CHEMSTEER: CONTROLLABLE MOLECULAR GENERATION VIA LOCALLY LINEAR STEERING IN CHEMICAL LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

The Linear Representation Hypothesis (LRH) holds that concepts are encoded along linear directions in the activation space of pretrained models. It was formulated for semantic concepts in natural language, where a concept and its negation define a natural axis. Chemical properties are more complex because they are determined by valence, ring systems, scaffold symmetries and synthesizability constraints. As such, a single global direction cannot control these properties across the entire expanse of chemical space. We argue that property control in chemical language models is better described as locally linear because the representation manifold is curved, so a linear direction is valid only within a limited neighborhood, and different structural families require different directions. We test this directly and find that the angle between local steering directions increases systematically as the chemical similarity between their molecular neighborhoods decreases, across datasets and properties. This motivates ChemSteer, which uses spectral co-clustering to partition molecules into locally linear regions and latent dimensions into incoherent subspaces simultaneously, extracts a steering direction within each region, and superposes them into a single global vector with exactly zero interference. We validate on QM9, ZINC250K, MOSES and ChEMBL. ChemSteer increases QED to 0.92–0.95, reduces synthetic accessibility to ≈ 1.5, and steers effectively toward higher LogP, without retraining or fine-tuning.

1 INTRODUCTION

The computational discovery and design of novel small molecules as therapeutic candidates represent a paramount challenge in machine learning, primarily due to the astronomical scale of the chemical search space. The technical difficulty is compounded by the necessity of solving a high-dimensional, multi-objective optimization problem; a viable drug candidate must not only bind with high specificity to a biological target but also satisfy stringent constraints regarding toxicity, metabolic stability, and synthetic accessibility Hoogeboom et al. (2022). Traditional high-throughput screening methods often fail to navigate this non-convex landscape effectively, leading to high attrition rates in the drug development pipeline. Generative AI and geometric deep learning offers a promising path forward by representing molecular structures within structured subspaces.

Our work builds on two central ideas in machine learning: incoherent subspaces and linear representations. Incoherent subspaces rely on the geometric property that vectors from different subspaces are approximately orthogonal, enabling high-dimensional superposition where a single global vector can encode many local constituents as distributed representations Kleyko et al. (2022); Tropp (2004). This perspective underlies the "Superposition Hypothesis" in LLMs Elhage et al. (2022), where the residual stream acts as a high-dimensional container storing many semantic features in superposition. To preserve fidelity, models learn to embed features into nearly orthogonal (incoherent) subspaces Yu et al. (2023). From this view, transformer attention can be interpreted as dynamic projection operators that regulate information density by filtering the global state to recover relevant subspaces with high precision. The second foundation is the Linear Representation Hypothesis (LRH), which posits that semantic and functional concepts in pretrained LLMs are encoded along approximately linear directions in representation space. We expand upon the LRH in the chemical space, and motivate the need for local rather than global LRH instantiations.

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.

The LRH holds for semantic concepts in natural language, where a concept and its negation define the two ends of an axis, and linear interpolation between them is meaningful. Even within language the assumption is not universal. Cyclic concepts such as days of the week or months of the year have been found to occupy curved rather than linear structures Engels et al. (2025). For example, steering from January toward October along a straight line does not correspond to traversing the calendar, though steering between adjacent months approximates it well, in the same way that a short chord approximates an arc. Similarly, chemical properties are more complex than semantic concepts with natural opposites. They are emergent consequences of atomic composition, connectivity, ring systems, functional-groups, etc., and any change to a molecule must respect valence and chemical validity. Hence, a single direction across the expansive chemical activation space cannot increase a given property uniformly while remaining feasible. If the manifold on which these properties vary smoothly is curved, then a linear direction is a tangent approximation valid only near the point where it was estimated, and the distance over which it remains valid is set by the local curvature.

We find that the LRH holds locally in chemical datasets (at least for the three properties we examine) and degrades with distance (the steering magnitude after a certain distance). The practical consequence is that a method should partition the representation space into regions where local linear directions suffice. This is precisely the design of ChemSteer, with spectral co-clustering partitioning the molecules (which region of the manifold) and latent dimensions (which coordinates carry the property signal there) simultaneously.

