LEARNING MOLECULES AS CONFORMER SETS FOR NOVEL-SCAFFOLD SELECTION AT BILLION SCALE

Anonymous authors
Paper under double-blind review

ABSTRACT

Molecular representations based on 1D, 2D, or a single 3D structure do not capture the conformational flexibility that governs binding, limiting the discovery of active compounds with novel scaffolds. We propose ZAO, a foundation model that represents each molecule as a conformer set and encodes 10 conformers per molecule as a single input. It has 238 million parameters and is pretrained on a corpus of 11 billion conformers, 12.4 times the size of the largest published conformer corpus. ZAO's frozen representation is stable under conformer regeneration, remains effective in regions structurally distant from known active compounds, and is complementary to 2D fingerprints.

We evaluate this representation prospectively for the discovery of novel active compounds. From a 1.1-billion-molecule virtual library, we select 9,220 compounds across 31 targets, score them with absolute binding free-energy calculations, and compare the outcome with that of a Boltz-2 plus ESPSim pipeline. To the best of our knowledge, no prior evaluation of a molecular foundation model has measured the ability to select novel active compounds at this scale. Median per-target hit rates ($\Delta G \leq -10$ kcal/mol) are 26.4% versus 16.0%, with ZAO's rate higher on 22 of 30 targets. ZAO's advantage is greatest on targets regarded as difficult, including protein-protein interactions. At $\Delta G \leq -12$ kcal/mol, the gap widens: 14.6% versus 5.2%. ZAO and Boltz-2 agree only weakly, and the hit rate for compound-target pairs that both methods judge active is 53.0%, indicating that conformer-set representations can complement structure-based scoring. We also release ZAO's weights.

1 INTRODUCTION

Molecular shape and charge distribution govern biological activity (Nicholls et al., 2010 [54]; Bauer and Mackey, 2019 [8]). However, a drug-like molecule has no single fixed shape, and no one structure can stand in for it (Hawkins, 2017 [32]). The heavy-atom root-mean-square deviation (RMSD) between conformers of the same molecule averages 1.81 Å (Appendix A.1), longer than a carbon-carbon single bond (1.54 Å) and close to the 2.0 Å threshold conventionally used to judge docking success (Cole et al., 2005 [17]). Choosing one conformer discards this variation, and taking the most stable one as representative is unreliable too [32]. Therefore, the unit a molecular representation should handle is the conformer set. Yet existing models have no mechanism for aggregating that set inside the representation. Even when models handle sets, they average per-conformer predictions or learn the aggregation itself in the supervised stage (Section 5). Once that step sits outside the representation, what stays frozen holds the information of a single conformer and nothing more. We present ZAO, which places the aggregation inside the representation, reading 10 conformers per molecule in one pass to produce a 2,048-dimensional embedding. We train a 238-million-parameter model on 1.1 billion molecules and 11 billion conformers (Section 2).

Candidate selection from a large library divides into two routes: a ligand-centric route that selects candidates by matching their 3D shapes and electrostatic potentials to those of known active compounds, and a structure-based route that predicts complex structures and binding affinities. When shape-based ranking was compared with seven docking programs across 21 protein systems using the same active compounds and decoys, it was more consistent and often performed better [31]. We represent the ligand-centric route with ESPSim [10], which also measures similarity of electrostatic

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

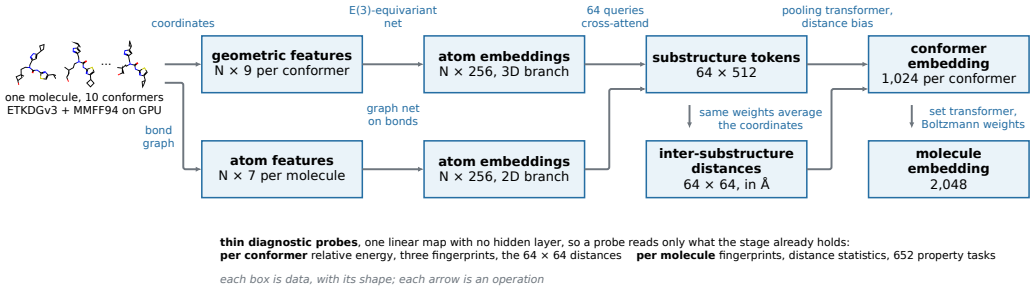

Figure 1: Architecture of the model. Each box is data, labeled with its tensor shape (N is the atom count of the molecule), and each arrow is an operation. The bottom block is what the auxiliary diagnostic heads read.

potential. On the structure-based side, the current representative is Boltz-2 [57]. It is reported as the first deep learning model to approach the accuracy of free-energy perturbation in binding-affinity estimation at more than 1,000 times the speed, and in a retrospective evaluation it is reported to outperform every method submitted to the CASP16 affinity task. We put ESPSim and Boltz-2 in series as our comparator (Section 3). ZAO never reads the protein structure.

This paper makes four contributions to molecular representation learning for drug discovery:

1. No prior evaluation of a molecular foundation model has measured selection performance at scale. We select candidates for 31 targets from 1.1 billion molecules and score them with absolute binding free-energy calculations. Candidate selection driven primarily by ZAO embeddings never reads the protein structure and outperforms the Boltz-ESPSim group, which does read it.
2. The hard part is finding active compounds structurally unlike the known ones. Of the 2,772 hit pairs identified by ZAO, the median 2D similarity to the nearest training positive is 0.300, and the fraction without a scaffold match has a per-target median of 100%.
3. Existing features do not retain much of the information that comes from 3D structure. In a linear-probe evaluation of 36 metrics, ZAO outperforms all three public 3D models, both language models, and the 2D descriptors.
4. Pretraining on conformer sets at this scale has not been attempted. We pretrain on 1.1 billion molecules and 11 billion conformers, 12.4 times the largest conformer count publicly reported for a dataset used in molecular-representation pretraining. We also release the pretrained weights of the 238-million-parameter ZAO model.

2 Model and Pretraining

The input is a conformer set: for each molecule we generate 10 conformers with ETKDGv3 [62, 75] and optimize them using MMFF94 [30]. With 11 billion conformers to generate, we do not adopt a learned conformer generator [35]. We give the model atom features and geometric features computed on the fly from the coordinates (see Appendix A.1 for the breakdown). Under SE(3) transformations of the coordinates, the input features are invariant, and under reflection only the single signed-volume dimension changes sign (Section 6).

As shown in Figure 1, the path that reads coordinates directly is an E(3)-equivariant graph network [65] and the path that reads only the bond graph is a 2D graph network [79]; hereafter we call them the 3D path and the 2D path. Since 2D connectivity is the same for all conformers, we compute it only once per molecule. Sixty-four learnable queries cross-attend to the atom representations of the 3D path; with those attention weights we read representations from both the 3D and the 2D paths and concatenate them.

With the same attention weights, a weighted average of the atom coordinates yields one centroid per query in 3D space, and the distances between these centroids form the 64×64 matrix of inter-substructure distances. Conformer Pooling passes a classification (CLS) token and the 64 substructure tokens through a transformer [74]; the distance matrix passes through a Gaussian basis and

is then added to the attention logits as a bias. The conformer-set aggregation applies Boltzmann weights derived from relative energies to the 10 conformer representations and then passes them through a transformer. Since this aggregation has no positional encoding, no conformer numbering, and no learnable parameters that vary with the conformer count, the released weights do not require 10 conformers (the 1-to-50 series in Section 4 relies on this property). During training, to prevent the model from memorizing a conformer set as a fixed unit, we draw a conformer count uniformly between 5 and 10 for each batch (see Appendix A.2 for the setting). We take the 2,048-dimensional output as the molecular representation.

The pretraining objective comprises four kinds of loss. Three are unsupervised: 1) per-conformer relative energy and the 64×64 distance matrix, 2) three fingerprint types at both the conformer and molecule levels, and 3) a term that constrains embedding-space distances to reproduce structural closeness as measured by fingerprints and shape. The fourth, multi-task prediction over 1.3 million molecules and 652 tasks, uses publicly available supervised data (ChEMBL [80], BindingDB [26]) (Appendices A.2, A.3, and E.1). In the implementation, these objectives are divided into 12 terms, and the total loss is $L(t) = \sum_k w_k(t)L_k(t)/s_k(t)$. Here, L_k is the loss for each objective, s_k is the exponential moving average of the magnitude of that loss (fixed at 1 for the six fingerprint terms), and w_k is a weight updated every 1,000 steps based on performance on the validation data, with larger weights for worse-performing terms. Although the losses span orders of magnitude, division by s_k absorbs those differences, eliminating the need to adjust the weights between objectives by hand. We record the breakdown of the terms, the update formula, and the fixed coefficients in Appendix A.2.

We train on the entire corpus using 80 H100s. We run 234,000 steps at an average of 7.5 conformers per molecule, for a total of about 13.5 billion conformers seen during training.

3 Novel-Scaffold Selection at Billion Scale

3.1 Targets and Scoring

When the target changes, the binding mechanism and the effective chemotypes change with it; only by measuring across diverse targets does the usable range of a foundation model become visible. In allocating scale, we prioritize the number of targets, and for each of 31 targets we select compounds believed to be active against that target. At this scale and granularity, binding free-energy calculations are the practical option for estimating binding strength across a large number of novel compounds. However, their reliability varies greatly by target, and for that reason we restrict the campaign to targets where these calculations correlate well with measured activity. We evaluate 40-50 ligands for each of 121 targets, narrow the target set to the 40 targets with both the Pearson and rank correlations above 0.4, and, within those, take as our subjects the 31 targets with ZAO group calculations (median correlation coefficient 0.528 over the 29 targets where correlation can be measured). Appendix B.1 outlines target selection and why we choose absolute binding free-energy calculations, and Appendix B.3 sets out the accepted range and pooling rules.

3.2 Selection Pipeline and Comparison Groups

As shown in Figure 2a, for each target we fit two regression models independently, a gradient-boosted tree (CatBoost [60]) and ridge regression via a Nyström approximation [76] of the radial basis function (RBF) kernel, and take their union. We build these models from the 2,048-dimensional ZAO embedding, four RDKit [61] fingerprint families, and 217 physicochemical descriptors. We impose the novelty requirement at the candidate-selection stage. In all, the ZAO group comprises 9,288 pairs across 31 targets (9,220 compounds), and it uses no protein 3D structure in candidate selection. For training labels see Appendix A.4; for model construction, the threshold at each stage, and the steps we use to control false positives at this scale, see Appendix C.1. Throughout this paper, all 2D similarities are Tanimoto coefficients on Morgan fingerprints (radius 2; 2,048 bits) [63].

For the Boltz-ESPSim comparison group, we narrow first with ESPSim [10], which measures the agreement of electrostatic potential and molecular shape, and then take the candidates that Boltz-2 [57], a model that predicts complex structure and binding affinity, judges to bind. Only the top 1,000 candidates per target after ESPSim narrowing go to Boltz-2, and since there is no 2D similarity constraint, candidates lacking novelty can be selected. We compare selection pipelines; we do

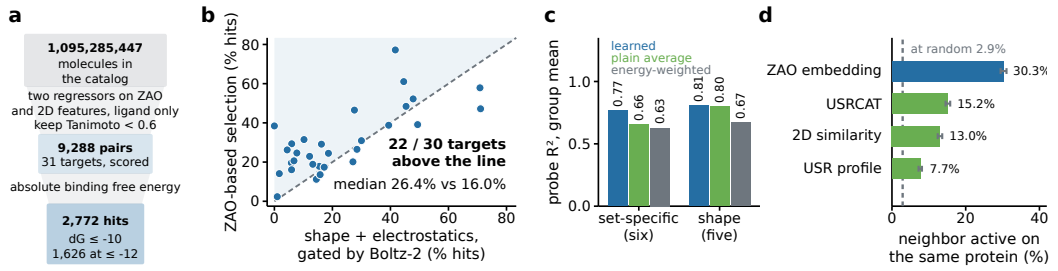

Figure 2: Selection results and content of the frozen representation. (a) The funnel from 1.1 billion molecules through the 9,288 pairs (31 targets) submitted for scoring down to the 2,772 pairs with $\Delta G \leq -10$ kcal/mol, of which 1,626 reach $\Delta G \leq -12$ kcal/mol. (b) Paired comparison of per-target hit rates. Each dot is one target. (c) Effect of conformer-set aggregation. Learned aggregation, the plain average of the same 10 conformer representations, and their energy-weighted average are compared on how well a linear probe recovers (coefficient of determination) six metrics definable only over a conformer set (set-specific metrics) and five size-independent shape metrics. (d) For 12,401 active compounds, the fraction of cases where the closest partner in embedding space, from partners with 2D similarity < 0.20, is a compound active against the same protein. The dashed line represents the chance level.

Table 1: Group vs. group comparison. The top two rows are hit rates and the bottom row is the median ΔG. The left two columns are per-target medians over the 30 targets with at least 20 compounds evaluated in each group, and the third column is the number of those targets where ZAO is better. The rightmost column is the pooled value over the same 30 targets. The subset that both judge active contains 1,329 pairs.

Metric	ZAO group	Boltz-ESPSim group	Targets where ZAO is better	Subset supported by both ZAO and Boltz-2
$\Delta G \leq -10$ kcal/mol	26.4%	16.0%	22 of 30 targets	53.0%
$\Delta G \leq -12$ kcal/mol	14.6%	5.2%	25 of 30 targets	37.8%
Median ΔG	-7.58	-7.16	19 of 30 targets	-10.82

not compare ZAO with Boltz-2 under identical conditions (Appendix C.2). Due to differences in computational requirements, the populations searched are not the same across the groups (1.1 billion molecules for the ZAO group versus some 11,000-96,000 per target for the Boltz-ESPSim group). Breadth does not confer an advantage, and narrowing by ESPSim, which confines candidates to the neighborhood of the reference compounds, works, if anything, in favor of the Boltz-ESPSim group.

As a positive control on whether the scoring system responds in the expected direction, we take from a pool of about 200 million molecules the compounds whose 2D similarity to known active compounds lies between 0.6 and 0.99 (median 0.66). This gives an analog group of 4,226 pairs across 38 targets. When this group is evaluated under the same scoring system, the median per-target hit rate across the 27 targets in the paired comparison is 64.8% (Appendix B.4).

The hit thresholds were fixed before any per-group results were examined. Since the groups do not share a target composition and the hit-rate level also ranges from 2.3% to 77.2% across targets, we make all judgments as paired comparisons at the target level, and both the reported difference and the quantity tested are the median of the per-target differences.

3.3 Hit Rates and Novelty

The ZAO group outperforms the Boltz-ESPSim group (Figure 2, Table 1): in the paired comparison over 30 targets, the hit-rate ratio widens to 1.65× at $\Delta G \leq -10$ kcal/mol and 2.81× at $\Delta G \leq -12$ kcal/mol (Table 1). Taken per target, the median difference is +5.7 points, and a two-sided exact

Wilcoxon signed-rank test against the null hypothesis that the differences are symmetric about zero yields W = 97 and p = 0.0043.

The predictions of ZAO and Boltz-2 barely correlate over the same compounds (median rank correlation 0.047); so the subset that both methods judge active is worth examining on its own. Among the compounds ZAO selects from 1.1 billion molecules, the 1,329 novel-scaffold pairs where Boltz-2 also judges the compound to bind (binding probability ≥ 0.5) have a hit rate of 53.0% and a strong-hit rate of 37.8%.

In four target classes, the gap between the groups is large (Table 4, Appendix D.1). ZAO performs better on all targets in the protein-protein interaction, protease, G protein-coupled receptor (GPCR), and hydrolase classes, with median differences of +16.5 to +21.9 points; protein-protein interactions have long been regarded as hard to target with small molecules [4]. In contrast, for the 10 kinase targets the improvement is small (+1.7 points, ahead on 7 of 10 targets), and the only targets where ZAO is clearly behind the Boltz-ESPSim group are NR3C1, AR, SLC6A4, and ITK, three of which are nuclear receptors or transporters. Since Boltz-2 relies on predicted complex structures in candidate selection, it is increasingly disadvantaged as accurate structures become harder to obtain, while ZAO selects candidates by looking only at the ligand, and its selection does not depend on complex-structure prediction (Appendix B.2).

Measures that need no 2D fingerprint threshold also confirm that the compounds we select have novel scaffolds. At two levels, side-chain-stripped scaffolds [9] and ring systems, we count the fraction of compounds that match nothing in a target's training positives. In the ZAO group, the per-target median is 100.0% at the scaffold level and 81.9% at the ring-system level, where a compound counts if it contains a ring system that appears in none of the training positives. The scaffold figure stays at 100.0% for the 2,772 hit pairs and for the 1,626 strong-hit pairs. For examples of the compounds actually selected, and how close they are to known compounds, see Appendix D.4. In addition, proximity to known active compounds does not account for the advantage (median similarity: 0.170 for the ZAO group and 0.206 for the Boltz-ESPSim group; Appendix D.2).

4 Analysis of the Frozen Representation

Since the selection in Section 3 draws its hits from compounds with low 2D similarity to known active compounds, 2D fingerprint similarity is not enough to explain that result. This section shows, in turn, the information the representation contains (Section 4.1), where that information comes from (Section 4.2), and what it makes possible in selection (Section 4.3). To isolate the effect of the conformer set, we pretrain a pair of models with the same architecture and parameter count, one on one conformer per molecule and the other on 10 (Appendix G). Owing to computational constraints, both are trained on about a seventh of the data and are hereafter called the 1-conformer control and the 10-conformer control.

4.1 Geometry Beyond 2D Descriptors

First, to examine what information the features contain, we sort 36 quantities into five categories: conformer energy, shape, 3D distances between pharmacophores (hereafter, pharmacophore distance), quantities definable only over a conformer set (hereafter, set-specific metrics), and 2D properties. For 20,000 molecules, we use linear probes [3] to extract the quantities from the frozen representation and compare the conditions by the coefficient of determination. As controls we take 13 feature sets: three public pretrained models that use conformers, two language models that read SMILES, 2D descriptors and 2D fingerprints [63], and our two models for the controlled analysis; with ZAO, we compare 14 feature sets (listed in Table 2; see Appendix E.1 for probe design and control conditions).

