

Definition 4.58. A parameter ϑ_0 is *locally identifiable* through $\mathcal{O}$ if some neighbourhood U of ϑ_0 exists in which $\mathcal{O}(\vartheta) = \mathcal{O}(\vartheta_0)$ implies $\vartheta = \vartheta_0$. Structural identifiability concerns the map $\mathcal{O}$. Practical identifiability also concerns noise, design and sample size.

Theorem 4.59 (9.1 — rank condition). *If $\mathcal{O} : \mathbb{R}^p \to \mathbb{R}^r$ is continuously differentiable near ϑ_0, $r \geq p$, and $J_\mathcal{O}(\vartheta_0)$ has column rank p, then ϑ_0 is locally identifiable through $\mathcal{O}$.*

Proof. Full column rank means that some p rows of $J_\mathcal{O}(\vartheta_0)$ form a nonsingular square matrix. Let $\widetilde{\mathcal{O}}$ keep those components. The inverse function theorem makes $\widetilde{\mathcal{O}}$ one-to-one near ϑ_0. Equal $\mathcal{O}$ implies equal $\widetilde{\mathcal{O}}$, hence equal parameters. □

The rank condition is sufficient. It is not necessary, because $\mathcal{O}(\vartheta) = \vartheta^3$ is injective near zero with a zero derivative there. A small singular value of $J_\mathcal{O}$ says that some parameter combination changes the observables only weakly to first order. It does not show exact nonidentifiability.

Proposition 4.60 (9.2 — covariance of a linearised inverse). *Let $\widehat{\mathcal{O}} = \mathcal{O}(\vartheta_0) + J_\mathcal{O}(\vartheta - \vartheta_0) + \epsilon$ with $\mathbb{E}\epsilon = 0$, $\mathrm{Cov}(\epsilon) = \Sigma_\mathcal{O}$ and $J_\mathcal{O}$ of full column rank. The generalised least-squares estimator has covariance $(J_\mathcal{O}^\top \Sigma_\mathcal{O}^{-1} J_\mathcal{O})^{-1}$. If the smallest eigenvalue of $\mathcal{I}_\mathcal{O} = J_\mathcal{O}^\top \Sigma_\mathcal{O}^{-1} J_\mathcal{O}$ tends to zero, the largest variance diverges.*

Proof. The estimator is $(J^\top \Sigma^{-1} J)^{-1} J^\top \Sigma^{-1} \widehat{\mathcal{O}}$ plus a constant, and its covariance follows from $\mathrm{Cov}(\epsilon) = \Sigma_\mathcal{O}$. The largest eigenvalue of the inverse is the reciprocal of the smallest eigenvalue of $\mathcal{I}_\mathcal{O}$. □

The matrix $\mathcal{I}_\mathcal{O}$ is a design information matrix for the chosen observables. It equals the Fisher information only under a likelihood whose score gives the same expression. A new observable adds information only along a direction that the others constrain weakly, so designs are compared by the eigenvectors of $\mathcal{I}_\mathcal{O}$, because the trace sums the directions and hides a weak one.

Proposition 4.61 (9.3 — stability of a top-N set). *Let sorted scores satisfy $S_{(1)} \geq \cdots \geq S_{(M)}$ and let the margin be $\Delta_{\mathrm{sel}} = S_{(N)} - S_{(N+1)} > 0$. If every perturbation obeys $\|\delta S\|_\infty < \Delta_{\mathrm{sel}}/2$, the selected set is unchanged.*

Proof. For a selected i and an unselected j, $S_i - S_j \geq \Delta_{\text{sel}}$. After the perturbation $(S_i + \delta S_i) - (S_j + \delta S_j) \geq \Delta_{\text{sel}} - 2\|\delta S\|_\infty > 0$. □

An admission indicator $A = \mathbf{1}\{S \geq \tau\}$ breaks the smooth delta method. If S has a density f_S continuous at τ, then $\mathrm{d}\,\mathbb{P}(S \geq \tau)/\mathrm{d}\tau = -f_S(\tau)$, so a high density at the threshold makes the yield sensitive even when the score is accurate. When the margin Δ_{sel} is small, uncertainty is propagated by rerunning the whole selector on bootstrap or posterior score draws.

