Supervised Cross-Entropy for Conformer Generation

Minimizing CE Between Teacher and Student Graphs by Optimizing Y

Abstract

This document summarizes how to use a supervised cross-entropy (CE) objective over pairwise-distance graphs to train molecular conformer coordinates Y directly. The method replaces or augments the usual unsupervised CE(P, Q(Y)) with CE(Q*, Q(Y)), where Q* is derived from a teacher conformation (ground truth or ETKDG+MMFF). It is rigid-body invariant, integrates smoothly with manifold-learning priors and lightweight physics, and identifies the target conformation when applied on the full pair set.

Generated on 2025-08-23

1. What is this?

We propose a supervised training scheme for small-molecule 3D conformer generation where the optimization variable is the 3D coordinates Y (N×3). The objective is the cross-entropy between a TEACHER graph Q* (derived from a reference/teacher conformation) and a STUDENT graph Q(Y) (derived from the current coordinates Y):

CE(Q*, Q(Y)) = - Σ_{i,j} [Q*_{ij} log Q_{ij}(Y) + (1 - Q*_{ij}) log(1 - Q_{ij}(Y))].

Here Q_{ij}(Y) = 1 / (1 + a * d_{ij}(Y)^{2b}) with distance d_{ij}(Y) = || y_i - y_j ||_2. Parameters (a, b) are set from (min_dist, spread) as in UMAP-style kernels. Optionally combine supervised CE with unsupervised CE(P, Q(Y)) from D1+hop and with light physics (bond springs, repulsion).

2. Why CE on graphs (not MSE on coords)?

Rigid-body invariance: CE depends only on pairwise distances ⇒ no alignment needed.

Well-conditioned gradients: bounded probabilities provide stable attractive/repulsive signals.

Identifiability: CE→0 on full pairs ⇒ identical distance matrices ⇒ same conformation up to rigid transform / mirror.

Partial supervision: supervise hop≤3 to focus on local geometry; leave long-range to manifold/physics.

Seamless with manifold learning: same CE form as your unsupervised objective; code reuse.

Easy to mix with physics: add small bond/1–3/1–4 springs and mild nonbonded repulsion.

3. Mathematics

Teacher: Q*_{ij} = 1 / (1 + a d*_{ij}^{2b}) from teacher coords Y*. Student: Q_{ij}(Y) = 1 / (1 + a d_{ij}(Y)^{2b}).

Total loss: L = w_sup CE(Q*,Q(Y)) + w_unsup CE(P,Q(Y)) + λ_b L_bond(Y) + λ_r L_rep(Y).

Bond springs: Σ (||y_i - y_j|| - L_ij)^2 over bonds or hop≤3 pairs. Repulsion: penalty when hop>k pairs are closer than a cutoff.

Early Exaggeration, smooth-k, fuzzy-union from your pipeline apply unchanged.

4. Training recipe

1) Build Q*: from ground truth Y* (or ETKDG+MMFF) with same (a,b); optionally mask hop≤3.

2) Build P: from D1 with hop gating, smooth-k per-row bandwidths, fuzzy union.

3) Loop: compute CE_sup & grad from (Q*,Y); optionally CE_uns from (P,Y); add physics grads; update Y with small trust-region step; center+rescale; adapt (a,b) occasionally.

4) Evaluate: RMSD vs teacher (Kabsch), bond RMSE, clashes, CE curves.

5. Advantages over alternatives

vs MSE: alignment-free training, scale-aware via (a,b), bounded loss, repulsive gradients via (1-Q*).

vs energy-only: injects global relational info; add physics as regularizer rather than sole driver.

vs contrastive: dense chemically meaningful supervision with hop-aware masks/weights.

6. Practical tips

Teacher choice: ETKDG + MMFF is a reasonable weak teacher; mask to hop≤3 if noisy.

Hop-aware weighting: multiply targets by (1 + α1*(hop==1) + α2*(hop==2)).

Plateau fixes: row-sharpening, CE-step backtracking line search, per-atom step cap (~0.2 Å), LR decay.

Chirality: add tiny chiral-volume/dihedral-sign regularizer if necessary.

Schedules: EE 10–30; w_sup warm-up 50–200; linear warm-up for λ_b, λ_r.

a,b: derive from min_dist∈[0.35,0.60], spread≈3; refit using 60th percentile of positive-pair distances.

7. Minimal drop-in code

```
# Precompute teacher graph:

Y_teacher = teacher_coords(mol, X_ref=None)

Q_star    = Q_from_coords(Y_teacher, a, b)

Q_star    = Q_star * (hop <= 3)  # optional


# In training:

CE_sup, g_sup = ce_and_grad(Q_star, Y, a, b)

CE_uns, g_uns = ce_and_grad(P_eff,  Y, a, b)  # optional

r = min(1.0, epoch / max(1, sup_warmup))

w_sup = w_sup0 * (1.0 + r*(w_sup_factor - 1.0))

g_bond = bond_spring_grad(Y, mol, bond_targets)

g_rep  = repulsion_grad(Y, H, cutoff=repulsion_cutoff, exclude_hop_le=exclude_hop_le)

g = w_sup*g_sup + w_unsup*g_uns + λ_b*g_bond + λ_r*g_rep

Y = Y - lr * g; Y = center_and_rescale(Y)
```

8. Evaluation & limitations

Metrics: RMSD (Kabsch), bond RMSE, ring planarity, clashes, CE curves, MMFF/UFF energy proxy, torsion profiles.

Limitations: mirror ambiguity for chiral molecules; full-matrix supervision can be brittle with noisy teachers.

Extensions: amortized decoder g_phi(h)->Y trained with CE; learned per-node bandwidths σ_i; soft hop gates; later add EGNN.

9. Summary

Supervised CE on distance graphs provides a principled, rigid-body-invariant way to train coordinates Y using a teacher conformation, while remaining compatible with your manifold prior and physics. When applied to the full pair set, CE→0 identifies the same conformation up to rigid transform/mirror; in practice, local (hop≤3) supervision with warm-ups and mild regularizers yields robust and chemically plausible geometries.