In all three categories that require 3D, ZAO's representation scores highest, and the largest margin is on set-specific metrics. All models improve when 10 conformers are averaged, but ours, which places aggregation inside pretraining, is still ahead. What matters is less the averaging operation than whether it sits inside training. By contrast, the 2D descriptors recover 2D properties almost perfectly and shape hardly at all. 3D information is not an extension of 2D features and has to be sought separately, which explains why hits come from the low-2D-similarity region in Section 3.

Table 2: Coefficients of determination over 5 categories, 36 metrics, and 14 feature sets (19,974 molecules). All are computed from the same 10 conformers. Dim is each condition's number of input columns for the linear probe, and the maximum in each column is in bold. Each category column is the mean coefficient of determination over that category's metrics (metric list in Table 8), and the rightmost column is the mean over all 36 metrics re-measured after reducing four of the 10-conformer conditions and the 2D fingerprint to 512 dimensions with the same Gaussian random projection (projection conditions and preprocessing in Appendix E.1).

Feature set	Dim	Conformer energy 4 metrics	Shape 7 metrics	Pharmacophore distance 6 metrics	Set-specific metrics 6 metrics	2D properties 13 metrics	Mean over all 36 metrics	Mean after matching 512 dim
ZAO 10 conformers	2,048	**0.422**	**0.858**	**0.856**	**0.805**	0.939	**0.830**	**0.768**
Control pair (10 conformers)	2,048	0.416	0.843	0.818	0.768	0.911	0.803	—
Control pair (1 conformer)	2,048	0.242	0.615	0.706	0.428	0.856	0.644	—
Control pair (1 conformer), mean of 10 conformers	2,048	0.254	0.729	0.765	0.471	0.876	0.692	—
ConforFormer 1 conformer	512	0.163	0.506	0.672	0.348	0.550	0.485	—
ConforFormer mean of 10 conformers	512	0.202	0.660	0.768	0.447	0.635	0.582	0.579
Uni-Mol2 1 conformer	1,536	0.299	0.672	0.767	0.551	0.900	0.709	—
Uni-Mol2 mean of 10 conformers	1,536	0.313	0.774	0.814	0.586	0.911	0.748	0.688
Uni-Mol 1 conformer	512	0.243	0.626	0.722	0.463	0.772	0.625	—
Uni-Mol mean of 10 conformers	512	0.272	0.790	0.798	0.539	0.819	0.702	0.701
MoLFormer-XL	768	0.283	0.525	0.726	0.517	0.878	0.658	—
ChemBERTa-77M	384	0.264	0.481	0.683	0.464	0.840	0.617	—
2D descriptors	216	0.317	0.479	0.710	0.498	**0.995**	0.689	—
2D fingerprints	2,048	0.298	0.467	0.690	0.485	0.825	0.617	0.484

Only the four conformer-energy metrics remain low (0.422): the mean energy over 10 conformers is recovered at 0.727 and the entropy of the Boltzmann distribution at 0.750, while the standard deviation reaches only 0.106 and the range only 0.105. The representative value of the set survives the aggregation, while its spread does not. This is consistent with the aggregation's use of Boltzmann weighting (Appendix E.2).

When conformers are regenerated, a feature built from one conformer is not uniquely determined by the molecule, regardless of the encoder. The rate at which a molecule is retrieved from the original feature set falls short of 80% with one conformer and approaches 90% when ZAO reads the conformer set. Features that ignore conformers are perfect by definition. However, reproducibility is not the only thing to look at: plain averaging of single-conformer embeddings scores even higher on reproducibility, but since averaging erases the differences between conformers, it gives up the set-specific metrics. The 10-conformer gain comes from reducing occasional large failures, and it grows with molecular flexibility and size (downstream stability and sensitivity to subtle conformer differences in Appendix E.2).

4.2 The Source of the Gain and the Place of Aggregation

Here, we evaluate what the conformer set adds by comparing the 1-conformer control with the 10-conformer control.

The 10-conformer control is ahead in all categories, but not by the same amount (Figure 3a). The gain is largest on quantities that describe the spread between conformers, while the rise in quantities that only the 2D graph determines is small. The latter move because reading several conformers reduces estimation variance; the coordinates add nothing to what the 2D graph already fixes. A conformer set contributes the information in the coordinates, which 2D descriptors and fingerprints leave out, and that is why it is worth adding to existing 2D features.

The gap also widens as the molecule becomes more flexible (Figure 3b). Both arms lose accuracy as the number of rotatable bonds increases, but the single-conformer arm loses more, and the gap between them widens for the more flexible molecules. One conformer is enough to approximate an individual structure. It is not enough to capture the differences between conformers. Binding requires a compound to adopt the shape the target accepts, and how easily it reaches that shape depends on the range of shapes it can take. A single-conformer representation lacks that range; it can show whether a shape fits, but not whether the compound can get there. Drug discovery works

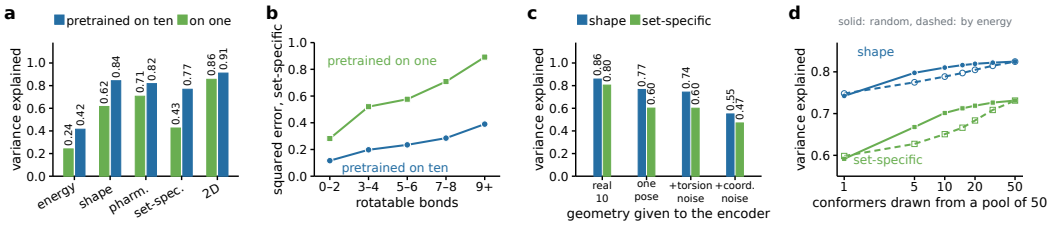

Figure 3: Panels (a) and (b) show results from the 1-conformer control and the 10-conformer control, and panels (c) and (d) show how the ZAO model behaves as the geometry and conformer count are varied. (a) Category-wise coefficients of determination. (b) Squared error on the six set-specific metrics with each metric's variance held fixed, split by rotatable-bond count. Lower is better. (c) Controls that vary the geometry given to the encoder: the real 10 conformers, the single most stable conformer, 10 conformers with noise added to the torsions, and 10 with noise added to the coordinates. (d) Two series varying the number of conformers drawn from a pool of 50, at random (solid) and in ascending energy order (dashed). The vertical axis starts at 0.55.

with large, flexible molecules, which is consistent with the selection in Section 3 paying off in that region.

In the controls that rebuild the geometry, we add noise to the torsion angles until the variation between conformers matches that of a real conformer set. Even so, what can be extracted falls short of what the real 10 conformers yield, and it falls short of what the single most stable conformer yields on its own. Adding noise to the coordinates lowers it further (Figure 3c). Only real, force-field-optimized conformers produce the gain.

As the conformer count grows, so does the information that can be extracted from the features, but the shape metrics flatten between 5 and 10 conformers (Figure 3d). On the other hand, the set-specific metrics go on rising to 50. The count is not all that matters: 10 conformers drawn at random match 30 drawn in ascending energy order. The 10 conformers ZAO uses sit where the random-draw series begins to flatten (Table 10, Appendix E.2).

Where the conformer set is aggregated also matters for performance (Figure 2c). We pass the same 10 conformers through the same encoder and change only how the set is aggregated. Learned aggregation, a transformer over the per-conformer representations, is then ahead of both plain averaging and energy-weighted averaging on all six set-specific metrics. Its advantage over plain averaging vanishes, however, for the five size-independent shape metrics (Appendix E.3). The asymmetry follows from the form of the averaging. A design that emits an independent scalar per conformer and then averages uniformly can express only permutation-invariant functions that decompose additively across conformers, so standard deviations and ranges across conformers are in principle out of reach. The advantage of learned aggregation ends at that boundary (see Appendix E.3 for the derivation and its scope).

Even in public encoders that we did not train, the same asymmetry appears. Plain averaging of per-conformer representations improves all three public encoders on all five categories (Table 2). Since the coordinate-independent 2D properties also increase (0.772 to 0.819 for Uni-Mol), part of that improvement is reduced estimation variance. Nevertheless, on set-specific metrics the best averaged result, Uni-Mol2, is still only 0.586 (Table 2). ConforFormer, pretrained on the assumption of conformer sets, is lowest of the three on set-specific metrics at 0.447.

4.3 Scaffold Jumps

Among pairs with low 2D fingerprint similarity, ZAO's representation still finds active compounds (Figure 2d). Taking the closest partner in embedding space from those with 2D similarity below 0.20 yields a compound active against the same protein in 30.3% of cases, 10 times the random rate of 2.9%. The best existing measure built from the same conformers is ultrafast shape recognition with CREDO atom types (USRCAT) [67], with its descriptors averaged over the 10 conformers; it reaches 15.2%, half the ZAO rate (Appendix E.4). These pairs share shape and pharmacophore arrangement, not scaffolds (see Appendix E.4 for molecule counts, control construction, and agreement values).

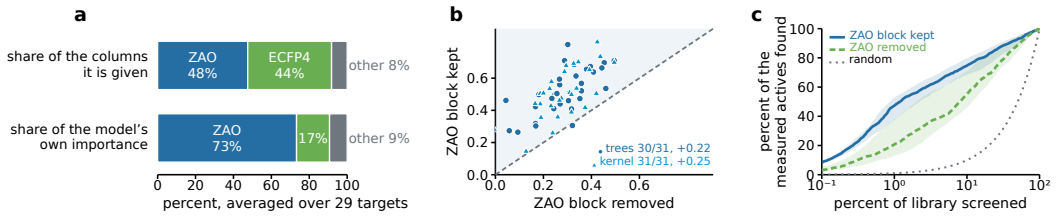

Figure 4: Contribution of the ZAO features. (a) Column composition of the features given to the selection model in Section 3 and the composition of feature importance (mean over the 29 targets whose selected configuration includes the ZAO block). The three thin blocks are merged into "other." (b) Rank correlation (Spearman) when retrained with and without these features on the same data, split, and hyperparameters. Each dot is one target-learner pair. Circles are gradient-boosted trees. Triangles are kernel ridge regression. (c) Fraction of experimentally active compounds recovered within the top k% of the library when screening with the same two settings. The band is the interquartile range over 31 targets.

The overlap between ZAO embeddings and 2D fingerprints is small, so ZAO supplements what 2D fingerprints already provide. Linear centered kernel alignment (CKA; Kornblith et al., 2019 [40]) is 0.948 between two settings of the same 2D fingerprint and 0.290 between ZAO embeddings and 2D fingerprints, and sweeping the fingerprint radius and bit count never takes it above 0.29 (Appendix E.4).

On the same 31 targets, we retrain with the configuration shared by all targets, removing only ZAO's 2,048 columns (Figure 4, Appendix D.3). Early enrichment (BEDROC20 [73]) falls by about 40% at the median for both gradient-boosted trees [60] and kernel ridge regression (Table 6). It is the top of the ranking that suffers: the number of experimentally active compounds recovered within the top 1% of the library falls to about a third (Figure 4c). Feature importance in the deployed model is dominated by ZAO too, well ahead of the extended-connectivity fingerprint (ECFP4), which is also 2,048-dimensional (Figure 4a).

5 Related Work

Single-conformer 3D encoders (SchNet [68], EGNN [65], Equiformer [43], MACE [7]) handle one geometry per forward pass, and 2D encoders pretrained on geometry (GraphMVP [44], 3D Infomax [70]) have no geometry at all during inference. GEM [23] still computes and inputs bond angles at inference, but that geometry comes from a single conformer. What sets us apart from prior work on conformer sets is where aggregation takes place. MARCEL [82] evaluates a strategy that aggregates per-conformer embeddings with a set encoder, but that aggregation is relearned under task-specific supervision. Uni-Mol [81] encodes each conformer and treats multiple conformers as fine-tuning data augmentation. ConforFormer [38] adds a contrastive loss to pretraining, taking different conformers of the same molecule as positive pairs. Whether conformer sets improve performance remains unsettled, as some report that a single conformer suffices [5] or that the properties that benefit are limited [15], while MARCEL [82] reports the opposite. Table 2 lines up the released encoders that can read the same 10 conformers at inference; GraphMVP [44] and 3D Infomax [70] fall outside that comparison because they have no geometry at inference, and the 1D MoLFormer [64] and ChemBERTa-2 [2] are added to show the level a representation attains without reading coordinates. Closest to ours in pretraining scale is Uni-Mol2 [34], at about 884 million conformers, and our 11 billion conformers are 12.4 times that. For the scale of the 1D and 2D corpora, how the denominator is defined, and the scope we check, see Appendix A.3.

Large-scale searching itself has many precedents, with populations spanning 12 million to more than 1 billion [28, 49, 72, 78]. Those studies all end with an assay of acquired or synthesized compounds for final evaluation, with 23-590 compounds tested against one or two protein targets or a single cell phenotype. Prospective evaluation by physics-based calculations has also been reported: more than 30,000 [18] and some 12,000 [42] compounds scored with absolute binding free energy, and 12 targets evaluated prospectively with relative free-energy calculations [66]. However, all of these reports measure their scale by the number of compounds, and among the reports gathered here the

number of targets in absolute binding free-energy studies is at most three. We compare candidate-selection methods under a fixed, existing scoring protocol and measure our scale by the number of targets (Appendix B.1).

6 Limitations and Conclusion

This paper has four limitations.

1. The scores in Section 3 are computed values; no compound was assayed (validity in Appendix B.4).
2. Enantiomers cannot be distinguished. Since one input dimension, the signed volume, changes sign under reflection, the impossibility result for E(3)-invariant features (Dumitrescu et al., 2025 [20]) does not apply, yet the trained model does not separate mirror images (Appendix A.1). That dimension is the only input that can respond to reflection, and zeroing it at inference leaves the embedding almost unchanged, which shows that the route to the representation is unobstructed and that training has not drawn on it. Of the pretraining objectives, only the chirality-aware functional-class fingerprint changes under reflection, and training has no reason to use this dimension. We leave the design of an objective function that discriminates chirality to future work.
3. Since retraining is difficult at this scale, and since practical use calls for supervision that covers as much as possible, we release and evaluate only a single set of weights. We use the public activity data in full as supervision, so we do not run leaderboard comparisons on public benchmarks (on the 22 ADMET leaderboards of the Therapeutics Data Commons, only 3 of the 10 top-ranked methods survive an audit in any case; Koleiev et al., 2026 [39]). Instead we build linear probes (Alain and Bengio, 2017 [3]), a measurement that does not go through a benchmark, and ask whether geometric quantities can be extracted linearly.
4. We restrict the inputs to a molecular weight of 100-1,000 Da, a single neutralized protonation state, and force-field-optimized conformers; nothing outside this range has been verified (for conformer quality, see [1]).

Whether a molecular representation is useful in drug discovery depends on whether it can actually pick out new compounds. We run inference over all 1.1 billion molecules, requiring 2D similarity to the training positives < 0.6 for each candidate. The novel-scaffold candidates selected this way span 31 targets and 9,220 compounds, a scale beyond the reach of synthesis and testing, and their hit rate exceeds that of the Boltz-ESPSim group (26.4% versus 16.0%, on 22 of 30 targets). From the frozen representation, a linear probe can predict the set-specific metrics with a coefficient of determination of up to 0.805, compared with 0.586 for the best public encoder.

Reproducibility and Release Statement

We release the weights of the 238-million-parameter ZAO model. We also provide scikit-learn-compatible implementations: ZAOTransformer, which extracts embeddings, and ZAORegressor, which implements the selection procedure of Section 3, along with runnable examples for both. Terms of use are stated separately. Of the selection model's training data, the part that comes from commercial activity databases cannot be reproduced, but even from public data alone, 50.9% of the pairs and all 140 targets remain (Appendix A.4). For Section 3, we put the accepted range and pooling rules in Appendix B.3, the selection model's feature blocks and each screening stage's threshold in Appendix C.1, and the per-group compound counts and the target list in Appendix D.3.

AI Use Statement

In conducting this research and writing this paper, we used large language model assistance within the following scope: (a) implementation of the model and analysis code and data wrangling; (b) drafting, revising, and translating this paper, and producing its figures and tables; and (c) advice on experimental design and manuscript structure, proposing and refining hypotheses, literature review and summarization, and drafting interpretations of the results. Large language models were not used

for synthetic data generation, theoretical frameworks, mathematical claims or proofs, or qualitative analysis, none of which apply to this study. Every number reported in this paper was checked against a result file or an execution log, and confirmed to be recomputable. The code was verified by the authors against execution results, translations were checked by the authors against the original text, and cited references were confirmed against the original sources. Only the authors decided the research questions, whether to run experiments, how to interpret results, and the scope of this paper's claims. The authors bear responsibility for the final content of this paper, including any text, claims, or artifacts produced with the aid of generative AI.

REFERENCES

[1] Keir Adams and Connor W. Coley. The impact of conformer quality on learned representations of molecular conformer ensembles. arXiv:2502.13220, 2025.

[2] Walid Ahmad, Elana Simon, Seyone Chithrananda, Gabriel Grand, and Bharath Ramsundar. ChemBERTa-2: towards chemical foundation models. In ELLIS Machine Learning for Molecule Discovery Workshop, 2021.

[3] Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. In International Conference on Learning Representations Workshop Track, 2017.

[4] Michelle R. Arkin, Yinyan Tang, and James A. Wells. Small-molecule inhibitors of protein-protein interactions: progressing toward the reality. Chemistry & Biology, 21(9):1102-1114, 2014.

[5] Simon Axelrod and Rafael Gómez-Bombarelli. Molecular machine learning with conformer ensembles. Machine Learning: Science and Technology, 4(3):035025, 2023.

[6] Pedro J. Ballester and W. Graham Richards. Ultrafast shape recognition to search compound databases for similar molecular shapes. Journal of Computational Chemistry, 28(10):1711-1723, 2007.