Proposition 4.62 (9.4 — bounded and Lipschitz observables). *If* $\mathrm{TV}(P, Q) \leq \delta$ *and* $a \leq T \leq b$, *then* $|\mathbb{E}_P T - \mathbb{E}_Q T| \leq (b - a)\delta$. *If* T *is* L_T*-Lipschitz under the ground metric of the Wasserstein distance* W_1, *then* $|\mathbb{E}_P T - \mathbb{E}_Q T| \leq L_T W_1(P, Q)$.

Proof. Scale T to $f = (T - a)/(b - a) \in [0, 1]$. The variational definition of total variation gives $|\mathbb{E}_P f - \mathbb{E}_Q f| \leq \delta$. The second bound is the Kantorovich–Rubinstein duality for Lipschitz functions, used as stated. □

4.10. Standing of the results

Every result in this supplement is proved in full for the stated conditions, except Theorem 4.14, which is given as a sketch, and the classical theorems that are used as stated and marked cited. These are the Markov property of the reversed process (Theorem 3.6), the Wasserstein duality (Section S1.5), the Woodbury identity and the Karush–Kuhn–Tucker condition (Section S1.7), the summation bound (Proposition 8.6) and the Kantorovich–Rubinstein duality (Proposition 9.4).

Theorem 4.7 establishes unbiased normaliser estimation with carried weights. Theorem 4.10 gives the exact-twist result. Example 4.8 and Propositions 4.12 and 4.13 distinguish unbiasedness from convergence and describe the limitations of systematic resampling. The kernel, potential and horizon are written K_j, Ψ_j and M throughout.

5. Scope, Ethics, Licence, and Declarations

5.1. Scope, Dual Use, and Licensing

Scope. The work is computational and contains no wet-lab step. Predictions of the full system remain predictions until an experiment tests them. A comparison with another group uses its published numbers, names the differences in protocol and states that the panels are not matched.

Dual use and biosafety. A tool that designs binders can be misused, as can any generative model for biochemistry. A drug-discovery model proposed toxic molecules once its objective was inverted [43]. AI-redesigned variants of proteins of concern evaded the screens of nucleic acid synthesis providers until those screens were patched [44]. Synthesis is therefore the control point, and one proposal keeps synthesis records in repositories that are queried only in emergencies [45]. The study takes three positions. The base model contains no component that conditions on a target or emits an antibody sequence. The manuscript prints no designed sequence, and the code package releases none: its only data are synthetic complexes. Any later release of an antigen-conditioned generator would first need a review of misuse and a screen of requested targets against curated