Existing ML methods for small molecules define a latent space using training data for the desired properties. Our proposed method relies on the representation of small molecules in the high dimensional spaces from a pre-trained LLM. There is no need for compression; the power of the method emerges due to a hierarchy of representations: first incoherent subspaces and then linear vectors for chemical properties of interest. We show that property-specific directions defined locally within individual subspaces can be superposed into a single global direction. Leveraging these results, we introduce a post hoc method to improve target properties without retraining or fine-tuning the underlying generative model. Our thesis is that pretrained language models organize molecular representations into incoherent subspaces that admit linear steering toward targeted chemical properties, and our findings establish incoherent subspace geometry as a foundation for controllable small molecule design using pretrained language models. Our main contributions are as follows:

- We argue that the LRH, formulated for semantic concepts in language, describes chemical property control only locally, because the representation manifold is curved (Section 3.1). We test this directly, showing that the angle between local steering directions grows systematically as the chemical similarity of their molecular neighborhoods decreases (Section 3.3).
- We prove ChemSteer recovers the affinity's block partition. Under an alignment condition we test empirically, the recovered partition is also CAA-optimal.
- We validate our framework on multiple molecular objectives, and benchmark against similar approaches (Section 5). ChemSteer achieves consistent improvements on QM9, ZINC250K, MOSES, and ChEMBL, increasing QED to the 0.92–0.95 range (0.04 - 0.30 absolute improvement upon baselines), reducing synthetic accessibility (SA) to ≈ 1.5 (vs. 1.8–2.5 for baselines), and most effectively steering towards increased LogP values.

2 Related Work

Prior approaches have shown molecular property control via latent space traversal in pretrained generative models Du et al. (2022); Wei et al. (2024). ChemSpacE Du et al. (2022) frames molecular property optimization in the latent space of a variational autoencoder. ChemFlow Wei et al. (2024) models property optimization as a continuous latent vector field, allowing nonlinear and state-dependent directions. LIMO Eckmann et al. (2022), optimizes properties via latent gradient directions. These methods assume a single coherent latent space. We show that molecular representations decompose into incoherent subspaces, each with its own locally linear property structure.

The application of pre-trained LLMs to the chemical sciences marks a transition from traditional quantitative structure-property relationship (QSPR) modeling to a more flexible, linguistic understanding of molecular data. Chemical Language Models (CLMs) treat molecular representations, most commonly SMILES (Simplified Molecular Input Line Entry System) strings, as a specialized

language, using transformer architectures and self-supervised pre-training on large molecular repositories to learn the underlying "grammar" of chemical structure Chithrananda et al. (2020); Fabian et al. (2020). By pre-training on millions of unlabeled molecules, models like MolFormer develop a structured latent space where chemical similarity maps geometric proximity Ross et al. (2022). Representations can then be fine-tuned for specific tasks, often requiring much less labeled data than traditional supervised methods. Recently, multi-modal models have aligned chemical structures with natural language.Deng et al. (2025) introduce ChemLML, a modular adapter-based framework that links pretrained language models with molecular generators to enable text-guided molecule design while preserving domain-specific embedding spaces Edwards et al. (2022).

Recent advances in latent flow-based models demonstrate the utility of mapping discrete molecular entities into continuous latent spaces, where generative trajectories are modeled as smooth flows that respect the synthesizable frontier Haddad et al. (2024); Wei et al. (2024). Equivariant Diffusion Models generate molecular structures by treating atoms as particles in 3D space, ensuring that the learned representations remain invariant to global rotations and translations Gangwal & Lavecchia (2024); Hoogeboom et al. (2022). Human-in-the-loop interpretability is addressed by platforms like ChemSpacE, which provide interactive frameworks for exploring these high-dimensional manifolds and identifying optimized candidates within specific structural "neighborhoods" Du et al. (2022). Furthermore, Graph Neural Networks (GNNs) continue to serve as a foundational predictive backbone, modeling the topological connectivity of atoms to predict properties without the "brittleness" associated with string-based tokenization Stokes et al. (2020). Concurrent work by Dimitrievikj & Wells (2025) applies contrastive activation addition to molecular generation, showing that a single global steering direction can be insufficient. They address this with a state-dependent vector field that recomputes the intervention direction from nearby contrastive examples. In contrast, ChemSteer models locality in the representation space itself, enabling a static inference-time intervention that simultaneously retains multiple local directions and explains geometrically why local steering is needed in chemical space.