[7] Ilyes Batatia, Dávid Péter Kovács, Gregor N. C. Simm, Christoph Ortner, and Gábor Csányi. MACE: higher order equivariant message passing neural networks for fast and accurate force fields. In Advances in Neural Information Processing Systems 35, pp. 11423-11436, 2022.

[8] Matthias R. Bauer and Mark D. Mackey. Electrostatic complementarity as a fast and effective tool to optimize binding and selectivity of protein-ligand complexes. Journal of Medicinal Chemistry, 62(6):3036-3050, 2019.

[9] Guy W. Bemis and Mark A. Murcko. The properties of known drugs. 1. Molecular frameworks. Journal of Medicinal Chemistry, 39(15):2887-2893, 1996.

[10] Giovanni Bolcato, Esther Heid, and Jonas Boström. On the value of using 3D shape and electrostatic similarities in deep generative methods. Journal of Chemical Information and Modeling, 62(6):1388-1398, 2022.

[11] Simon Boothroyd, Pavan Kumar Behara, Owen C. Madin, David F. Hahn, Hyesu Jang, Vytautas Gapsys, Jeffrey R. Wagner, Joshua T. Horton, David L. Dotson, Matthew W. Thompson, Jessica Maat, Trevor Gokey, Lee-Ping Wang, Daniel J. Cole, Michael K. Gilson, John D. Chodera, Christopher I. Bayly, Michael R. Shirts, and David L. Mobley. Development and benchmarking of Open Force Field 2.0.0: the Sage small molecule force field. Journal of Chemical Theory and Computation, 19(11):3251-3275, 2023.

[12] Stefan Boresch, Franz Tettinger, Martin Leitgeb, and Martin Karplus. Absolute binding free energies: a quantitative approach for their calculation. The Journal of Physical Chemistry B, 107(35):9535-9551, 2003.

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

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

[15] Bryan Cheng, Austin Hanchi Jin, and Jasper Zhang. When does conformer geometry help? Selective complementarity of 3D ensemble statistics and 2D fingerprints. In Proceedings of the 17th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics (BCB '26), pp. 1-9, 2026.

[16] John D. Chodera and Michael R. Shirts. Replica exchange and expanded ensemble simulations as Gibbs sampling: simple improvements for enhanced mixing. The Journal of Chemical Physics, 135(19):194110, 2011.

[17] Jason C. Cole, Christopher W. Murray, J. Willem M. Nissink, Richard D. Taylor, and Robin Taylor. Comparing protein-ligand docking programs is difficult. Proteins: Structure, Function, and Bioinformatics, 60(3):325-332, 2005.

[18] Jordan E. Crivelli-Decker, Zane Beckwith, Gary Tom, Ly Le, Sheenam Khuttan, Romelia Salomon-Ferrer, Jackson Beall, Rafael Gómez-Bombarelli, and Andrea Bortolato. Machine learning guided AQFEP: a fast and efficient absolute free energy perturbation solution for virtual screening. Journal of Chemical Theory and Computation, 20(16):7188-7198, 2024.

[19] Jörg Degen, Christof Wegscheid-Gerlach, Andrea Zaliani, and Matthias Rarey. On the art of compiling and using 'drug-like' chemical fragment spaces. ChemMedChem, 3(10):1503-1507, 2008.

[20] Alexandru Dumitrescu, Dani Korpela, Markus Heinonen, Yogesh Verma, Valerii Iakovlev, Vikas Garg, and Harri Lähdesmäki. E(3)-equivariant models cannot learn chirality: field-based molecular generation. In International Conference on Learning Representations, 2025.

[21] Joseph L. Durant, Burton A. Leland, Douglas R. Henry, and James G. Nourse. Reoptimization of MDL keys for use in drug discovery. Journal of Chemical Information and Computer Sciences, 42(6):1273-1280, 2002.

[22] Peter Eastman, Raimondas Galvelis, Raúl P. Peláez, Charlles R. A. Abreu, Stephen E. Farr, Emilio Gallicchio, Anton Gorenko, Michael M. Henry, Frank Hu, Jing Huang, Andreas Krämer, Julien Michel, Joshua A. Mitchell, Vijay S. Pande, João P. G. L. M. Rodrigues, Jaime Rodriguez-Guerra, Andrew C. Simmonett, Sukrit Singh, Jason Swails, Philip Turner, Yuanqing Wang, Ivy Zhang, John D. Chodera, Gianni De Fabritiis, and Thomas E. Markland. OpenMM 8: molecular dynamics simulation with machine learning potentials. The Journal of Physical Chemistry B, 128(1):109-116, 2024.

[23] Xiaomin Fang, Lihang Liu, Jieqiong Lei, Donglong He, Shanzhuo Zhang, Jingbo Zhou, Fan Wang, Hua Wu, and Haifeng Wang. Geometry-enhanced molecular representation learning for property prediction. Nature Machine Intelligence, 4(2):127-134, 2022.

[24] Mudong Feng, Germano Heinzelmann, and Michael K. Gilson. Absolute binding free energy calculations improve enrichment of actives in virtual compound screening. Scientific Reports, 12:13640, 2022.

[25] Shikun Feng, Yuyan Ni, Yanyan Lan, Zhi-Ming Ma, and Wei-Ying Ma. Fractional denoising for 3D molecular pre-training. In Proceedings of the 40th International Conference on Machine Learning, PMLR 202:9938-9961, 2023.

[26] Michael K. Gilson, Tiqing Liu, Michael Baitaluk, George Nicola, Linda Hwang, and Jenny Chong. BindingDB in 2015: a public database for medicinal chemistry, computational chemistry and systems pharmacology. Nucleic Acids Research, 44(D1):D1045-D1053, 2016.

[27] Alberto Gobbi and Dieter Poppinger. Genetic optimization of combinatorial libraries. Biotechnology and Bioengineering, 61(1):47-54, 1998.

[28] Christoph Gorgulla, Andras Boeszoermenyi, Zi-Fu Wang, Patrick D. Fischer, Paul W. Coote, Krishna M. Padmanabha Das, Yehor S. Malets, Dmytro S. Radchenko, Yurii S. Moroz, David A. Scott, Konstantin Fackeldey, Moritz Hoffmann, Iryna Iavniuk, Gerhard Wagner, and Haribabu Arthanari. An open-source drug discovery platform enables ultra-large virtual screens. Nature, 580(7805):663-668, 2020.

[29] Oleksandr O. Grygorenko, Dmytro S. Radchenko, Igor Dziuba, Alexander Chuprina, Kateryna E. Gubina, and Yurii S. Moroz. Generating multibillion chemical space of readily accessible screening compounds. iScience, 23(11):101681, 2020.

[30] Thomas A. Halgren. Merck molecular force field. I. Basis, form, scope, parameterization, and performance of MMFF94. Journal of Computational Chemistry, 17(5-6):490-519, 1996.

[31] Paul C. D. Hawkins, A. Geoffrey Skillman, and Anthony Nicholls. Comparison of shape-matching and docking as virtual screening tools. Journal of Medicinal Chemistry, 50(1):74-82, 2007.

[32] Paul C. D. Hawkins. Conformation generation: the state of the art. Journal of Chemical Information and Modeling, 57(8):1747-1756, 2017.

[33] Araz Jakalian, David B. Jack, and Christopher I. Bayly. Fast, efficient generation of high-quality atomic charges. AM1-BCC model: II. Parameterization and validation. Journal of Computational Chemistry, 23(16):1623-1641, 2002.

[34] Xiaohong Ji, Zhen Wang, Zhifeng Gao, Hang Zheng, Linfeng Zhang, Guolin Ke, and Weinan E. Exploring molecular pretraining model at scale. In Advances in Neural Information Processing Systems 37, pp. 46956-46978, 2024.

[35] Bowen Jing, Gabriele Corso, Jeffrey Chang, Regina Barzilay, and Tommi Jaakkola. Torsional diffusion for molecular conformer generation. In Advances in Neural Information Processing Systems 35, pp. 24240-24253, 2022.

[36] William L. Jorgensen, Jayaraman Chandrasekhar, Jeffry D. Madura, Roger W. Impey, and Michael L. Klein. Comparison of simple potential functions for simulating liquid water. The Journal of Chemical Physics, 79(2):926-935, 1983.

[37] Kerstin Kläser, Błażej Banaszewski, Samuel Maddrell-Mander, Callum McLean, Luis Müller, Ali Parviz, Shenyang Huang, and Andrew Fitzgibbon. MiniMol: a parameter-efficient foundation model for molecular learning. In International Conference on Machine Learning Workshop on Accessible and Efficient Foundation Models for Biological Discovery, 2024.

[38] Mas Pieter Klein, Irina Rudenko, Evgeny A. Pidko, and Ivan Bushmarinov. ConforFormer: representation for molecules through understanding of conformers. Digital Discovery, 5(6):2639-2649, 2026.

[39] Ihor Koleiev, Roman Stratiichuk, Nazar Shevchuk, Mykola Melnychenko, Alex Nyporko, Daniil Todoryshyn, Vladyslav Husak, Sergii Starosyla, Semen Yesylevskyy, and Alan Nafiiev. An end-user audit of reproducibility, data leakage, and overfitting of the top-ranked ADMET prediction models in TDC leaderboards. Journal of Chemical Information and Modeling, 66(14):8045-8058, 2026.

[40] Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In Proceedings of the 36th International Conference on Machine Learning, PMLR 97:3519-3529, 2019.

[41] Gregory A. Landrum and Sereina Riniker. Combining IC50 or Ki values from different sources is a source of significant noise. Journal of Chemical Information and Modeling, 64(5):1560-1567, 2024.

[42] Zhe Li, Chengkun Wu, Yishui Li, Runduo Liu, Kai Lu, Ruibo Wang, Jie Liu, Chunye Gong, Canqun Yang, Xin Wang, Chang-Guo Zhan, and Hai-Bin Luo. Free energy perturbation-based large-scale virtual screening for effective drug discovery against COVID-19. The International Journal of High Performance Computing Applications, 37(1):45-57, 2023.

[43] Yi-Lun Liao and Tess Smidt. Equiformer: equivariant graph attention transformer for 3D atomistic graphs. In International Conference on Learning Representations, 2023.

[44] Shengchao Liu, Hanchen Wang, Weiyang Liu, Joan Lasenby, Hongyu Guo, and Jian Tang. Pre-training molecular graph representation with 3D geometry. In International Conference on Learning Representations, 2022.

[45] Ilya Loshchilov and Frank Hutter. SGDR: stochastic gradient descent with warm restarts. In International Conference on Learning Representations, 2017.

[46] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In International Conference on Learning Representations, 2019.

[47] Shuqi Lu, Zhifeng Gao, Di He, Linfeng Zhang, and Guolin Ke. Data-driven quantum chemical property prediction leveraging 3D conformations with Uni-Mol+. Nature Communications, 15:7104, 2024.

[48] Shengjie Luo, Tianlang Chen, Yixian Xu, Shuxin Zheng, Tie-Yan Liu, Liwei Wang, and Di He. One Transformer can understand both 2D & 3D molecular data. In International Conference on Learning Representations, 2023.

[49] Jiankun Lyu, Sheng Wang, Trent E. Balius, Isha Singh, Anat Levit, Yurii S. Moroz, Matthew J. O'Meara, Tao Che, Enkhjargal Algaa, Kateryna Tolmachova, Andrey A. Tolmachev, Brian K. Shoichet, Bryan L. Roth, and John J. Irwin. Ultra-large library docking for discovering new chemotypes. Nature, 566(7743):224-229, 2019.

[50] Jiankun Lyu, John J. Irwin, and Brian K. Shoichet. Modeling the expansion of virtual screening libraries. Nature Chemical Biology, 19(6):712-718, 2023.

[51] James A. Maier, Carmenza Martinez, Koushik Kasavajhala, Lauren Wickstrom, Kevin E. Hauser, and Carlos Simmerling. ff14SB: improving the accuracy of protein side chain and backbone parameters from ff99SB. Journal of Chemical Theory and Computation, 11(8):3696-3713, 2015.

[52] Yvonne Connolly Martin. Let's not forget tautomers. Journal of Computer-Aided Molecular Design, 23(10):693-704, 2009.

[53] David L. Mobley and Michael K. Gilson. Predicting binding free energies: frontiers and benchmarks. Annual Review of Biophysics, 46:531-558, 2017.

[54] Anthony Nicholls, Georgia B. McGaughey, Robert P. Sheridan, Andrew C. Good, Gregory Warren, Magali Mathieu, Steven W. Muchmore, Scott P. Brown, J. Andrew Grant, James A. Haigh, Neysa Nevins, Ajay N. Jain, and Brian Kelley. Molecular shape and medicinal chemistry: a perspective. Journal of Medicinal Chemistry, 53(10):3862-3886, 2010.

[55] Ramaswamy Nilakantan, Norman Bauman, J. Scott Dixon, and R. Venkataraghavan. Topological torsion: a new molecular descriptor for SAR applications. Comparison with other descriptors. Journal of Chemical Information and Computer Sciences, 27(2):82-85, 1987.

[56] NVIDIA. nvMolKit: a high-performance, GPU-accelerated library for key computational chemistry tasks, such as molecular similarity, conformer generation, and geometry relaxation. https://github.com/NVIDIA-BioNeMo/nvMolKit, 2025.

[57] Saro Passaro, Gabriele Corso, Jeremy Wohlwend, Mateo Reveiz, Stephan Thaler, Vignesh Ram Somnath, Noah Getz, Tally Portnoi, Julien Roy, Hannes Stark, David Kwabi-Addo, Dominique Beaini, Tommi Jaakkola, and Regina Barzilay. Boltz-2: towards accurate and efficient binding affinity prediction. bioRxiv, 2025. doi:10.1101/2025.06.14.659707.

[58] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: machine learning in Python. Journal of Machine Learning Research, 12(85):2825-2830, 2011.

[59] Emanuele Perola and Paul S. Charifson. Conformational analysis of drug-like molecules bound to proteins: an extensive study of ligand reorganization upon binding. Journal of Medicinal Chemistry, 47(10):2499-2510, 2004.

[60] Liudmila Prokhorenkova, Gleb Gusev, Aleksandr Vorobev, Anna Veronika Dorogush, and Andrey Gulin. CatBoost: unbiased boosting with categorical features. In Advances in Neural Information Processing Systems 31, 2018.

[61] RDKit: Open-source cheminformatics, version 2025.03.1. https://www.rdkit.org, 2025.

[62] Sereina Riniker and Gregory A. Landrum. Better informed distance geometry: using what we know to improve conformation generation. Journal of Chemical Information and Modeling, 55(12):2562-2574, 2015.

[63] David Rogers and Mathew Hahn. Extended-connectivity fingerprints. Journal of Chemical Information and Modeling, 50(5):742-754, 2010.

[64] Jerret Ross, Brian Belgodere, Vijil Chenthamarakshan, Inkit Padhi, Youssef Mroueh, and Payel Das. Large-scale chemical language representations capture molecular structure and properties. Nature Machine Intelligence, 4(12):1256-1264, 2022.

[65] Víctor Garcia Satorras, Emiel Hoogeboom, and Max Welling. E(n) equivariant graph neural networks. In Proceedings of the 38th International Conference on Machine Learning, PMLR 139:9323-9332, 2021.

[66] Christina E. M. Schindler, Hannah Baumann, Andreas Blum, Dietrich Böse, Hans-Peter Buchstaller, Lars Burgdorf, Daniel Cappel, Eugene Chekler, Paul Czodrowski, Dieter Dorsch, Merveille K. I. Eguida, Bruce Follows, Thomas Fuchß, Ulrich Grädler, Jakub Gunera, Theresa Johnson, Catherine Jorand Lebrun, Srinivasa Karra, Markus Klein, Tim Knehans, Lisa Koetzner, Mireille Krier, Matthias Leiendecker, Birgitta Leuthner, Liwei Li, Igor Mochalkin, Djordje Musil, Constantin Neagu, Friedrich Rippmann, Kai Schiemann, Robert Schulz, Thomas Steinbrecher, Eva-Maria Tanzer, Andrea Unzue Lopez, Ariele Viacava Follis, Ansgar Wegener, and Daniel Kuhn. Large-scale assessment of binding free energy calculations in active drug discovery projects. Journal of Chemical Information and Modeling, 60(11):5457-5474, 2020.

[67] Adrian M. Schreyer and Tom Blundell. USRCAT: real-time ultrafast shape recognition with pharmacophoric constraints. Journal of Cheminformatics, 4:27, 2012.

[68] Kristof T. Schütt, Pieter-Jan Kindermans, Huziel E. Sauceda, Stefan Chmiela, Alexandre Tkatchenko, and Klaus-Robert Müller. SchNet: a continuous-filter convolutional neural network for modeling quantum interactions. In Advances in Neural Information Processing Systems 30, pp. 992-1002, 2017.

[69] Michael R. Shirts and John D. Chodera. Statistically optimal analysis of samples from multiple equilibrium states. The Journal of Chemical Physics, 129(12):124105, 2008.

[70] Hannes Stärk, Dominique Beaini, Gabriele Corso, Prudencio Tossou, Christian Dallago, Stephan Günnemann, and Pietro Liò. 3D Infomax improves GNNs for molecular property prediction. In Proceedings of the 39th International Conference on Machine Learning, PMLR 162:20479-20502, 2022.

[71] Nikolaus Stiefl, Ian A. Watson, Knut Baumann, and Andrea Zaliani. ErG: 2D pharmacophore descriptions for scaffold hopping. Journal of Chemical Information and Modeling, 46(1):208-220, 2006.

[72] Jonathan M. Stokes, Kevin Yang, Kyle Swanson, Wengong Jin, Andres Cubillos-Ruiz, Nina M. Donghia, Craig R. MacNair, Shawn French, Lindsey A. Carfrae, Zohar Bloom-Ackermann, Victoria M. Tran, Anush Chiappino-Pepe, Ahmed H. Badran, Ian W. Andrews, Emma J. Chory, George M. Church, Eric D. Brown, Tommi S. Jaakkola, Regina Barzilay, and James J. Collins. A deep learning approach to antibiotic discovery. Cell, 180(4):688-702.e13, 2020.