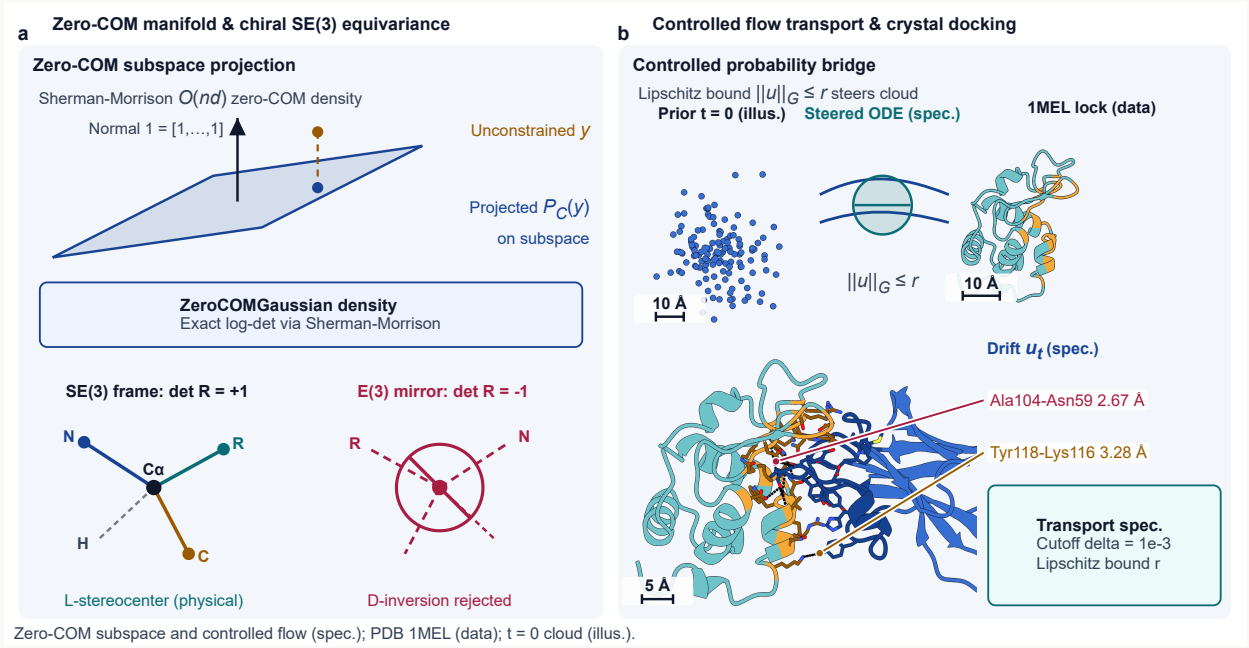


Figure 1: Zero-COM Sherman–Morrison subspace projection preserves chiral SE(3) geometry along trust-ball flow transport to PDB 1MEL. a, Helmert zero-center-of-mass projection alongside stereochemical handedness ($\det R = +1$ preserving L-amino-acid centers against rejected D-inversion) and the Sherman–Morrison inverse covariance block for shared translations. **b,** Controlled flow transport trajectory from Gaussian prior ($t = 0.00$) through $t = 0.50$ within Riemannian trust ball $\|u\|_G \leq r$ converging to the PDB 1MEL interface lock (Ala104–Asn59 at 2.67 Å). Algebraic identities match formal Lean propositions and native coordinates reflect PDB 1MEL, while intermediate frames illustrate the flow geometry. COM, center of mass; PDB, Protein Data Bank.

lists of controlled pathogen accessions. These positions reduce risk and do not amount to a safety claim.

Licences and data. The code package is licensed under the GNU Affero General Public License, version 3 (AGPL-3.0). Its section 13 requires that a modified version offered to users over a network make its source available. The licence covers code and grants nothing on weights or data. Third-party libraries keep their own licences, and the main ones are named in Technical Report I (Section "Computing, software and availability"). Third-party papers appear as citations only. The licence of trained weights follows the licences of the training data (Technical Report I (Section "Sources and licences")). The base model reads public protein structures only and no personal, clinical or patient data, and each downloaded file is stored with its SHA-256 digest (Technical Report I (Section "Shard format and sampling")).

5.2. Declarations

Author contributions. D.K. is the sole author and carried out the conception, design, code and writing of the work.

Data and code availability. The training code, its tests, the configuration files and the container definition are supplied as Supplementary Information and described in Supplementary Section S4. The code is licensed under the GNU Affero General Public License, version 3. Each dataset stays with its provider, and this manuscript redistributes none. The sources and their licences are listed in Technical Report I (Section "Data") and Supplementary Section S2.

Artwork. The protein pictures in Figures 1, 2 and 3 (and the accompanying plates of Technical Report I) were rendered with UCSF ChimeraX from deposited coordinates of the Protein Data

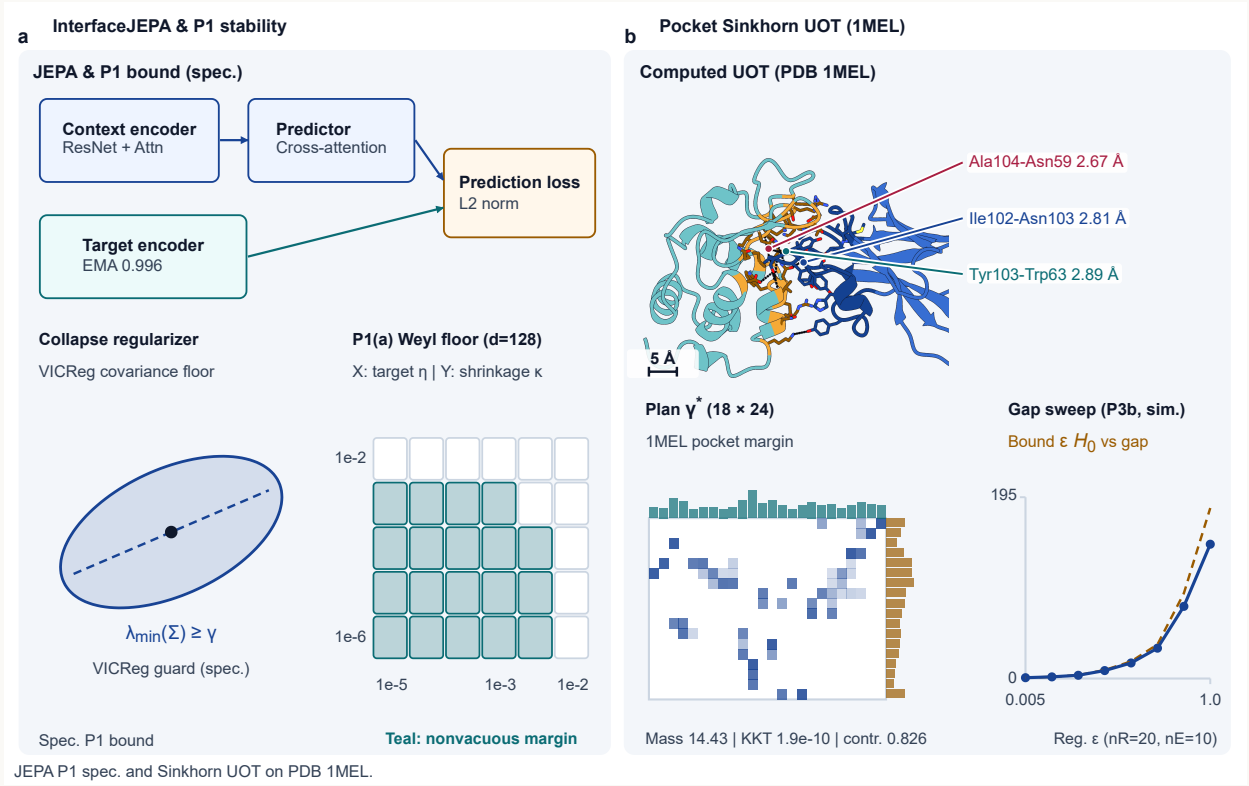
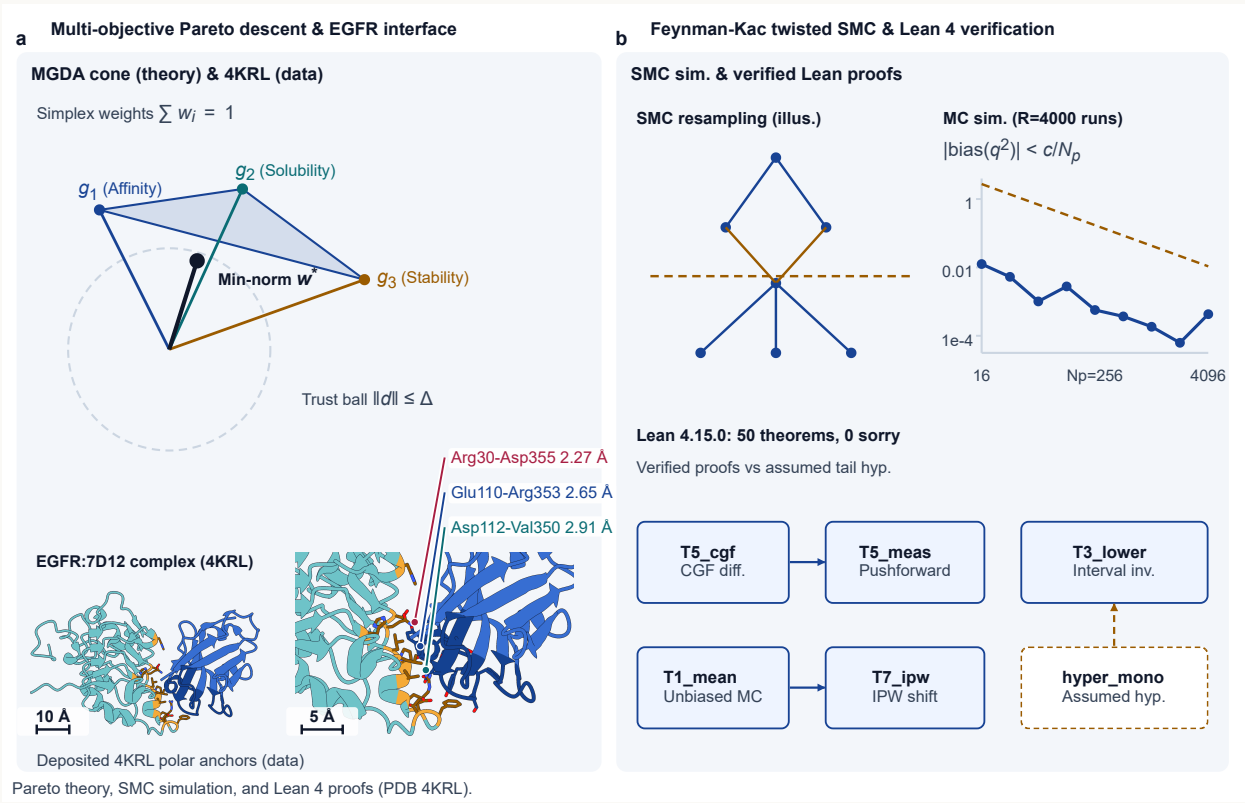