[73] Jean-François Truchon and Christopher I. Bayly. Evaluating virtual screening methods: good and bad metrics for the 'early recognition' problem. Journal of Chemical Information and Modeling, 47(2):488-508, 2007.

[74] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Advances in Neural Information Processing Systems 30, pp. 5998-6008, 2017.

[75] Shuzhe Wang, Jagna Witek, Gregory A. Landrum, and Sereina Riniker. Improving conformer generation for small rings and macrocycles based on distance geometry and experimental torsional-angle preferences. Journal of Chemical Information and Modeling, 60(4):2044-2058, 2020.

[76] Christopher K. I. Williams and Matthias Seeger. Using the Nyström method to speed up kernel machines. In Advances in Neural Information Processing Systems 13, pp. 682-688, 2001.

[77] Jeremy Wohlwend, Gabriele Corso, Saro Passaro, Noah Getz, Mateo Reveiz, Ken Leidal, Wojtek Swiderski, Liam Atkinson, Tally Portnoi, Itamar Chinn, Jacob Silterra, Tommi Jaakkola, and Regina Barzilay. Boltz-1: democratizing biomolecular interaction modeling. bioRxiv, 2025 (version posted 6 May 2025). doi:10.1101/2024.11.19.624167.

[78] Felix Wong, Erica J. Zheng, Jacqueline A. Valeri, Nina M. Donghia, Melis N. Anahtar, Satotaka Omori, Alicia Li, Andres Cubillos-Ruiz, Aarti Krishnan, Wengong Jin, Abigail L. Manson, Jens Friedrichs, Ralf Helbig, Behnoush Hajian, Dawid K. Fiejtek, Florence F. Wagner, Holly H. Soutter, Ashlee M. Earl, Jonathan M. Stokes, Lars D. Renner, and James J. Collins. Discovery of a structural class of antibiotics with explainable deep learning. Nature, 626(7997):177-185, 2024.

[79] Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? In International Conference on Learning Representations, 2019.

[80] Barbara Zdrazil, Eloy Felix, Fiona Hunter, Emma J. Manners, James Blackshaw, Sybilla Corbett, Marleen de Veij, Harris Ioannidis, David Mendez Lopez, Juan F. Mosquera, Maria Paula Magarinos, Nicolas Bosc, Ricardo Arcila, Tevfik Kizilören, Anna Gaulton, A. Patrícia Bento, Melissa F. Adasme, Peter Monecke, Gregory A. Landrum, and Andrew R. Leach. The ChEMBL Database in 2023: a drug discovery platform spanning multiple bioactivity data types and time periods. Nucleic Acids Research, 52(D1):D1180-D1192, 2024.

[81] Gengmo Zhou, Zhifeng Gao, Qiankun Ding, Hang Zheng, Hongteng Xu, Zhewei Wei, Linfeng Zhang, and Guolin Ke. Uni-Mol: a universal 3D molecular representation learning framework. In International Conference on Learning Representations, 2023.

[82] Yanqiao Zhu, Jeehyun Hwang, Keir Adams, Zhen Liu, Bozhao Nan, Brock Anton Stenfors, Yuanqi Du, Jatin Chauhan, Olaf Wiest, Olexandr Isayev, Connor W. Coley, Yizhou Sun, and Wei Wang. Learning over molecular conformer ensembles: datasets and benchmarks. In International Conference on Learning Representations, 2024.

A Model, Data, and Training

This appendix focuses on the information needed to reproduce the released weights. Only Appendix A.4 departs from the foundation model's configuration, covering instead the activity labels used by the selection model in Section 3 and the downstream learners we recommend to readers.

A.1 Inputs: Standardization, Conformers, and Features

Standardization. We remove salts and solvents, neutralize charges, and restrict molecules to a molecular weight of 100-1,000 Da. Additionally, we remove molecules with more than 100 atoms after adding hydrogens, and for protonation states and tautomers [52] we take a single state.

Conformer generation. We generate 10 conformers per molecule with ETKDGv3 [62, 75] and optimize each with MMFF94 [30]. We sample 2,000 molecules and measure all 45 pairwise heavy-atom RMSDs across the 10 conformers of each molecule, with symmetry correction; the mean per molecule is 1.81 Å. Generation uses the GPU implementation of nvMolKit [56] (throughput of the GPU and CPU implementations is compared in Appendix F.2).

Atom features. Each atom is described by seven features (one dimension each): atomic number, MMFF94 partial charge, in-ring, aromatic, degree, hydrogen count, and hybridization. Partial charges are computed per molecule from the MMFF94 force field and set to 0 when unavailable. The bond connectivity fixes all seven features, and none of them changes with the conformer. Each edge has a five-dimensional one-hot encoding of bond type (none, single, double, triple, aromatic). As input, the 2D path receives only the same atom features and the bond graph, never the coordinates.

Geometric features. Of the nine dimensions recomputed from coordinates each time, one is a signed volume over the three nearest atoms; four are the mean, standard deviation, minimum, and maximum of the six angles formed by the four nearest atoms; and four are the mean, standard deviation, minimum, and maximum of the dihedral angles that those same four atoms form. These features have no learnable parameters and are invariant under SE(3) transformations of the coordinates. Only the signed-volume dimension flips sign under reflection, which is why they are not E(3)-invariant.

Chirality. For the 1,953 molecules with a stereocenter, the cosine distance between a molecule's embedding and that of its mirror image, obtained by inverting one coordinate axis of the conformer, has a median of 0.0002, and no molecule goes above the conformer-regeneration noise level of 0.041 (the mean over 146 molecules measured separately). For the 498 molecules without a stereocenter, the figure is 0.0001, which is no different. Only the signed-volume dimension can respond to reflection, and zeroing it at inference removes the response exactly. Zeroing it moves the embedding by only 0.0047, which is 0.8% of the 0.605 that separates unrelated molecules. The information is in the input and the route to the representation is unobstructed, but the model barely uses it.

A.2 Model, Objectives, and Optimization

Model dimensions and training settings. The 3D path's E(3)-equivariant graph network has a hidden dimension of 256 and 3 layers, with edges over the 16 nearest atoms, and the 2D path's graph network also has 3 layers. There are 64 learnable queries, and the substructure embeddings have 256 dimensions; the Conformer Pooling transformer has 3 layers, 8 attention heads, and an intermediate dimension of 512, and distances are encoded by a 16-function Gaussian basis (maximum 20 Å) and added to the attention logits as a bias. Per-conformer representations are 1,024-dimensional; the aggregation transformer has 6 layers and 8 attention heads with no positional encoding, and the temperature of the Boltzmann weights built from relative energies is 1.0 kcal/mol; the molecular representation is the 2,048-dimensional CLS output, and dropout is 0.1 for encoders and 0.3 for auxiliary heads. For training regularization, Gaussian noise with a standard deviation of 0.02 Å is applied to the coordinates of the supervised-data molecules and of the BRICS-recombined compounds, and label smoothing with a standard deviation of 0.05 is applied to the multi-task prediction targets.

Optimization. The optimizer is AdamW [46]; the learning rate peaks at 1e-4 after a linear warmup from 1e-6 over 1,000 steps, then decays by cosine annealing (with warm restarts [45], period 2,000 steps, floor 1e-7). Gradient norms are clipped at 1.0; weight decay is 0.01 for the multi-task prediction head and 0.0001 elsewhere. The batch size is 96 per GPU and the total step count is 234,000.

Learning objectives. The auxiliary heads at the conformer stage predict relative energy, three fingerprints (Morgan radius 4 / 128 bits, functional-class fingerprint radius 4 / 128 bits, topological pharmacophore fingerprint 512 bits [27]), and the 64×64 distance matrix. The auxiliary heads after aggregation predict the same three fingerprints plus the Boltzmann-weighted mean and variance of the same distance matrix. Losses are squared error, except binary cross-entropy for the fingerprints. In all, the loss comprises 12 terms: per conformer, relative energy, the 64×64 distance matrix, and three fingerprints; per molecule, three fingerprints, the Boltzmann-weighted mean and variance of the same distance matrix, multi-task prediction (counted as two terms, one for dense tasks and one for sparse tasks), and the similarity term. For the similarity term, the correlation is taken between the cosine similarity in the embedding space and a target value, where the target is a weighted

average of the Tanimoto coefficients of the three fingerprints and the shape similarity of ultrafast shape recognition (USR) [6]. The USR shape similarity is built from a 24-dimensional vector, the mean and the standard deviation of the 12-dimensional USR descriptor computed per conformer, with the L2 distance turned into `1/(1+d)`. The weights in that average are 1.0 for each of the three fingerprints and 2.0 for USR. Multi-task prediction over 652 public assay tasks is added to these objectives. All heads are linear maps; only the multi-task prediction head has one hidden layer.

Loss weighting. The s_k in the main text is an exponential moving average of the loss magnitude, with a decay of 0.999 and a plain running mean for the first 10 steps. It is not applied to the six fingerprint terms, for which $s_k = 1$, because these terms would create a positive feedback loop: a larger value increases the denominator and reduces the gradient. We determine the w_k values at validation every 1,000 steps. For the performance p_k of each term (coefficient of determination for regression, correlation for similarity, and F1 rescaled to 0 to 1 for fingerprints), $1 - p_k$ is taken, clamped to 0.1 to 3.0, normalized so that the mean over the 12 terms is 1, clamped again to 0.3 to 3.0, and smoothed with the previous value at a ratio of 0.6 to 0.4, so the worse a term performs, the larger its weight. These values are multiplied by fixed task-specific factors: 1.5 for the sparse tasks, 1.2 for the two distance terms and the similarity term, and 1.0 for the rest, followed by normalization to a mean of 1. Neither s_k nor w_k passes a gradient.

Conformer subsampling. For each batch we draw a count uniformly from 5 to 10 and take that many of the 10 conformers at random, using the same draw for every molecule in the batch, so a molecule is not always seen as the same fixed set. The mean of that draw is the 7.5 conformers per molecule quoted in Section 2.

Training stabilization. Any term exceeding the mean loss over the preceding 1,000 steps plus three standard deviations is clamped to that value before entering the total loss, and clamped values are excluded from the history. Any batch with a NaN or infinite term is discarded on all nodes.

Targets for the distance matrices. The targets of the 64×64 distance predictions per conformer and per molecule are the model's own distances between substructures, detached from the gradient; they are not labels supplied externally.

A.3 Pretraining Corpus and Scale Comparison

Corpus composition. The pretraining corpus is 1.1 billion molecules and 11 billion conformers, consisting of 994.71 million molecules from the Enamine REAL enumerated catalog [29] as we retrieved it, a catalog that carries no measured labels, together with 100.57 million molecules recombined by BRICS [19] decomposition of the public databases that supply the supervised data; hereafter, we call the former the enumerated catalog. In addition, 1.3 million molecules and 652 tasks derived from public supervised data (ChEMBL [80], BindingDB [26]) are mixed in as supervision for multi-task prediction. These 1.3 million molecules are not part of the 1.1 billion molecules and 11 billion conformers.

Mixing. For every four enumerated-catalog entries, we interleave one supervised-data entry and one BRICS-recombined compound in alternation. For sparse multi-task prediction tasks, we use the output of a separately trained activity predictor (gradient-boosted quantile regression) as pseudo-labels to fill missing values. Pseudo-labels are given only to rows with a confidence above 0.95, derived from the prediction interval. They apply to supervised data and BRICS-recombined compounds; the enumerated catalog never receives them. Molecules used by the probes in Section 4 come from the test split taken from this supervised portion.

Denominator for the scale comparison. The Uni-Mol2 [34] dataset comprises about 884 million entries; that paper describes them as either 3D conformers or molecular structures and does not state a conformer count per molecule. Training uses 838 million conformers sampled from that set, and to compare corpus scale we take 884 million as the denominator and report 12.4×. Using the 838 million conformers actually used for training would give 13.1×, but then the numerator would still be the pre-sampling corpus and the comparison would not be like-for-like. To support our claim of the largest publicly reported number, we check Uni-Mol+ [47], Frad [25], Transformer-M [48], GraphMVP [44], 3D Infomax [70], MoLFormer [64], ChemBERTa-2 [2], MiniMol [37], and related work from 2025 to 2026 and find no conformer count exceeding that of Uni-Mol2; the next largest is Uni-Mol [81] at 209 million conformers.

A.4 Labels and Learners for the Selection Model

Sources of activity labels. For the selection model in Section 3, the activity labels come from ChEMBL [80], BindingDB [26], and one commercial activity database that requires a subscription. Of the 1,474,453 pairs, 724,157 come only from that commercial source. Even rebuilt from public data only, the set has 750,296 pairs and 522,564 molecules, retaining 50.9% of the pairs, 51.6% of the rows used for training, and all 140 targets.

Measurement types and consolidation of duplicate measurements. Only three measurement types are accepted (IC50, Ki, and Kd); we exclude EC50, which we judge too noisy. Only measurements reported in nanomolar units are taken and converted to `-log10(M)`. Where ChEMBL provides a curated pchembl value, that value is used instead. Measurements with inequality signs are kept with their sign marked, but the regression uses no censoring loss: the numerical value accompanying the inequality is used directly as the target. Molecules are identified by SMILES passed through the same standardization procedure that ZAO uses. For consolidating duplicate measurements, rows for each target-molecule pair are ranked into four tiers (identical protein without inequality, identical protein with inequality, homologous protein without inequality, homologous protein with inequality), and the label is the median over the topmost tier's rows only. Homologous-protein expansion requires at least 40% identity and at least 50% coverage to enter training. Identity of 70% or above is called the core, and 40% to 70% the relaxed set; relaxed entries are used only for targets where the core, on its own, yields < 30 active compounds. Training sample weights are 1.0 for exact matches and the square of identity otherwise (floor 0.1). Rows used in training: 1,092,230 pairs; a positive is defined as `-log10(M)` ≥ 6.0 (1 μM), which covers 690,038 pairs.

Level of label noise. Under minimal curation, even for the same target and compound, values measured in different assays disagree by 0.3 log or more in about 65% of cases and by 1 log or more in 27% of cases (Landrum and Riniker, 2024 [41]), and we treat that as the resolution limit of any downstream metric.

Downstream learners. As the recommended default configuration, we use three families of learner. For gradient-boosted trees, we use CatBoost [60] with 1,000 iterations and random strength 2, averaged over five random seeds. On scaled features, we use scikit-learn [58] for an RBF-kernel support vector machine (SVR with C=1, gamma=scale, epsilon=0.1 for regression; SVC with C=1, gamma=scale for classification) and L2-regularized linear models (Ridge with alpha=1 for regression, logistic regression with C=1 for classification). We run no search for any of the three families; all settings not specified here are left at their defaults. For the selection model in Section 3, the learners are separate, and Appendix C.1 describes them.

Settings by stage. Three kinds of configuration appear in this paper. In the deployed configuration, a per-target search runs over feature blocks and hyperparameters; this is the configuration used for selection in Section 3 and shown in Figure 4a (Appendix C.1). The common configuration does not run a per-target search and applies the same settings to all targets, but the feature block still changes with the learner (Appendix C.1); the removal experiments in Appendix D.3 and Figures 4b and 4c use it. In the recommended default configuration, the learners are fixed to the values provided here and the feature blocks to all six defined in Appendix C.1.

B Binding Free Energy Scoring

This appendix describes Section 3's scoring system proper.

B.1 Scope and Choice of Protocol

Scope of the comparison. The scoring system is a fixed, existing procedure; we compare only the selection step. Before we saw a single result from this campaign's binding free-energy calculations, the candidates, the selection model, and the screening thresholds were all fixed. Only the narrowing of targets is based on prior calculations involving known active compounds.

We compare both selection pipelines as they were actually run, not a controlled experiment placing ZAO and Boltz-2 under identical conditions (Appendix C.2).

Target selection. The campaign is built by narrowing 121 targets to 40 on the criterion of Pearson correlation > 0.4 and Spearman rank correlation > 0.4 (all rank correlations in this paper are Spearman correlations). Of those, we report on the 31 targets with ZAO group calculations.

Choice of absolute binding free-energy calculations. Since we score compounds that share no scaffold with known active compounds, a relative free-energy calculation cannot be constructed; that method presupposes a series with reference compounds. It is relative free energy, which computes differences between closely related compounds, that the field has come to trust (Mobley and Gilson, 2017 [53]). By contrast, absolute calculations converge poorly and are computationally expensive, the more so when they are applied across structurally diverse compounds (Crivelli-Decker et al., 2024 [18]). On the absolute side, there is a prior report that used absolute binding free energy for screening and obtained enrichment surpassing that of docking (Feng et al., 2022 [24]). Indeed, the Boltz-2 paper itself, describing the pipeline we compare against, closes its exploration with an absolute binding free-energy calculation (TYK2; Passaro et al., 2025 [57]).

B.2 Protocol and Pose Generation

Calculation protocol. Absolute binding free energies are computed with the absolute binding protocol of OpenFE 1.12 (OpenMM 8.4 [22]): the ligand is annihilated in stages along two legs, in solvent and in complex, with sampling by replica exchange [16] and estimation by the multistate Bennett acceptance ratio [69]. The force fields are ff14SB [51] for protein, TIP3P [36] for water, and OpenFF 2.2.1 [11] for ligands, with partial charges from NAGL (openff-gnn-am1bcc-1.0.0 [33]), and ions are Na+ and Cl- at 0.15 M. Ligand orientation is held by Boresch-style [12] restraints; the radial force constant (in kJ/mol/nm^2) is 2,000 for neutral ligands and 418 for charged ligands. The complex side uses 30 λ windows; the solvent side uses 14 for neutral ligands and 24 for charged ligands. A 100,000-step minimization precedes equilibration; production sampling runs 2 ns per leg with three independent replicates per complex. The calculations run on RTX 6000 Pro GPUs, taking about 72 GPU-hours per complex, and these settings are identical across all groups.

Pose generation. For all groups, protein and ligand are co-folded with Boltz-2, producing one pose per compound. In addition, protonation states and tautomers are decided the same way across groups. Errors arising from pose generation are already reflected in the correlations reported in Appendix B.4, which are measured by running the entire pipeline from pose generation to scoring.

Since poses produced by co-folding models, including Boltz-1 [77] and Chai-1 [14], are not necessarily physically valid, we inspect the inputs themselves for physical validity with PoseBusters [13] (22 binary checks in dock mode; a target-stratified random sample of 1,153 complexes). The fraction of complexes passing all checks is 99.6% for the ZAO group, 99.7% for the Boltz-ESPSim group, 99.6% for the analog group, and 98.7% for known active compounds. However, PoseBusters looks at the physical validity of a structure and not at whether a pose is the correct binding mode. In particular, nothing here guarantees that co-folding succeeds equally on novel scaffolds and on analogs of known active compounds.

B.3 Acceptance, Pooling, and Pairing

Accepted range. The range admitted to the ΔG pooling is -20 to 0 kcal/mol; calculations falling outside it are excluded. Out-of-range fractions: ZAO group 9.2%, Boltz-ESPSim group 4.6%, analog group 13.4%, known active compounds 16.6%. Most out-of-range values are above the upper limit in the ZAO group and below the lower limit in the analog group and among known active compounds. As a result, truncation at the lower limit removes compounds computed to bind strongly, which leaves the hit rates of the analog group and of the known active compounds underestimated.

Pooling. Since each group has one pose per compound, a compound's value is the mean over its in-range replicates. When a compound has no in-range replicate, it is marked failed: it remains in the hit-rate denominator but is dropped from the ΔG median (failure rates: ZAO group 3.6%, Boltz-ESPSim group 1.7%, analog group 8.0%). The hit thresholds $\Delta G \leq -10$ kcal/mol and ≤ -12 kcal/mol are set with reference to the distribution of known active compounds. When known active compounds are passed through the same scoring system, 57.9% of their 1,419 pairs over the 30 paired-comparison targets meet the first threshold and 45.7% the second (per-target medians 59.2% and 41.8%).

Table 3: Why the target count changes from measurement to measurement, with a column listing which targets are removed from the 31.

Targets	Which measurement	Targets dropped from the 31	Reason for narrowing
31	Campaign scope (Table 7)	—	Targets for which the ZAO group has results from binding free-energy calculations
30	Group-vs-group paired comparison and both-supported subset (Tables 1 and 4)	EPHA2	Only pairs where both groups compute at least 20 compounds on the same target; the ZAO group has just nine compounds on EPHA2
30	Representation overlap and neighbor retrieval (Appendix E.4)	OGG1	Targets holding at least one active compound from public databases
29	Correlation between scoring system and measured activity (Appendix B.4)	SLC6A4, RARG	Targets with at least 20 known active compounds and for which correlation can be measured
29	Mean feature importance (Appendix D.3, Figure 4a)	The two targets where per-target search does not select the ZAO block	Targets whose deployed configuration includes the ZAO block
27	Paired comparison with the analog group (Appendix B.4)	EPHA2 and the three targets where the analog group has fewer than 20 compounds	Only pairs where both groups compute at least 20 compounds on the same target

Changing the pooling rule leaves the sign of the difference unchanged. The median over all replicates with no range cut moves the difference to +6.1 points (the ZAO group is higher on 22 of 30 targets, p = 0.0007), and the minimum over replicates to +11.4 points (24 targets, p = 0.0001), against +5.7 points for the current rule, which is the smallest of the three. Both alternatives leave all compounds in the denominator; neither manufactures the difference by dropping compounds.

Comparison procedure. Since each group covers a different set of targets, each verdict is a paired comparison at the target level, and two groups are paired on a target only when both compute at least 20 compounds there. The unweighted median per-target hit rate is 26.4% versus 16.0%, and the median of the per-target differences is +5.7 points. Applying Boltz-2's cutoff to both groups widens the gap to +16.5 points; for per-target pair counts and hit rates, see Table 7.

Definition of the target set. The 9,288 figure scored in the campaign counts compound-target pairs; the number of distinct compounds is 9,220. Below we write 9,220 when counting compounds and 9,288 when counting scored pairs, and the both-supported set is counted the same way, with 1,329 pairs.

Furthermore, the target count changes from measurement to measurement; Table 3 summarizes why.

B.4 Validity of the Scoring

Validity of the scoring system. On the 29 of this paper's 31 targets that have at least 20 known active compounds and allow correlation to be measured, the median Pearson correlation between computed ΔG and measured activity is 0.528, and the rank correlation is 0.519. Correlations are computed between the activity values (converted to `-log10(M)`) and ΔG, with the sign of ΔG reversed so that the association between more negative ΔG and stronger activity is positive. We confirm with the analog group that the scoring system responds in the expected direction: analogs of known active compounds scored under the same protocol have a median per-target hit rate of 64.8% over the 27 paired-comparison targets. This level of correlation suffices for reading differences in hit rate between groups, but not for reading a single compound's ΔG as that compound's affinity; every judgment in this paper is made on group comparisons, and the individual values attached to Figure 5 are illustrative.

Note that the correlation used for narrowing and the correlation reported here are different measurements. Screening uses values computed against the known active compounds available before this campaign was assembled (Appendix B.1). The values here are re-measured on the set of known active compounds that are finally scored. Therefore, some targets selected on a > 0.4 criterion fall below 0.4 when re-measured (NR3C1 0.360, AR 0.207).

Scoring reliability by target class. Of the 29 targets with measurable correlation, the medians for classes with at least three targets are: protein-protein interactions (3 targets) 0.704, kinases (10 targets) 0.582, GPCRs (3 targets) 0.503, proteases (3 targets) 0.483, receptor extracellular domains (3 targets) 0.468. Four of the five classes fall within 0.06 of the 0.528 median over all 29 targets, and only protein-protein interactions sit 0.176 above it. Classes with one or two targets (hydrolases 0.700, chaperones 0.548, transferases 0.519, nuclear receptors 0.284, cytokines 0.202) are individual values, not medians.

Scoring reliability and ZAO's differences in Appendix D.1 do not order the target classes the same way. Proteases, the class with the largest difference of the five, rank second from the bottom on scoring reliability, and kinases, the class with the smallest, rank second from the top. The differences between target classes are not a reflection of how well the scoring system works on each class.

Stratification by scoring reliability. The ZAO versus Boltz-ESPSim difference is +5.7 over all targets (30 targets), +10.8 in both the > 0.4 and > 0.5 strata (23 and 19 targets), +15.3 in the > 0.6 stratum (6 targets), and +21.3 in the > 0.7 stratum (3 targets). The strata are nested, and the three targets that remain in the > 0.7 stratum are also among the targets where ZAO's difference is largest. All that can be read from this is that the sign is consistent; no reverse trend appears. Of the eight targets where ZAO is lower, the median correlation of the six measurable ones is 0.562, no different from the 0.530 measured over the 28 of the 30 paired-comparison targets with measurable correlation. The four largest shortfalls are SLC6A4 (not measurable), NR3C1 (0.360), AR (0.207), and ITK (0.594); the relationship between these shortfalls and scoring reliability is inconsistent.

C Building the Selected Groups

This appendix explains how the 10,000 or so compounds submitted for scoring are screened down from 1.1 billion molecules. Appendix C.1 describes how the ZAO group is built, and Appendix C.2 describes how both comparison groups are built.

C.1 Selection Model, Filters, and Validation

Selection model and screening. We build the ZAO group with the settings below, taking the union of the two models' predictions to reduce false negatives from a single model. The selection model's features are the 2,048-dimensional ZAO embedding, four RDKit [61] fingerprint families (ECFP4 as Morgan radius 2 / 2,048 bits [63], topological torsion [55] folded to 2,048 bits, ErG 315 [71], the 166 MACCS keys [21] in RDKit's 167-bit vector), and 217 physicochemical descriptors. In what follows, these six blocks, including ZAO, are called feature blocks. Which blocks are used is decided by a per-target search, and the adopted composition varies by target (Figure 4a shows the breakdown); the common configuration is the winner of that search, used unchanged. For gradient-boosted trees, the fixed inputs are ECFP4, physicochemical descriptors, and ZAO. For kernel ridge regression, topological torsion, ErG, physicochemical descriptors, and ZAO are passed in order through a MaxAbsScaler, a 1,024-dimensional Nyström approximation of the RBF kernel [76], and Ridge regression. Screening has six stages, applied in this order: physicochemical properties (logP $\leq$ 5.0, at most five aromatic rings); Tanimoto < 0.6 against all training positives; applicability domain; predicted activity with `-log10(M)` $\geq$ 7.0, i.e., 100 nM (it is sufficient for either of the two models to pass this threshold); ECFP4 bit count within the mean $\pm$ 2 standard deviations of the target's known active compounds; and one compound per chemotype cluster. For the applicability domain, the requirement is that the maximum cosine similarity to the training set in ZAO embedding space be at or above a per-target threshold (the 10th-40th percentile of the distribution of nearest-neighbor similarities computed inside the training set) and that the fingerprint Tanimoto be at least 0.15.

False positives at scale. As the scale increases, more molecules are ranked artificially high by systematic error in the selection score, as has been measured previously for docking scoring functions

(Lyu et al., 2023 [50]), and the same can happen to this paper's selection-model predictions. Among the six stages above, the physicochemical and size limits, the applicability-domain floor, and the limit of one compound per chemotype cluster are the stages that prevent this. The last of these stops the selected compounds from concentrating on one scaffold even if false positives concentrate there. Adding a novelty floor (a Tanimoto similarity cap) to these stages sets two effects against each other. While the novelty floor favors compounds far from the training set, the applicability-domain floor favors those near it. Passing both filters leaves only regions that are far from the training set in 2D but reachable in the representation's space.

C.2 Construction of the Comparison Groups

Candidates for the Boltz-ESPSim group are drawn, for each target, from the designated partition of the enumerated catalog [29] and from a baseline pool common to all targets. Known active compounds are assigned to each group (up to 400 per target, with actual counts ranging from 2 to 400), and the set assigned in this way is called the distributed reference set below. As ESPSim references, we take 10 compounds from that set, picked to avoid electrostatic fields too similar to one another. Candidates are narrowed by how closely their electrostatic potential and shape [10] match those of these 10 compounds, with no 2D similarity constraint. After narrowing by shape and electrostatics, the pool is thinned to at most 5 per Bemis-Murcko scaffold, and the top 1,000 per target go to Boltz-2 [57]. Of these, the compounds that Boltz-2 judges to bind (binding probability ≥ 0.5) are taken in order of decreasing predicted affinity, up to 500 per target. The analog group takes compounds from a pool of some 200 million molecules with a 2D similarity to known active compounds between 0.6 and 0.99 and applies the same Boltz-2 filters to them.

Search-space breadth varies across groups by orders of magnitude: 1.1 billion molecules for the ZAO group, about 200 million for the analog group, and some 11,000-96,000 per target for the Boltz-ESPSim group. That breadth is not automatically an advantage: a predictive model gets less accurate the further it is taken outside the range it applies to, and widening the population also works to lower the hit rate. Narrowing by ESPSim holds the candidates passed to Boltz-2 close to the reference compounds, and in this respect it favors the Boltz-ESPSim group.

D Campaign Results

This appendix provides a breakdown of the results scored using binding free-energy calculations.

D.1 Breakdown by Target Class

For the 30 paired-comparison targets, Table 4 breaks down the per-target-class differences described in Section 3.3 of the main text; the values are per-target hit-rate differences, and the per-group hit-rate levels themselves are shown in Table 7. Ten of the 11 classes have at most three targets, and only the kinases, at 10, have more; the ranking between classes cannot be read at these counts. The breakdown does support two things: the advantage is not confined to one class, and all three nuclear receptors sit together on the side where ZAO is behind.

The five targets with the largest differences span several target classes: F10 and DPP-4 (proteases), DRD1 (GPCR), BRD4 (protein-protein interaction), and OGG1 (hydrolase) (Table 7). ZAO is lower on 8 of the 30 targets: three nuclear receptors (NR3C1, AR, RARG), three kinases (ITK, EGFR, FLT3), one transporter (SLC6A4), and one receptor extracellular domain (CD274). Four targets have differences within about 3 points on the side where ZAO is lower (CD274 0.6, FLT3 2.0, RARG 2.2, EGFR 3.3) and five on the side where it is higher, which means the count of 22 of 30 targets includes targets inside this band.

Target classification follows the campaign's target ledger (EPHA2, NTRK1 and CD274 are assigned to the receptor extracellular domain class because the construct scored is the extracellular domain).

D.2 Novelty: Similarity and Scaffolds

Closeness to known active compounds. The similarity measure and the reference set originally recorded vary from group to group, and the Boltz-ESPSim group has no 2D similarity on record.

Table 4: Breakdown by target class (30 paired-comparison targets). The difference column is the median of the per-target hit-rate differences.

Target class	Targets	Median hit-rate difference	Targets where ZAO is higher
Protease	3	+21.9 points	3
Protein-protein interaction	3	+18.9 points	3
Hydrolase	2	+17.5 points	2
GPCR	3	+16.5 points	3
Receptor extracellular domain	2	+6.2 points	1
Transferase	1	+5.8 points	1
Chaperone	1	+3.0 points	1
Kinase	10	+1.7 points	7
Cytokine	1	+1.3 points	1
Nuclear receptor	3	-10.3 points	0
Transporter	1	-23.9 points	0

We recompute the same measure for each compound: the maximum Tanimoto similarity (Morgan radius 2 / 2,048 bits [63]) to that target's known active compounds put through the same scoring system (the same set as the compounds in the known-active-compound column of Table 7; 42-50 per target over the 31 targets); the ZAO and Boltz-ESPSim group medians are 0.170 versus 0.206. When we divide the compounds into the same similarity bands, the ZAO group's point estimate on the target-unstratified pooled values is higher in all seven bands, by +6.9 to +34.3 points, though the two outermost bands, which contain few compounds, have intervals spanning zero. Within a group, it is not similarity that makes compounds hit. For this comparison, the reference set is the training positives, so these values are not directly comparable to the 0.170 above: the median is 0.300 for hits, 0.284 for non-hits, and 0.307 even for strong hits, values that are essentially the same (the selection condition requires similarity < 0.6, putting the ceiling at 0.598).

Novelty at the scaffold level. For scaffold-level novelty and overlap with training data, see Table 5. The three structural levels are side-chain-stripped Bemis-Murcko scaffolds [9], graph frameworks (from which atom types and bond orders are also stripped), and ring systems per fused-ring unit. Switching the reference set from training positives to the distributed reference set (Appendix C.2) raises all the groups' numbers. The order of the ZAO and Boltz-ESPSim groups does not change, but the analog group and the known active compounds swap at the graph-framework level and at new ring systems (43.8% against 44.7%, and 18.2% against 28.6%); only the values referenced to training positives are shown here. In this table, the ZAO group is the only one to reach 100.0% at the scaffold level, and the figure is 99.4% even at the graph-framework level. Overlap with the training data is counted by exact structural match after standardization; of the ZAO group's 9,220 compounds, zero appear in the selection model's training labels or in the training split of the supervised pretraining set.

The numbers in this section support the novelty described in Section 3.3. Selection imposes 2D similarity to the training positives below 0.6 as a condition, but novelty is measured by scaffold match, a different quantity from that threshold; at the per-target median, 100.0% of the ZAO group's hits have a scaffold absent from the training positives. This is not a re-measurement of the quantity imposed as a condition. Moreover, the zero structural matches against the training data reported above rule out the possibility that the selection model reselected compounds it had seen in training.

D.3 Feature-Block Ablation and Per-Target Results

Removing the ZAO block. To remove it, we keep the common configuration of Appendix A.4 unchanged (gradient-boosted trees: ECFP4, physicochemical descriptors, ZAO; kernel ridge regression: topological torsion, ErG, physicochemical descriptors, ZAO; Appendix C.1), take out ZAO's 2,048 columns, and retrain on the same data and the same split. Hyperparameters chosen on the assumption that ZAO is present remain unchanged.

On early enrichment, the configuration including ZAO is ahead for 28 of 31 targets with gradient-boosted trees and for 29 of 31 with kernel ridge regression (see Table 6 for the per-target values).

Table 5: Structural-level novelty by group and subset. Values are the fractions of compounds whose structure at that level appears in none of the target's training positives, counted per target with the median reported. `Pairs` is the number of compound-target pairs, and `Targets` is the number of targets with at least one compound in that subset. Medians are computed only over targets with at least 20 compounds, so the denominator is smaller than the number in the Targets column (12 targets for strong hits in the Boltz-ESPSim group).

Group	Subset	Targets	Pairs	Novel scaffold	Novel graph framework	Contains novel ring system	Shares no ring system
ZAO group	All	31	9,288	100.0%	99.4%	81.9%	21.2%
ZAO group	Hits	31	2,772	100.0%	99.4%	81.1%	17.0%
ZAO group	Strong hits	31	1,626	100.0%	100.0%	77.7%	15.9%
Boltz-ESPSim group	All	39	10,681	97.9%	84.4%	24.2%	2.9%
Boltz-ESPSim group	Hits	37	2,287	94.7%	77.0%	19.6%	3.1%
Boltz-ESPSim group	Strong hits	34	1,240	92.2%	71.0%	16.7%	1.8%
Analog group	All	38	4,226	51.2%	31.9%	8.5%	0.0%
Analog group	Hits	36	2,519	43.9%	26.7%	6.8%	0.0%
Analog group	Strong hits	35	1,851	42.5%	27.3%	6.7%	0.0%
Known active compounds	All	39	1,816	12.2%	8.3%	4.0%	2.1%
Known active compounds	Hits	39	1,002	4.7%	3.7%	0.0%	0.0%
Known active compounds	Strong hits	38	780	3.6%	3.6%	0.0%	0.0%

The fold-enrichment of active compounds in the top 1% (a multiple of the random baseline) falls from 41.0× to 14.4× and from 39.6× to 12.0×, respectively. On rank correlation, the configuration including ZAO is higher on 30 targets for gradient-boosted trees and 31 for kernel ridge regression, with median per-target differences of +0.220 and +0.253 (Figure 4b).