Figure 2: Unbalanced log-domain Sinkhorn transport couples PDB 4KRL pocket geometry with VICReg-regularized Interface-JEPA representations. a, Unbalanced optimal transport formulation applied to the PDB 4KRL 7D12–EGFR pocket (Arg30–Asp355 at 2.27 Å, Glu110–Arg353 at 2.65 Å, and Asp112–Val350 at 2.91 Å) beside the simulated dual-gap convergence trace (P1). **b,** Dual online and exponential-moving-average target encoder architecture for Interface-JEPA with typed prediction heads, eigenvalues above the $\lambda_{\min}(\Sigma) \geq 0.10$ VICReg variance-covariance margin (P3 simulation), and the 55-pair PDB 1MEL native contact matrix. Structural contacts derive from PDB 4KRL and PDB 1MEL; P1 and P3 traces come from deterministic simulations over synthetic fixtures without trained neural weights. EMA, exponential moving average; JEPA, joint-embedding predictive architecture; PDB, Protein Data Bank; UOT, unbalanced optimal transport.



Pareto theory, SMC simulation, and Lean 4 proofs (PDB 4KRL).

Figure 3: Pareto MGDA guidance and Feynman–Kac particle steering connect PDB 4KRL pocket geometry to 50 verified Lean 4 theorems. a, Minimum-norm common descent cone inside the decoder pullback metric $G_\phi = J_h^\top W_h J_h + \tau I$ paired with PDB 4KRL interface anchors (Arg30–Asp355 at 2.27 Å, Glu110–Arg353 at 2.65 Å, and Asp112–Val350 at 2.91 Å). **b,** Feynman–Kac sequential Monte Carlo particle trajectory fan with stratified resampling below the effective sample size floor $N_p/2$ beside the verified `lean/Omniverso` proof ledger (50 theorems and 50 definitions across six files with zero uses of `sorry`). Pocket coordinates come from PDB 4KRL and proof counts match live Lean files; guidance cones and particle branches illustrate algorithmic behavior. ESS, effective sample size; MGDA, multiple-gradient descent algorithm; PDB, Protein Data Bank; SMC, sequential Monte Carlo.

Bank (PDB entries 1MEL and 4KRL).

References

[1] Mansoor Ahmed and Murray Patterson. AgForce Enables Antigen-conditioned Generative Antibody Design. *arXiv*, 2026. URL: https://arxiv.org/abs/2605.21610.

[2] Lénaïc Chizat, Gabriel Peyré, Bernhard Schmitzer, and François-Xavier Vialard. Scaling algorithms for unbalanced optimal transport problems. *Mathematics of Computation*, 87(314):2563–2609, 2018. doi: 10.1090/mcom/3303. arXiv:1607.05816.

[3] Gabriel Peyré and Marco Cuturi. Computational Optimal Transport. *arXiv*, 2018. URL: https://arxiv.org/abs/1803.00567.

[4] Thibault Séjourné, Jean Feydy, François-Xavier Vialard, Alain Trouvé, and Gabriel Peyré. Sinkhorn Divergences for Unbalanced Optimal Transport. *arXiv*, 2019. URL: https://arxiv.org/abs/1910.12958.

[5] Kieran Didi, Zuobai Zhang, Guoqing Zhou, Danny Reidenbach, Zhonglin Cao, Sooyoung Cha, Tomas Geffner, Christian Dallago, Jian Tang, Michael M. Bronstein, Martin Steinegger, Emine Kucukbenli, Arash Vahdat, and Karsten Kreis. Scaling Atomistic Protein Binder Design with Generative Pretraining and Test-Time Compute. *arXiv*, 2026. URL: https://arxiv.org/abs/2603.27950.

[6] James Dunbar, Konrad Krawczyk, Jinwoo Leem, Terry Baker, Angelika Fuchs, Guy Georges, Jiye Shi, and Charlotte M. Deane. SAbDab: the structural antibody database. *Nucleic Acids Research*, 42(D1): D1140–D1146, 2014. doi: 10.1093/nar/gkt1043. PMID:24214988.

[7] Gemma L. Gordon, Henriette L. Capel, Bora Guloglu, Eve Richardson, Ryan L. Stafford, and Charlotte M. Deane. A comparison of the binding sites of antibodies and single-domain antibodies. *Frontiers in Immunology*, 14:1231623, 2023. doi: 10.3389/fimmu.2023.1231623. PMID:37533864.

[8] Yu-Xi Tsai, Ning-En Chang, Klaus Reuter, Hao-Ting Chang, Tzu-Jing Yang, Sören von Bülow, Vidhi Sehrawat, Noémie Zerrouki, Matthieu Tuffery, Michael Gecht, Isabell Louise Grothaus, Lucio Colombi Ciacchi, Yong-Sheng Wang, Min-Feng Hsu, Kay-Hooi Khoo, Gerhard Hummer, Shang-Te Danny Hsu, Cyril Hanus, and Mateusz Sikora. Rapid simulation of glycoprotein structures by grafting and steric exclusion of glycan conformer libraries. *Cell*, 187(5):1296–1311.e26, 2024. doi: 10.1016/j.cell.2024.01.034. PMID:38428397.

[9] Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-Supervised Learning from Images with a Joint-Embedding Predictive Architecture. *arXiv*, 2023. URL: https://arxiv.org/abs/2301.08243.

[10] S. Rabinowitz, P. Nigam, N. Santolla, and C. T. Ford. Kukulu: Diffusion-Based Reconstruction of Antibody CDR Loops using a Structure-Aware Joint Embedding Predictive Architecture. *bioRxiv*, 2026. doi: 10.64898/2026.08.03.742168.

[11] Ralph Allan Bradley and Milton E. Terry. Rank Analysis of Incomplete Block Designs: I. The Method of Paired Comparisons. *Biometrika*, 39(3/4):324, 1952. doi: 10.2307/2334029.

[12] Adrien Bardes, Jean Ponce, and Yann LeCun. VICReg: Variance-Invariance-Covariance Regularization for Self-Supervised Learning. *arXiv*, 2021. URL: https://arxiv.org/abs/2105.04906.

[13] Michael S. Albergo, Nicholas M. Boffi, and Eric Vanden-Eijnden. Stochastic Interpolants: A Unifying Framework for Flows and Diffusions. *arXiv*, 2023. URL: https://arxiv.org/abs/2303.08797.

[14] Tomas Geffner, Kieran Didi, Zhonglin Cao, Danny Reidenbach, Zuobai Zhang, Christian Dallago, Emine Kucukbenli, Karsten Kreis, and Arash Vahdat. La-Proteina: Atomistic Protein Generation via Partially Latent Flow Matching. *arXiv*, 2025. URL: https://arxiv.org/abs/2507.09466.

[15] Georgios Arvanitidis, Lars Kai Hansen, and Søren Hauberg. Latent Space Oddity: on the Curvature of Deep Generative Models. *arXiv*, 2017. URL: https://arxiv.org/abs/1710.11379.

[16] Jean-Antoine Désidéri. Multiple-gradient descent algorithm (MGDA) for multiobjective optimization. *Comptes Rendus. Mathématique*, 350(5-6):313–318, 2012. doi: 10.1016/j.crma.2012.03.014.

[17] Ozan Sener and Vladlen Koltun. Multi-Task Learning as Multi-Objective Optimization. *arXiv*, 2018. URL: https://arxiv.org/abs/1810.04650.

[18] Yinghua Yao, Yuangang Pan, Jing Li, Ivor Tsang, and Xin Yao. PROUD: PaRetO-gUided diffusion model for multi-objective generation. *Machine Learning*, 113:6511–6538, 2024. doi: 10.1007/s10994-024-06575-2. arXiv:2407.04493.

[19] Christian A. Naesseth, Fredrik Lindsten, and Thomas B. Schön. Elements of Sequential Monte Carlo. *Foundations and Trends® in Machine Learning*, 12(3):187–306, 2019. doi: 10.1561/2200000074. arXiv:1903.04797.