Losing 2,048 columns is not enough to account for this. When the measurement is repeated with the influence of dimension-coupled hyperparameters removed, the direction is unchanged: the median early-enrichment difference is +0.285 in a version where the kernel width is set according to dimensionality and +0.209 in a version where the gradient-boosted trees' split candidate count is matched by column count. We read the deployed gradient-boosted tree models directly, and feature importance summed by block is 73.3% for ZAO and 17.4% for ECFP4, averaged over the 29 targets with a ZAO block in the model. Given the same 2,048 columns, what the model actually uses is the ZAO side.

The exceptions are few. With gradient-boosted trees the configuration including ZAO falls behind for MAP4K1, IL17A and BCL6; with kernel ridge regression, for BCL6 and CD274. Of those five, BCL6 with gradient-boosted trees and CD274 with kernel ridge regression differ by 0.02 or less, leaving the comparison effectively unchanged.

Table 6: Per-target contribution of ZAO. For early enrichment (BEDROC20), the table shows the setting common to all targets alongside a configuration retrained with only ZAO's 2,048 columns removed on the same data, split, and hyperparameters (GBT, gradient-boosted trees: ECFP4, physicochemical descriptors, ZAO; KRR, kernel ridge regression: topological torsion, ErG, physicochemical descriptors, ZAO; Appendix C.1).

Target	Training positives	GBT common	Same, no ZAO	KRR common	Same, no ZAO
BCL2	1,236	0.766	0.637	0.765	0.573
OGG1	442	0.726	0.517	0.683	0.284
SLC6A4	15,916	0.713	0.488	0.725	0.346
BACE1	15,204	0.701	0.519	0.631	0.496
NTRK1	8,832	0.700	0.465	0.716	0.551
ADORA2A	20,454	0.690	0.520	0.606	0.450
BTK	16,030	0.686	0.553	0.677	0.359
PDE4D	8,223	0.663	0.258	0.652	0.259
BRD4	10,430	0.654	0.321	0.673	0.299
MAP4K1	13,241	0.639	0.667	0.688	0.622
EGFR	17,063	0.634	0.323	0.569	0.254
ABL1	4,629	0.627	0.301	0.568	0.160
IRAK4	12,743	0.613	0.528	0.648	0.383
TYK2	17,945	0.605	0.212	0.631	0.277
EPHA2	649	0.604	0.081	0.538	0.098
RARG	449	0.604	0.441	0.692	0.400
PARP1	9,020	0.590	0.353	0.622	0.362
CDK4	7,155	0.587	0.447	0.494	0.377
HSP90AA1	1,968	0.567	0.263	0.533	0.176
DPP-4	8,151	0.545	0.237	0.504	0.386
IL17A	2,418	0.536	0.676	0.316	0.256
F10	16,661	0.521	0.271	0.473	0.449

Target	Training positives	GBT common	Same, no ZAO	KRR common	Same, no ZAO
ITK	3,542	0.502	0.312	0.526	0.212
BCL6	990	0.494	0.498	0.400	0.505
FLT3	7,800	0.490	0.277	0.439	0.184
CD274	8,692	0.484	0.469	0.541	0.559
MKNK2	3,025	0.479	0.237	0.532	0.111
AR	6,694	0.474	0.308	0.431	0.196
DRD1	4,207	0.435	0.240	0.426	0.177
NR3C1	9,641	0.433	0.350	0.445	0.333
HTR2B	5,994	0.427	0.174	0.377	0.143

Per-target results. Pair counts and hit rates for each target appear in Table 7. Since selection cuts at about the same count on each target, the ZAO group has a median of 331 pairs per target and 27 of the 31 targets have at least 250. The four that fall below are EPHA2 (9), OGG1 (54), DRD1 (114), and BCL6 (248), while the Boltz-ESPSim group ranges from 20 to 499. Moreover, hit-rate levels range from 2.3% to 77.2% across targets, a spread an order of magnitude larger than the difference between groups (+5.7 points at the per-target median). A value that pools compounds across targets is governed by how the groups' target compositions differ, which is why Appendix B.3 restricts all verdicts to per-target paired comparisons.

Similarly, hit rates for the known active compounds range from 17.0% to 93.8%; the targets themselves are not equally difficult. The ZAO group exceeds the known active compounds' hit rate on 4 of the 31 targets (DRD1, HTR2B, NR3C1, SLC6A4). However, this cannot be read as the selected compounds being better than the known active compounds, because the baseline is not fixed: it moves from target to target.

Table 7: Per-group pair counts and hit rates ($\Delta G \leq -10$ kcal/mol) over the 31 targets with ZAO group calculations. Paired comparisons use the 30 targets with at least 20 compounds in both groups, so EPHA2 drops out of the median computation, and the total is 9,288 pairs. Target names are gene symbols (only DPP-4 retains its conventional notation). When the construct is the extracellular or kinase domain, the domain is noted in parentheses.

Target	Class	ZAO group	Hit rate	Boltz-ESPSim group	Hit rate	Known active compounds	Hit rate
DRD1	GPCR	114	77.2%	499	41.7%	47	66.0%
HTR2B	GPCR	344	61.0%	498	44.6%	44	54.5%
NR3C1	Nuclear receptor	330	57.9%	453	70.9%	49	57.1%
ADORA2A	GPCR	337	52.2%	23	47.8%	50	58.0%
HSP90AA1	Chaperone	318	48.4%	260	45.4%	45	62.2%
SLC6A4	Transporter	337	47.2%	356	71.1%	44	31.8%
BCL2	Protein-protein interaction	329	46.5%	29	27.6%	44	61.4%
AR	Nuclear receptor	340	39.1%	397	49.4%	47	57.4%
CD274 (extracellular domain)	Receptor extracellular domain	330	38.8%	132	39.4%	49	40.8%
F10	Protease	294	38.4%	32	0.0%	42	54.8%
OGG1	Hydrolase	54	31.5%	480	10.2%	45	64.4%
MAP4K1	Kinase	333	30.9%	20	30.0%	48	93.8%
BRD4	Protein-protein interaction	283	29.3%	486	6.0%	50	76.0%
NTRK1 (extracellular domain)	Receptor extracellular domain	333	29.1%	111	16.2%	49	38.8%
FLT3 (kinase domain)	Kinase	332	26.5%	260	28.5%	48	75.0%
DPP-4	Protease	312	26.3%	68	4.4%	45	48.9%
CDK4	Kinase	337	24.6%	297	7.7%	48	60.4%
PARP1	Transferase	331	24.5%	424	18.6%	50	58.0%
ABL1	Kinase	336	22.9%	313	12.1%	50	50.0%
TYK2	Kinase	344	20.6%	400	6.8%	48	62.5%
ITK	Kinase	333	20.1%	229	27.1%	49	83.7%
PDE4D	Hydrolase	337	19.6%	480	5.8%	49	40.8%
BACE1	Protease	318	18.9%	136	13.2%	44	22.7%
IRAK4	Kinase	342	17.8%	388	15.5%	50	68.0%
MKNK2	Kinase	299	17.4%	297	17.2%	44	65.9%
BTK	Kinase	317	16.1%	456	5.9%	49	65.3%
BCL6	Protein-protein interaction	248	14.1%	471	1.7%	49	87.8%
RARG	Nuclear receptor	346	13.6%	387	15.8%	47	17.0%
EGFR	Kinase	330	11.2%	159	14.5%	48	66.7%
EPHA2 (extracellular domain)	Receptor extracellular domain	9	11.1%	73	37.0%	49	49.0%

Target	Class	ZAO group	Hit rate	Boltz-ESPSim group	Hit rate	Known active compounds	Hit rate
IL17A	Cytokine	341	2.3%	400	1.0%	48	41.7%

D.4 Selected Compounds and Their Neighbors

To show, across a range of compounds and targets, how far the selected compounds are from the training data, we arrange them as follows. We restrict the set to targets where the calculations reproduce measured activity well (Pearson r ≥ 0.5), and we keep the hits ($\Delta G \leq -10$ kcal/mol) whose scaffold is absent from that target's training positives and that meet the drug-likeness and ring-saturation conditions. This leaves 14 targets and 80 compounds. Figure 5 presents a total of 21 compounds, three for each of the seven targets with the most qualifying compounds (ADORA2A 15, HSP90AA1 13, ABL1 10, IRAK4 6, BCL2 6, F10 5, CDK4 5). To the right of each compound are six of that target's training positives, taken in descending order of 2D similarity while skipping any whose Tanimoto to a neighbor already taken is 0.40 or above, because neighbors drawn from a single series would hide how wide that target's training data is. For all 21 compounds, the closest training positive lies between 0.143 and 0.329 in 2D similarity, and none of the 126 neighbors drawn is above 0.33. The absence of anything close in the training data looks the same across targets and across compounds.

E Measurements on the Frozen Representation

This appendix measures what can be extracted linearly from the frozen representation emitted by the released checkpoint. Section 4.1 of the main text corresponds to Appendices E.1 and E.2, Section 4.2 to the control conditions of Appendix E.1 and to Appendices E.2 and E.3, and Section 4.3 to Appendix E.4. One exception is the removal experiment that takes the ZAO block out of the selection model, reported in Appendix D.

E.1 Probe Design and Full Results

Across the probes, 37 quantities fall into five categories (conformer energy: 4; shape: 7; 3D pharmacophore distances: 6; set-specific metrics: 7; 2D properties: 13), listed individually in Table 8. For set-specific metrics, the 2,945 molecules of this appendix use all seven metrics in Table 8. The mean pairwise RMSD between conformers is not computed for some 20,000 molecules, so those measurements use the remaining six. The separate aggregation run in Appendix E.3 uses six. The number in each heading is the number of metrics actually used in that table.

Overlap with pretraining. Of the 37 quantities the probes extract, conformer energy, shape, and 2D properties overlap through auxiliary heads or as regression targets; pharmacophore distances and set-specific metrics overlap indirectly, in that auxiliary heads predict the distances between the 64 query centroids and the variance of those distances.

Molecule sets and probe settings. Molecules are drawn from the split of the supervised portion of pretraining that holds out measured labels (Appendix A.3): 2,945 used by all conditions plus a disjoint 20,000 for statistical power (22,945 in total). Measurements requiring conformer regeneration use the 22,913 molecules with successful generation. Within each measurement block, probe comparisons run over the molecules that have complete rows for all conditions of that block (counts in each table). There are 11 conditions including ZAO, and adding the three control-pair conditions of Appendix G brings the total to the 14 settings shown in Table 2. For the 2D-descriptor condition, we use 216 of the selection model's 217 physicochemical descriptor columns, dropping Ipc for its extreme range. As the head, we use ridge regression, and all coefficients of determination are computed from out-of-fold predictions. We run five-fold outer cross-validation; regularization strength is chosen per target variable within the training split by a further five-fold search over seven candidates (1e-1 to 1e5), and the test split is never touched. Only the fingerprint condition, which is sparse and binary, skips division by the column standard deviation: scaling sparse columns amplifies rare bits and would unfairly penalize the fingerprint baseline.

Equalization across conditions, and its limits. For the mean over metrics after matching dimensions to 512, the four 10-conformer conditions and the 2D fingerprints are each projected to 512

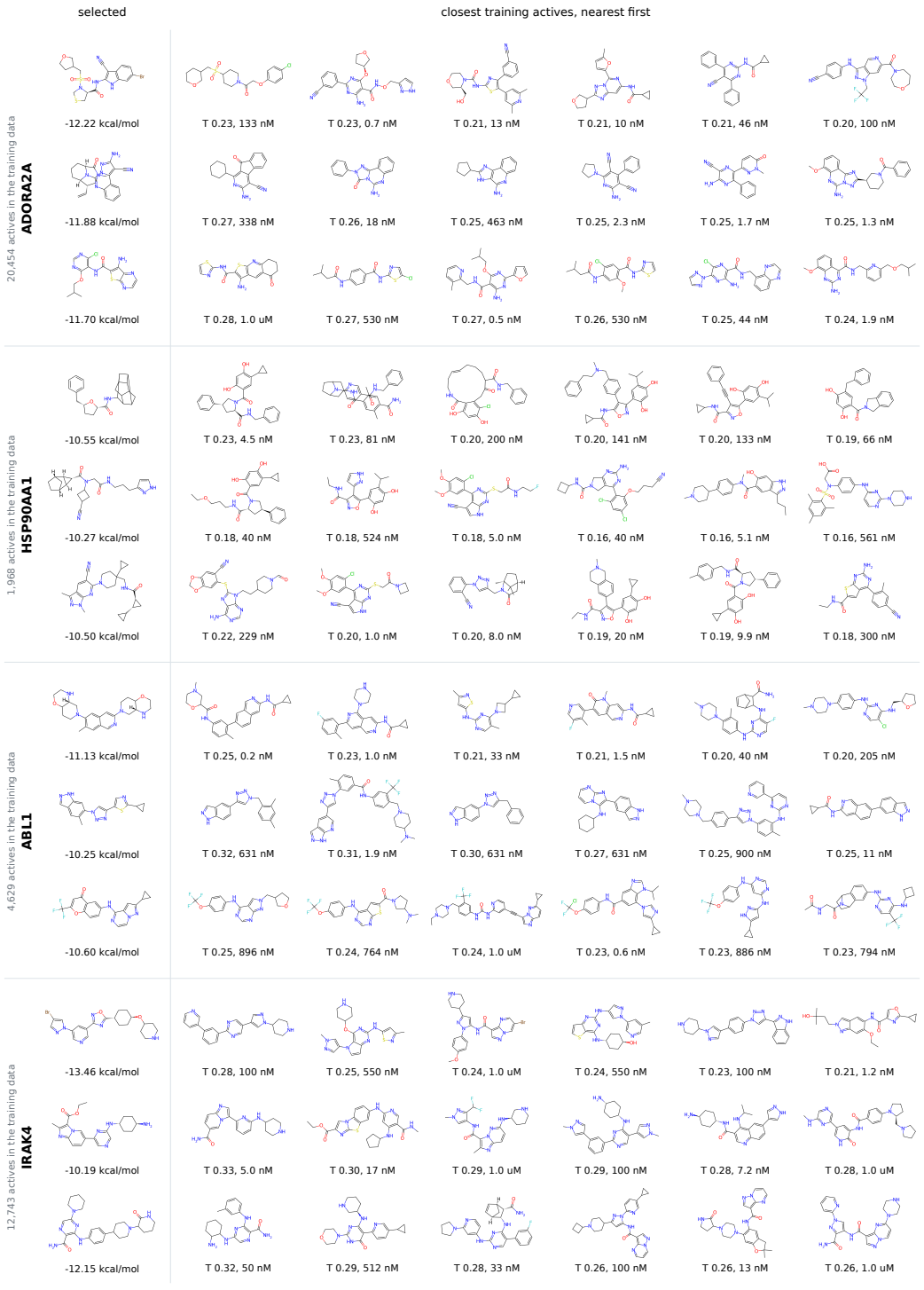

Figure 5: Selected compounds and their neighbors in the target's training data, three compounds for each of seven targets in 21 rows. In the left column are the compounds selected out of the 1.1 billion and calculated to be hits, labeled with the calculated binding free energy. The six columns to their right are that target's training positives (1 μM or better) in descending order of 2D similarity to the compound on the left, labeled with the 2D similarity T (defined in Section 3.2) and the activity measured for that molecule. In the left margin are the target and its total number of training positives. Targets are selected as described above, and the three compounds per target are examples, not a sample.

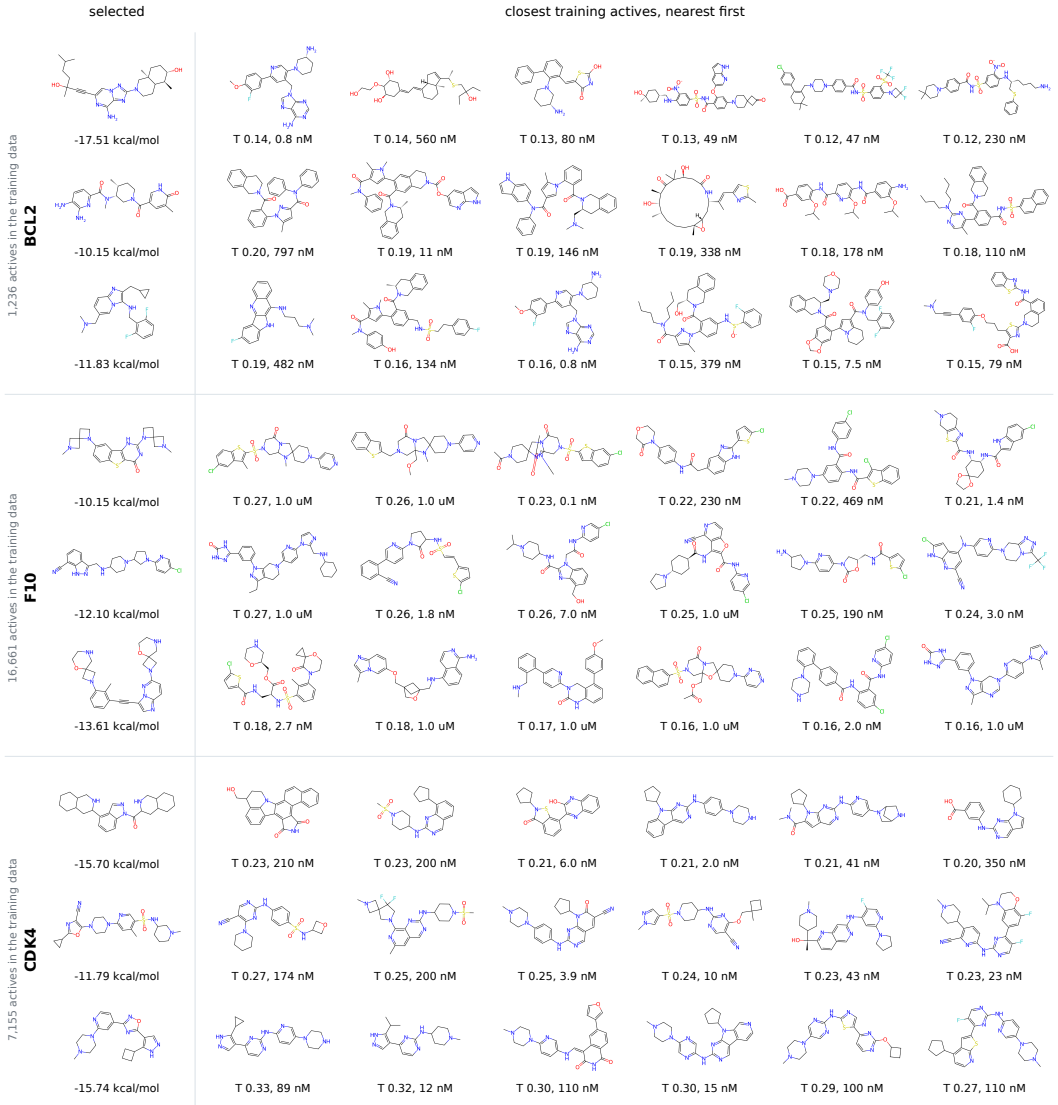

Figure 5 (continued).