[20] Luhuan Wu, Brian L. Trippe, Christian A. Naesseth, David M. Blei, and John P. Cunningham. Practical and Asymptotically Exact Conditional Sampling in Diffusion Models. *arXiv*, 2023. URL: https://arxiv.org/abs/2306.17775.

[21] Marta Skreta, Tara Akhound-Sadegh, Viktor Ohanesian, Roberto Bondesan, Alán Aspuru-Guzik, and Arnaud Doucet. Feynman-Kac Correctors in Diffusion: Annealing, Guidance, and Product of Experts. *arXiv*, 2025. URL: https://arxiv.org/abs/2503.02819.

[22] Konstantin Mark, Leonard Galustian, Maximilian P.-P. Kovar, and Esther Heid. Feynman-Kac-Flow: Inference Steering of Conditional Flow Matching to an Energy-Tilted Posterior. *arXiv*, 2025. URL: https://arxiv.org/abs/2509.01543.

[23] Erik Hartman, Jonas Wallin, Johan Malmström, and Jimmy Olsson. Controllable protein design with particle-based Feynman-Kac steering. *arXiv*, 2025. URL: https://arxiv.org/abs/2511.09216.

[24] Filippo Stocco, Michele Garibbo, and Noelia Ferruz. Steering generative models for protein design: Aligning and conditioning strategies. *Current Opinion in Structural Biology*, 98:103250, 2026. doi: 10.1016/j.sbi.2026.103250.

[25] Adhithyan Kalaivanan, Zheng Zhao, Jens Sjölund, and Fredrik Lindsten. Interacting particle guidance for sampling reward-tilted generative priors. *arXiv*, 2026. URL: https://arxiv.org/abs/2609.37227.

[26] Mathieu Gerber, Nicolas Chopin, and Nick Whiteley. Negative association, ordering and convergence of resampling methods. *The Annals of Statistics*, 47(4):2236–2260, 2019. doi: 10.1214/18-aos1746. arXiv:1707.01845.

[27] Víctor Elvira, Luca Martino, and Christian P. Robert. Rethinking the Effective Sample Size. *International Statistical Review*, 90(3):525–550, 2022. doi: 10.1111/insr.12500.

[28] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. SDEdit: Guided Image Synthesis and Editing with Stochastic Differential Equations. *arXiv*, 2021. URL: https://arxiv.org/abs/2108.01073.

[29] Junhao Xiong, Ishan Gaur, Maria Lukarska, Hunter Nisonoff, Luke M. Oltrogge, David F. Savage, and Jennifer Listgarten. ProteinGuide: On-the-fly property guidance for protein sequence generative models. *arXiv*, 2025. URL: https://arxiv.org/abs/2505.04823.

[30] Justin Sanders, Luca Giancardo, Lan Guo, Yue Zhao, Kemal Sonmez, Nina Cheng, and Melih Yilmaz. Conditional generation of antibody sequences with classifier-guided germline-absorbing discrete diffusion. *arXiv*, 2026. URL: https://arxiv.org/abs/2605.06720.

[31] Imee Sinha, Samuel Stanton, Stephen Lillington, Sarah A. Robinson, Santrupti Nerli, Karina Zadorozhny, Joseph Kleinhenz, Homa MohammadiPeyhani, Michael Dillon, Yongmei Chen, Jack Bevers, Yan Wu, Andrew M. Watkins, Henri Dwyer, Richard Bonneau, Kyunghyun Cho, Franziska Seeger, Vladimir Gligorijević, and Simon Kelow. CDR Conformation Aware Antibody Sequence Design with ConformAb. *bioRxiv*, 2025. doi: 10.1101/2025.11.12.688095.

[32] Gur Yaari, Jason A. Vander Heiden, Mohamed Uduman, Daniel Gadala-Maria, Namita Gupta, Joel N. H. Stern, Kevin C. O'Connor, David A. Hafler, Uri Laserson, Francois Vigneault, and Steven H. Kleinstein. Models of Somatic Hypermutation Targeting and Substitution Based on Synonymous Mutations from High-Throughput Immunoglobulin Sequencing Data. *Frontiers in Immunology*, 4, 2013. doi: 10.3389/fimmu.2013.00358. PMID:24298272.

[33] Jian Ye, Ning Ma, Thomas L. Madden, and James M. Ostell. IgBLAST: an immunoglobulin variable domain sequence analysis tool. *Nucleic Acids Research*, 41(W1):W34–W40, 2013. doi: 10.1093/nar/gkt382. PMID:23671333.

[34] David Prihoda, Jad Maamary, Andrew Waight, Veronica Juan, Laurence Fayadat-Dilman, Daniel Svozil, and Danny A. Bitton. BioPhi: A platform for antibody design, humanization, and humanness evaluation based on natural antibody repertoires and deep learning. *mAbs*, 14(1):2020203, 2022. doi: 10.1080/19420862.2021.2020203. PMID:35133949.

[35] Jasmin F. Sydow, Florian Lipsmeier, Vincent Larraillet, Maximiliane Hilger, Bjoern Mautz, Michael Mølhøj, Jan Kuentzer, Stefan Klostermann, Juergen Schoch, Hans R. Voelger, Joerg T. Regula, Patrick Cramer, Apollon Papadimitriou, and Hubert Kettenberger. Structure-Based Prediction of Asparagine and Aspartate Degradation Sites in Antibody Variable Regions. *PLoS ONE*, 9(6):e100736, 2014. doi: 10.1371/journal.pone.0100736. PMID:24959685.

[36] Edward P. Harvey, Jung-Eun Shin, Meredith A. Skiba, Genevieve R. Nemeth, Joseph D. Hurley, Alon Wellner, Ada Y. Shaw, Victor G. Miranda, Joseph K. Min, Chang C. Liu, Debora S. Marks, and Andrew C. Kruse. An in silico method to assess antibody fragment polyreactivity. *Nature Communications*, 13(1):7554, 2022. doi: 10.1038/s41467-022-35276-4. PMID:36477674.

[37] Xinyan Zhao, Yi-Ching Tang, Akshita Singh, Victor J Cantu, KwanHo An, Junseok Lee, Adam E Stogsdill, Ibraheem M Hamdi, Ashwin Kumar Ramesh, Zhiqiang An, Xiaoqian Jiang, and Yejin Kim. AbBiBench: A Benchmark for Antibody Binding Affinity Maturation and Design. *arXiv*, 2025. URL: https://arxiv.org/abs/2506.04235.

[38] Brian D.O. Anderson. Reverse-time diffusion equation models. *Stochastic Processes and their Applications*, 12(3):313–326, 1982. doi: 10.1016/0304-4149(82)90051-5.

[39] Daniel Kuhn, Peyman Mohajerin Esfahani, Viet Anh Nguyen, and Soroosh Shafieezadeh-Abadeh. Wasserstein Distributionally Robust Optimization: Theory and Applications in Machine Learning. *arXiv*, 2019. URL: https://arxiv.org/abs/1908.08729.

[40] G. L. Nemhauser, L. A. Wolsey, and M. L. Fisher. An analysis of approximations for maximizing submodular set functions—I. *Mathematical Programming*, 14(1):265–294, 1978. doi: 10.1007/bf01588971.

[41] Anastasios N. Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal Risk Control. *arXiv*, 2022. URL: https://arxiv.org/abs/2208.02814.

[42] Aaditya Ramdas, Peter Grünwald, Vladimir Vovk, and Glenn Shafer. Game-Theoretic Statistics and Safe Anytime-Valid Inference. *Statistical Science*, 38(4):576–601, 2023. doi: 10.1214/23-sts894. arXiv:2210.01948.

[43] Fabio Urbina, Filippa Lentzos, Cédric Invernizzi, and Sean Ekins. Dual Use of Artificial Intelligence-powered Drug Discovery. *Nature Machine Intelligence*, 4:189–191, 2022. doi: 10.1038/s42256-022-00465-9.

[44] Bruce J. Wittmann, Tessa Alexanian, Craig Bartling, Jacob Beal, Adam Clore, James Diggans, Kevin Flyangolts, Bryan T. Gemler, Tom Mitchell, Steven T. Murphy, Nicole E. Wheeler, and Eric Horvitz. Strengthening nucleic acid biosecurity screening against generative protein design tools. *Science*, 390:82–87, 2025. doi: 10.1126/science.adu8578.

[45] David Baker and George Church. Protein design meets biosecurity. *Science*, 383:349, 2024. doi: 10.1126/science.ado1671.