Table 8: The 37 quantities extracted by the probes.

Category	Quantity
Conformer energy (4)	Mean conformer energy, standard deviation of conformer energy, range of conformer energy, entropy of the Boltzmann distribution
Shape (7)	Radius of gyration, asphericity, eccentricity, sphericity, principal moment of inertia ratio 1, principal moment of inertia ratio 2, inertial shape factor
3D pharmacophore distances (6)	Maximum distance between hydrogen-bond donor and acceptor, mean of the same, maximum distance between aromatic rings, maximum distance between hydrophobic groups, maximum inter-heavy-atom distance, mean inter-heavy-atom distance
Set-specific metrics (7)	Across-conformer standard deviation of radius of gyration, of asphericity, of principal moment of inertia ratio 1, of principal moment of inertia ratio 2, across-conformer range of radius of gyration, across-conformer standard deviation of maximum inter-heavy-atom distance, mean pairwise RMSD between conformers
2D properties (13)	Molecular weight, cLogP, polar surface area, hydrogen-bond donor count, hydrogen-bond acceptor count, rotatable bond count, sp3 carbon fraction, ring count, aromatic ring count, stereocenter count, maximum partial charge, minimum partial charge, mean absolute partial charge

dimensions with the same Gaussian random projection, all 36 metrics are re-measured, and the results are averaged over five random seeds (seed-to-seed variation has a standard deviation of at most 0.0023). Uni-Mol and ConforFormer are natively 512-dimensional, and the results for those two barely change; the projection is not what destroys information. Projected fingerprints are neither sparse nor binary, and they are scaled like the other conditions here. Two things cannot be equalized. One is hydrogen handling, which follows the released checkpoints' own settings: ConforFormer and Uni-Mol2 strip hydrogens; ZAO and Uni-Mol do not. The other is that none of the three public encoders was trained to emit conformer-set embeddings, and the 10-conformer average is an operation that we impose.

Results for all conditions. Per-category, per-condition coefficients of determination are shown in Table 2; for the 19,974 molecules complete across all conditions, measurements use the same rows, splits, and target variables. With shuffled target variables, the floor in all five categories is near 0. Thus, the differences below reflect the content of the representations; they do not arise from the degrees of freedom available to the fit.

In the three categories that require 3D, ZAO leads the best public encoder by +0.068 on shape and +0.042 on pharmacophore distance, against +0.219 on set-specific metrics. On quantities determined by a single structure, the margin is small; on the set-specific metrics it is large. The table also shows the reverse direction: on 2D properties the 216-column 2D descriptors reach 0.995 against ZAO's 0.939, and those same descriptors reach only 0.479 on shape. Conditions differ in the kind of information they retain, and no single one leads in all categories; this split is the basis for Section 4.1's statement that 3D information is not an extension of 2D features.

Column count does not account for these results. At their native dimensions, the 2,048-column 2D fingerprints average 0.617 over all 36 metrics, below the 0.702 of the 512-column Uni-Mol 10-conformer average. Matching all conditions to 512 dimensions in the rightmost column leaves ZAO highest, at 0.768. Hence, the result does not come from handing the probe more columns.

Regenerated-conformer condition. In addition, we measure a condition that reads regenerated conformers of the same molecules, generated separately from the 10 conformers used to compute the target variables. Over the 19,966 molecules with rows complete across all conditions (including this one), category means are energy: 0.365; shape: 0.812; pharmacophore distance: 0.841; set-specific metrics: 0.700; 2D properties: 0.939, and the mean over all 36 metrics is 0.794. All four categories that use coordinates drop; the coordinate-independent 2D properties remain at 0.939. The Table 2 conditions come out higher since they read the same conformers from which the targets are computed, and the penalty is not small: for shape, the gap between the real 10 conformers and the regenerated ones is +0.046, about half of the +0.093 gap between the real 10 conformers and one; for set-specific metrics, the regenerated gap is +0.104 against a 10-versus-1 gap of +0.201. Therefore, absolute differences between conditions in Table 2 are upper bounds. As measured on

Table 9: Compute-matched controls and geometry-rebuilt controls (20,000 molecules).

Condition	Conformer energy 4 metrics	Shape 7 metrics	Pharmacophore distance 6 metrics	Set-specific metrics 6 metrics	2D properties 13 metrics
Real 10 conformers	0.421	0.858	0.856	0.805	0.939
Most stable 1 conformer only	0.339	0.766	0.822	0.602	0.927
Same conformer $\times 10$ (most stable)	0.340	0.766	0.821	0.598	0.927
Same conformer $\times 10$ (6th)	0.351	0.771	0.824	0.598	0.925
10 conformers with coordinate noise	0.192	0.550	0.572	0.470	0.573
10 conformers with torsion noise	0.316	0.743	0.797	0.601	0.875
Real 10 conformers, energy fixed at minimum	0.595	0.863	0.856	0.815	0.934
Real 10 conformers, energies permuted	0.436	0.858	0.855	0.805	0.939
Label shuffle	-0.007	0.002	-0.005	0.007	-0.002

ZAO, around half of the difference arises because the target and the condition share one conformer generation; nothing guarantees that this ratio is the same for the other conditions. Despite this penalty, the ordering of the conditions and the direction of the effects remain unchanged. When a conformer-independent effect size is needed, use Table 10, which varies only the count within the same pool.

The control conditions appear in Table 9 (20,000 molecules). None of them generates new conformers; all are built by perturbing or replicating the existing 10. Noise magnitude is calibrated per molecule to the actual mean pairwise RMSD of that molecule's 10 conformers (median 1.760 Å). Adding noise to the atomic coordinates themselves, the coordinate-noise condition preserves neither bond lengths nor bond angles, and the resulting structures are not chemically valid; we include it to make clear what the torsion-noise condition preserves and what it breaks. Finally, we place the one-conformer condition in this table, and all condition-vs-condition comparisons happen within it.

Of the six conditions, the two that replicate one conformer 10 times remain within 0.012 of the single most stable conformer in all categories (0.598 and 0.598 against 0.602 on set-specific metrics). Passing the same structure through the encoder 10 times does not change what can be read out; what separates the replicated conditions from the real 10 conformers is the content of the structures read, not the number of forward passes. Section 4.2 concludes from the geometry-rebuilt controls that force-field-optimized real conformers are what is needed; the replicated conditions close off, one step earlier, the alternative explanation that the gain comes from extra compute.

E.2 Conformer Count, Coverage, and Energy

Conformer count and coverage miss. Full results for both series appear in Table 10. We measure how broadly a conformer set covers the shapes a molecule can take by taking, for each of the 10 released conformers, the minimum superposition error relative to the sampled subset and averaging over the 10 (hereafter, coverage miss; smaller means broader coverage). Coefficients of regressions that include it as a covariate are shown in Table 11.

How the result improves with the count differs from category to category. In the random-draw series, shape gains 0.068, from 0.742 at one conformer to 0.810 at 10, then moves only 0.014 out to 50. Set-specific metrics rise 0.109 by 10 conformers and a further 0.030 from there to 50. 2D properties go from 0.925 to 0.940 and barely respond to the count. The more a quantity depends on coordinates, the more it responds to the count, and only the metrics that describe the spread between conformers fail to flatten by 50. The 10 conformers ZAO uses sit where shape has saturated and set-specific metrics are still rising. The saturation point that Section 4.2 reads off Figure 3d is the value in this table.

How the conformers are drawn matters too: 10 drawn at random (shape 0.810, set-specific metrics 0.701) match 30 drawn in ascending energy order (0.814, 0.709). Compared with a random draw at equal counts, ascending energy is worse: at 10 conformers, shape is 0.789 against 0.810 and set-

Table 10: Two series varying how many conformers are drawn from a pool of 50 (19,982 molecules): one takes conformers in ascending energy order, while the other shuffles them with a fixed per-molecule random permutation and takes the first N.

Selection	Count	Conformer energy 4 metrics	Shape 7 metrics	Pharmacophore distance 6 metrics	Set-specific metrics 6 metrics	2D properties 13 metrics
Ascending energy	1	0.337	0.748	0.816	0.598	0.927
Ascending energy	5	0.345	0.775	0.829	0.627	0.935
Ascending energy	10	0.354	0.789	0.834	0.651	0.937
Ascending energy	15	0.360	0.797	0.837	0.666	0.938
Ascending energy	20	0.364	0.805	0.840	0.684	0.939
Ascending energy	30	0.369	0.814	0.843	0.709	0.940
Ascending energy	50	0.373	0.824	0.845	0.731	0.940
Random	1	0.341	0.742	0.815	0.592	0.925
Random	5	0.356	0.798	0.835	0.668	0.937
Random	10	0.363	0.810	0.840	0.701	0.939
Random	15	0.367	0.816	0.842	0.713	0.940
Random	20	0.368	0.819	0.843	0.719	0.940
Random	30	0.371	0.822	0.844	0.726	0.940
Random	50	0.373	0.824	0.845	0.731	0.940

Table 11: Coefficients from a regression with per-molecule fixed effects. With per-molecule standardized squared error as the target variable, a negative coefficient means the factor reduces error, and a smaller coverage miss means broader coverage. In the left column the coefficient comes from a regression with conformer count as the only explanatory variable, while the middle and right columns are coefficients from the same regression with coverage miss added as a covariate.

Category	Count only	Count + coverage miss	Coverage-miss coefficient
Conformer energy 4 metrics	-0.0061	-0.0040	+0.0089
Shape 7 metrics	-0.0125	+0.0118	+0.1036
Pharmacophore distance 6 metrics	-0.0055	+0.0042	+0.0417
Set-specific metrics 6 metrics	-0.0235	+0.0162	+0.1697
2D properties 13 metrics	-0.0024	-0.0010	+0.0060

specific metrics are 0.651 against 0.701. Drawing from the low-energy side needs three times the count to catch up with a random draw.

Adding coverage miss as a covariate flips the sign of the count coefficient in the three categories that require 3D (shape from -0.0125 to +0.0118, pharmacophore distance from -0.0055 to +0.0042, set-specific metrics from -0.0235 to +0.0162). For 2D properties and conformer energy, which do not use coordinates, the coefficient remains negative and merely shrinks. In the three categories that flip, the coverage-miss coefficient is large, from +0.0417 to +0.1697: what the error corresponds to is not the count itself but how broadly that count covers the shapes. However, count and coverage miss cannot be varied independently, and this regression cannot decide which is the cause. Separately, the difference between the two series above shows that the same count leads to different results depending on how it is drawn.

Stability under regeneration. To measure how far structures move on regeneration, we take the superposition error between the leading conformers before and after regeneration; the median is 1.463 Å (the 1.760 Å reported in Appendix E.1 is the spread among conformers of the same molecule, a different quantity). We search the original full embedding set with each molecule's regenerated embedding and ask how often the same molecule ranks first (hereafter, self-identification rate); the result is 78.66% at one conformer and 88.36% at 10, with 45% of the misses disappearing at 10. Agreement within the top 10 is 86.5% and 94.8%. The median cosine similarity remains unchanged, 0.978 versus 0.977; the change appears in the lower tail (the 5th percentile moves from 0.696 to 0.809).

As shown in Table 12, the gain is monotone in flexibility and size and is larger in strata where one conformer does worse. In the stratum with nine or more rotatable bonds, the one-conformer self-identification rate falls to 59.33%, against 79.70% with 10 conformers. Between strata the spread narrows from 35.7 points at one conformer to 17.8 points at 10: reading a conformer set raises the

Table 12: Self-identification rate when searching the original full embedding set with regenerated embeddings, stratified by molecular property. Molecules with at most one ring are absent from this set.

Stratum	Value	Molecules	1 conformer	10 conformers	Difference
Rotatable bonds	0-2	2,991	95.02%	97.49%	+2.47
Rotatable bonds	3-4	7,833	84.97%	91.31%	+6.34
Rotatable bonds	5-6	6,165	76.20%	86.52%	+10.32
Rotatable bonds	7-8	3,239	68.97%	83.45%	+14.48
Rotatable bonds	9+	2,685	59.33%	79.70%	+20.37
Heavy atoms	≤ 20	3,350	94.60%	98.90%	+4.30
Heavy atoms	21-25	4,591	87.00%	92.96%	+5.96
Heavy atoms	26-30	6,344	78.34%	86.24%	+7.90
Heavy atoms	31-35	5,291	71.22%	83.29%	+12.07
Heavy atoms	36+	3,337	63.59%	83.49%	+19.90
Ring count	2	5,518	84.70%	92.13%	+7.43
Ring count	3	3,671	86.33%	93.35%	+7.02
Ring count	4+	13,724	74.18%	85.50%	+11.32

Table 13: Movement of features stratified by RMSD between conformers (5,000 molecules; 225,000 pairs). Values are relative to a cosine distance of 1 between unrelated molecules for each encoder, and larger values indicate greater sensitivity to conformer differences. Rank correlation is Spearman correlation between RMSD and cosine distance across pairs.

Encoder	<0.5 Å	0.5-1.0	1.0-1.5	1.5-2.0	2.0-3.0	>3.0 Å	Slope	Rank correlation
ConforFormer	0.028	0.165	0.267	0.361	0.491	0.635	0.178	0.626
Uni-Mol	0.041	0.162	0.248	0.319	0.417	0.560	0.149	0.661
Control pair (10 conformers)	0.027	0.095	0.147	0.204	0.279	0.358	0.099	0.542
ZAO	0.012	0.062	0.115	0.172	0.242	0.306	0.090	0.568
Control pair (1 conformer)	0.014	0.050	0.077	0.111	0.159	0.240	0.067	0.550
Uni-Mol2	0.013	0.031	0.026	0.034	0.039	0.060	0.012	0.530

average and also evens out how much the result hinges on the kind of molecule. Ring count is the one property that is not monotone, with three rings slightly higher than two (86.33% versus 84.70%); fused-ring count is no substitute for flexibility.

Failure modes. The 2D similarity to the molecule retrieved by mistake has medians of 0.407 and 0.506, above the 0.111 of random pairs, and in 23.0% and 30.4% of cases even the scaffold is the same: some of the remaining failures are not missing molecules but mistaken retrievals of close analogs.

Energy. One condition keeps the real conformers as they are but fixes the energy feature at its minimum, making the Boltzmann weights uniform; this condition outperforms the real 10 conformers in the energy category (Table 9; 0.595 versus 0.421), and the difference comes from the spread and range of the energies. This is consistent with a temperature of 1.0 kcal/mol concentrating weight on low-energy conformers and discarding the high-energy conformers that define the spread. That bound-state conformers deviate from the most stable conformer in solution has been documented for many years (Perola and Charifson, 2004 [59]), and an aggregation that concentrates weight on the low-energy side pulls against that fact. Note that a condition that merely permutes energies across conformers barely changes the result (0.421 versus 0.436 in the energy category and 0.805 versus 0.805 in set-specific metrics); what matters is the energy distribution, not which energy belongs to which conformer.

Sensitivity to subtle conformer differences. Whether an encoder reads coordinates shows up in how far its features move as the difference between conformers grows. For 5,000 molecules (random seed 20260925), we pair symmetry-corrected heavy-atom RMSD with the cosine distance between features obtained by reading one conformer at a time, for all 45 pairs from the 10 stored conformers. Across the 225,000 pairs the mean RMSD is 1.778 Å, and for each encoder the values are normalized by the cosine distance between unrelated molecules; the results appear in Table 13.

From the lowest to the highest RMSD interval, the numbers of pairs are 21,743, 24,044, 41,524, 50,097, 65,195, and 22,397. Only Uni-Mol2 is flat: its movement remains 0.060 even above 3 Å RMSD and is not even monotonic in the 1.0-1.5 Å interval. It increases only fivefold across all intervals, compared with 25-fold for ZAO and 23-fold for ConforFormer. Uni-Mol2 appears robust to the mild perturbation of replacing a single conformer, its features barely moving as the geometry changes, whereas the features of the other five conditions rise monotonically with RMSD.

E.3 Controls, Aggregation, and the Learned Queries

Scaffold-unit splits. Re-splitting by scaffold unit costs ZAO only 0.002-0.005 in the category mean (the largest drop is 0.025, for 2D fingerprints on shape). This rules out the possibility that the results of Appendices E.1 and E.2 are produced by the same scaffolds appearing in both the training and the test split.

Place of aggregation. Per-condition coefficients of determination for set-specific metrics, measured on 2,945 molecules, are: learned aggregation 0.769, plain averaging 0.657, energy-weighted averaging 0.627, Uni-Mol [81] one conformer 0.471, 216 2D descriptor columns 0.504, and 2D fingerprints 0.440. This section uses a separate run built for the aggregation comparison, and it measures set-specific metrics with its own set of six: the seven in Table 8 minus the across-conformer standard deviation of the maximum distance between heavy atoms. The Table 2 family's six are the same seven minus the mean pairwise RMSD, so the two sets differ in content. Learned aggregation performs better on all six metrics, by +0.029 to +0.192.

On the five size-independent shape metrics (asphericity, eccentricity, sphericity, and the two principal moment of inertia ratios), learned aggregation scores 0.808 against plain averaging 0.798 and energy-weighted averaging 0.673, and its advantage over plain averaging disappears. It disappears on these five because averaging the 10 representations is close to the optimal operation when the quantity predicted is the mean of the per-conformer values over the 10 conformers. Two metrics remain (radius of gyration and inertial shape factor); both scale with molecular size and are treated separately. This split, between the metrics where the advantage appears and those where it does not, coincides with the limit of the averaged construction shown next.

The limit of the averaged construction. A construction that emits an independent scalar per conformer and then averages uniformly can express only permutation-invariant functions that decompose additively across conformers. This follows almost immediately from the definition and can be checked with a set of just two elements: suppose G({x, y}) = (x - y)^2 could be written as [f(x) + f(y)]/2 for some f. Substituting x = y gives f(x) = 0 for every x, so f is identically 0; then G = 0 even for $x \neq y$, contradicting G > 0.

This limit constrains the average of the predictions; it does not apply to constructions that average embeddings and then pass them through a nonlinear map, or that use set-dependent weights. However, the probes in this paper are linear: applying a linear probe to an averaged embedding is identical to applying it to each conformer and then averaging, and is subject to the same constraint.

Nor can the descriptor columns' performance be explained as a proxy for size: with the same procedure, heavy-atom count on its own scores 0.188, rotatable-bond count 0.281, and the two columns together 0.308, against 0.504 for the 216 2D descriptor columns.

Attention of the 64 queries. We check whether the 64 queries respond to chemical substructures. For 1,968 molecules, we match each query's attention over atoms against the 43 of RDKit's 85 FragmentDescriptors functional-group patterns that occur often enough in this molecule set. We then make the same comparison for a control that only scrambles atom correspondence within molecules: 61 of the 64 queries show at least twofold enrichment over chance, with a maximum of 22.7-fold, while no query in the control reaches twofold. Cosine similarity between query pairs averages 0.006 (some pairs are aligned; the maximum is 0.835). Which query attaches to which functional group is decided by the initialization; what is established here is only that the correspondence exists.

E.4 Representation Overlap and Retrieval

Design. The subjects are 12,406 active compounds measured at $\leq$ 100 nM in public databases (spanning 30 targets). The pair population is built across targets; the property of sharing a target does

Table 14: Overlap between representations (12,406 molecules). A linear CKA closer to 1 means the same information, and the two correlation columns are Pearson correlations. After stratification, pair counts vary across fingerprint settings and are 112,404 for radius 2 / 2,048 bits.

Pair	Linear CKA	Similarity-decision correlation	Tanimoto-stratified correlation
ZAO × Morgan radius 2 · 2,048 bits	0.290	0.329	0.554
ZAO × Morgan radius 3 · 2,048 bits	0.279	0.328	0.551
ZAO × Morgan radius 2 · 1,024 bits	0.287	0.350	0.547
ZAO × Morgan radius 4 · 128 bits	0.241	0.638	0.802
ZAO (before L2 normalization) × Morgan radius 2 · 2,048 bits	0.235	—	—
Morgan radius 2 · 2,048 × radius 3 · 2,048	0.948	—	—
Morgan radius 2 · 2,048 × radius 2 · 1,024	0.962	—	—
Morgan radius 2 · 2,048 × radius 4 · 128	0.376	—	—
Morgan × Morgan with rows permuted	0.017	—	—

not arise from how the sample is constructed. Over all 76.95 million pairs, we compute Tanimoto on Morgan radius 2 / 2,048-bit fingerprints and cosine similarity between embeddings, restricting to pairs that are distant in 2D (Tanimoto < 0.20).

Full representation-overlap results are shown in Table 14, with linear CKA reported for each pairing of the ZAO embeddings with one of four Morgan fingerprint settings, and for pairs of fingerprints. Agreement between similarity judgments is measured on 1,999,839 random pairs; since the Tanimoto values of random pairs concentrate around a mean of 0.123, we also re-measure on 112,404 pairs stratified uniformly over Tanimoto. Linear probes are run as ridge regression with five-fold cross-validation; in cross-validated variance explained, fingerprints explain 49.2% of the embedding, the embedding explains 41.3% of the fingerprints, and, as a reference, one fingerprint explains 78.2% of another.

Within the table are its own upper and lower reference points. Re-measuring the same fingerprint with only the radius or the bit count changed returns 0.948 and 0.962, the values this metric takes when the same information is measured under another setting. Permuting the rows to break the correspondence between molecules leaves 0.017, the floor for unrelated representations. The ZAO-fingerprint pairs run from 0.241 to 0.290, clearly off the floor and far from the level of one information set measured two ways. Before L2 normalization, the value is 0.235; this is not an artifact of normalization. The variance the linear probes explain, reported above, separates the fingerprint-to-fingerprint comparison from the comparisons involving ZAO by the same margin. The measurement's position between these two reference points is the basis for Section 4.3's use of ZAO as an addition to 2D fingerprints, not a replacement.

However, among fingerprints themselves, the pair with Morgan radius 4 cut down to 128 bits is only 0.376. ZAO's 0.241 against that setting cannot be separated from the effect of the reduced bit count on its own. The claim above rests on the pairs with the remaining three settings, from 0.279 to 0.290.

Full retrieval results appear in Table 15. For the 12,401 molecules with computable 3D shape descriptors, we restrict each measure's neighbors to partners with 2D similarity < 0.20 and count how often the neighbor is a compound active against the same protein. The USRCAT row in this table ranks by descriptors averaged over 10 conformers and is computed differently from the agreement values in Table 16 (best over conformer combinations). The USR profile is the 24-dimensional vector formed by the mean and the standard deviation (over the 10 conformers) of the 12-dimensional USR descriptor [6], with pair similarity built from the L2 distance as `1/(1+d)`. This is the quantity the pretraining similarity loss uses as its target (Appendix A.2). USRCAT [67] is 60-dimensional, split by pharmacophore type, and does not enter the pretraining objective.

The fraction of molecules with a top-1 neighbor active against the same protein is highest for ZAO at 30.3%, against 15.2% for USRCAT (the best of the existing measures), 13.0% for 2D similarity, and 2.9% at random. Since candidates are restricted to partners distant in 2D, 2D similarity has almost nothing left to rank by in this range.

Table 15: Neighbor retrieval restricted to partners distant in 2D. Each value is the fraction of molecules whose neighbor is a compound active against the same protein. Parentheses give molecule counts.

Measure	Top 1
ZAO embedding cosine	30.3% (3,754)
2D similarity	13.0% (1,615)
USRCAT	15.2% (1,880)
USR profile	7.7% (951)
ZAO residual (after regressing out four properties)	16.9% (2,092)
Random	2.9% (364)

Scaffolds of the retrieved pairs. We take pairs with 2D similarity < 0.20 and order them by descending ZAO cosine. From the top 100,000, we draw 20,000 pairs at random with a fixed seed, and compare scaffolds for the 19,948 pairs (2,271 distinct molecules) that have 3D shape descriptors for both molecules. Bemis-Murcko scaffolds [9] match for zero pairs; even if we loosen to graph frameworks (which ignore atom types and bond orders), one pair matches.

Scope of the advantage. Over 2D-distant pairs, four descriptors (molecular weight, heavy-atom count, cLogP, and polar surface area) explain 62.7% of the variance of cosine. When pairs are reordered by the residual left after regression removes all four, the top-1 match drops to 16.9%, still far above the random baseline's 2.9%.

For the 3D shape agreement values see Table 16, where each stratum contains about 20,000 pairs. Only the top of the cosine range shows an effect: the mid-range stratum is indistinguishable from the random stratum on both 3D agreement and same-target fraction (USRCAT 0.2066 against 0.2055; same target 2.79% against 2.82%). Only in the stratum with the closest embeddings do these rise, to 0.2591 and 6.16%. Overall, embedding cosine does not track 3D closeness across the whole range; it tracks it at the close end.

The two 3D measures behave differently: moving from the random stratum to the stratum with the closest embeddings, USRCAT moves up by 26%, from 0.2055 to 0.2591, whereas the USR profile doubles, from 0.1115 to 0.2311. However, the two columns do not stand on equal footing. The USR profile is the very quantity the pretraining similarity loss uses as its target (Appendix A.2), and it is the heaviest term in that target; embedding cosine is explicitly trained to correlate with it. Therefore, part of the rise in the USR profile is the training objective showing through, and the rise is not independent evidence. USRCAT is the measure that does not enter the objective, and it is the one Section 4.3 cites. The claim that the retrieved pairs share shape and pharmacophore arrangement rests on USRCAT's 0.2591 against 0.2099 and 0.2055.

The check also runs the other way. Although the objective targets the USR profile, ranking by the USR profile and nothing else retrieves a partner active against the same target in 7.7% of cases at top 1 (Table 15), the lowest of the five measures other than the random baseline and a factor of four below ZAO's 30.3%. The embedding is not a proxy for the USR profile.

In the bottom two strata, the 98% figures cannot be read the same way as the ones above them: by definition, pairs close in 2D often belong to the same series, and these two strata are included to show the ceiling of the comparison.

Table 17 shows the sensitivity to where the cutoff for calling a pair 2D-distant is placed. Varying the threshold has little effect on the 3D shape agreement, but the same-target rates are sensitive to it.

Even at the strictest threshold, 0.15, the top-1 match is 21.5%, about nine times the random baseline's 2.39%.

Reverse control. Pairs close in 2D (Tanimoto ≥ 0.50) number only 116,390, or 0.15% of all pairs. Within the 1% of them with the farthest embeddings (1,163 pairs, cosine at most 0.500, mean 0.421), 96.1% are still pairs of compounds active against the same target, while USRCAT agreement remains at 0.260. The embedding does not override 2D closeness.

Limitations. Linear CKA can show no more than that the linear overlap is small. Additionally, small overlap is not a property unique to conformer sets: running the same procedure on one conformer

Table 16: 3D shape agreement and same-target fraction by pair stratum. Each value is computed per pair, unlike Table 15, which takes each molecule's single nearest neighbor.

Stratum	USRCAT	USR profile	Same target
Random (2D-distant pairs)	0.2055	0.1115	2.82%
Control matched on molecular weight, heavy-atom count, cLogP, polar surface area	0.2099	0.1103	3.15%
Distant in 2D, mid-range cosine	0.2066	0.1092	2.79%
Distant in 2D, nearest embedding	0.2591	0.2311	6.16%
Distant in 2D, nearest residual	0.2440	0.2773	10.49%
Close in 2D, far embedding	0.3866	0.2134	98.23%
Close in 2D, close embedding	0.4133	0.2342	98.88%

Table 17: Sensitivity to the 2D similarity threshold. The USRCAT column is the agreement for the retrieved pairs, with the agreement over random 2D-distant pairs in parentheses. The ZAO top-1 and residual top-1 columns are counted over all 12,406 active compounds, a different denominator from Table 15, which counts over the 12,401 molecules with computable 3D shape descriptors.

Threshold	Fraction of pairs	Random partner same target	ZAO top 1	Residual top 1	USRCAT (control)
0.15	80.07%	2.39%	21.5% (2,664)	12.4%	0.254 (0.203)
0.20	96.85%	2.93%	30.3% (3,756)	16.9%	0.259 (0.206)
0.25	99.11%	3.22%	39.2% (4,862)	21.0%	0.261 (0.206)
0.30	99.49%	3.37%	47.6% (5,900)	23.4%	0.262 (0.207)

from Uni-Mol [81], which we did not train, yields a linear CKA of 0.269 with Morgan fingerprints, comparable to ZAO's 0.298. This contrast is measured on the 2,945 molecules with embeddings from both models, a different molecule set from the one used for Table 14's 0.290. The CKA between ZAO and Uni-Mol is also low, at 0.369. This shows only that 3D encoders generally do not overlap linearly with 2D fingerprints, not that the effect comes from reading conformer sets.

F Recommended Usage and Inference Cost

This appendix sets out the procedure for readers using the released weights (Appendix F.1) and the cost of inference (Appendix F.2). For the list of released artifacts, see the Reproducibility and Release Statement.

F.1 Recommended Usage

ZAO has one mode of use, with no fine-tuning of the foundation model and no per-task adaptation. It is an addition to existing features, not a replacement, and it belongs in selection, at the stage of narrowing a broad population.

Applying the recommended usage to your own data takes five steps. (1) Standardize molecular structures (Appendix A.1). (2) Generate 10 conformers with ETKDGv3 [62, 75] and optimize with MMFF94 [30]; use the GPU implementation [56] for conformer generation (the throughput difference from the CPU implementation is reported in Appendix F.2). (3) Load the weights and pass all 10 conformers through the model in a single forward pass to extract a 2,048-dimensional representation. (4) If you already have other features, keep them and concatenate them with the representation. (5) Feed the resulting features to the downstream learners [58, 60] with the recommended default configuration (learners and feature blocks in Appendix A.4).

Table 18: Compute resources reported in this paper. Values are gathered from the sections where they originate, and nothing here is newly computed.

Stage	Hardware	Scale	Time
Pretraining	80× H100	234,000 steps	about 8,300 H100-hours
Conformer generation	about 18,000 conformers/s on 8× H200	11 billion conformers	about 1,360 GPU-hours
Same, CPU implementation	914 conformers/s on 224 vCPU	—	—
Embedding extraction (full-library inference)	40× H200	1.1 billion molecules	about 800 GPU-hours
End to end (per molecule, generation included)	1× H200	1,962 molecules	12.4 ms for 1 conformer, 20.5 ms for 10 conformers
Binding free-energy calculation (Appendix B.2)	RTX 6000 Pro	1 complex, 3 replicates	about 72 GPU-hours per complex

F.2 Inference Cost

Table 18 summarizes the compute resources this paper reports.

Even with 10 conformers instead of one, the overall cost remains within 1.65×. While the forward pass costs 4.90× more, conformer generation, which dominates the time, grows only 1.56× with the GPU implementation [56]. Additionally, single-conformer 3D models incur the conformer-generation cost. The full-library inference run over 1.1 billion molecules excludes both conformer generation and per-target scoring. At about 1,360 GPU-hours, conformer generation costs more than the 800 GPU-hours or so of embedding extraction, so generation dominates the cost.

G The Single-Conformer Pretraining Control

All the other measurements in Section 4 intervene at inference on the one set of weights we release; none of them can say whether pretraining on conformer sets is what produced the representation. To establish that and nothing else, we pretrain two models from random initialization that differ only in their pretraining input, and this appendix describes the pair.

G.1 How the Pair Was Built

Two models of the same architecture and the same 238,175,411 parameters are trained from random initialization to a common 280,000 steps. One reads real 10-conformer sets; the other reads the lowest-energy conformer of the same molecule copied into all 10 slots. Therefore, the only difference between them is how many conformers each reads. The molecules, their order, the seeds, the batch size, and the loss weights (which are fixed to static values and not adapted) are identical across both, and so are the token count and the number of forward passes.

Ten identical slots make the relative-energy target identically zero, so the one-conformer condition is degenerate in its energy signal as well as in its geometry. What separates the conditions is everything the conformer data contains, not geometry alone.

Compute limits the pair to a smaller configuration than the main model: about 143 million molecules (a seventh of the data), presented some 287 million times in total.

G.2 What the Probes Show

Both checkpoints are read with the probes of Appendix E.1: the same molecules, splits, target variables, and ridge protocol. Splitting by molecular size points the same way as splitting by flexibility: the gap between the one-conformer and 10-conformer arms in squared error on the set-specific metrics, with each metric's variance held fixed, widens monotonically across all five strata, from 0.196 at 20 heavy atoms or fewer to 0.425 at 36 or more. Table 19 lists the category means over 20,000

Table 19: The two conditions of the pair. The floor, with targets permuted, is between -0.005 and 0.000 in all categories.

Pretrained on	Read at inference	Conformer energy 4 metrics	Shape 7 metrics	Pharmacophore distance 6 metrics	Set-specific metrics 6 metrics	2D properties 13 metrics
10 conformers	10 conformers	0.415	0.843	0.818	0.768	0.911
10 conformers	mean of 10 conformers	0.296	0.868	0.835	0.622	0.919
1 conformer	1 conformer	0.241	0.616	0.707	0.426	0.856
1 conformer	mean of 10 conformers	0.254	0.729	0.764	0.469	0.876

molecules and 36 metrics. Each condition is read the way it is trained, and also as the mean over 10 conformers, which is the treatment Table 2 applies to the published encoders.

Giving the one-conformer condition that same mean-of-10 treatment from Table 2 leaves the shortfall in place. Quantities a single structure determines do rise, shape by 0.113 and pharmacophore distances by 0.057, but set-specific metrics rise by only 0.043 (0.426 to 0.469), and the gap to the 10-conformer condition narrows from 0.342 only as far as 0.299. Adding conformers and averaging them afterward does not fill in the deficit from not pretraining on conformer sets.

This is not because averaging is a weak operation; learned aggregation and plain averaging divide the work between them. On the 10-conformer condition, plain averaging is slightly ahead on quantities that are themselves means over the conformers (shape 0.868 against 0.843, pharmacophore distances 0.835 against 0.818, 2D properties 0.919 against 0.911), and learned aggregation is far ahead on quantities that are not (set-specific metrics 0.768 against 0.622, conformer energy 0.415 against 0.296). The direction of each difference matches the class of functions an averaging construction can express (Appendix E.3). In the one-conformer condition, what is missing is exactly what plain averaging cannot reach.