3 Molecular Representations and Incoherent Subspaces

We motivate ChemSteer with intuition about underlying CLM representation space geometry, and the complex nature of molecular structure and function.

3.1 Local Linearity on Curved Property Manifolds

Let $h \in \mathbb{R}^d$ be the representation of a molecule and y a chemical property. The LRH asserts a single direction v that increases y everywhere. This is the assumption for a single dataset-wide contrastive activation direction. Suppose instead that molecules populate a manifold $\mathcal{M} \subset \mathbb{R}^d$ and that y varies smoothly along it. The direction of steepest increase is then the gradient of y restricted to the tangent space at h, $v(h) = \nabla_{\mathcal{M}} y(h)$, which depends on h. A single global direction is recovered only when $\mathcal{M}$ is flat, or when curvature is small enough over the region of interest that one tangent direction serves throughout. Where curvature is appreciable, $v(h)$ rotates as h moves, and a linear step taken along a direction estimated far away both fails to increase y and leaves the manifold.

The linear approximation $y(h + \alpha v(h)) \approx y(h) + \alpha \|\nabla_{\mathcal{M}} y\|$ holds for α small relative to the local radius of curvature; beyond that the step departs from $\mathcal{M}$. Thus, the steering coefficient α is a property of the representation geometry, and it predicts that the optimum should differ across datasets whose chemical spaces have different curvature. We observe exactly this (Appendix K, Table 14). Moreover, if $v(h)$ is a tangent field on a curved manifold, then the angle between $v(h_i)$ and $v(h_j)$ should grow with the separation of h_i and h_j along $\mathcal{M}$. A monotone relationship between angular divergence and molecular-neighborhood dissimilarity cannot be produced by noise alone. Section 3.3 shows this, and Figure 1 shows the relationship for an example dataset–property pair.

3.2 Molecular Scaffolds, Symmetry, and Spectral Co-clustering

Small molecules are organized into scaffolds, groups of structurally related compounds that share a common core but may differ by functional-group substitutions, ring modifications, or stereochemical variations. This introduces a notion of molecular symmetry at the scaffold level where transformations can preserve the scaffold structure while varying peripheral features. We argue that pretrained

CLMs learn representations that reflect this scaffold-level symmetry, and that spectral co-clustering is a principled method for recovering these families from the learned representation space.

Spectral co-clustering. Spectral co-clustering Dhillon (2001) operates on the bipartite graph between molecules (rows) and latent representation dimensions (columns), partitioning both according to the spectral structure of the normalized adjacency matrix. Molecules that co-activate similar latent dimensions are grouped together, and the resulting clusters align with chemical scaffolds rather than arbitrary partitions (see Appendix B). This joint partition is what Section 3.1 calls for because the row blocks are regions of the manifold within which a single tangent direction suffices, and the column blocks are the coordinates carrying the property signal in each such region. Incoherence follows from the column partition. Each latent dimension is assigned to exactly one cluster, so the subspaces $\mathcal{S}_k = \mathrm{span}\{e_l : l \in D_k\}$ have disjoint coordinate supports and are therefore mutually orthogonal, giving worst-case operator-norm overlap $\varepsilon = \|P_i P_j\|_{\mathrm{op}} = 0$ exactly, for any dataset. This supplies exactly the incoherence required by the superposition argument in Section 4.2, Step 2.

In Appendix C we prove that spectral co-clustering exactly recovers A's block partition under mild structural conditions (Theorem 1). We show that the recovered partition is aligned with the optimal structure for locally linear steering (1). Appendix C.7 confirms that this holds for our chemical datasets. This evidence, together with the following analysis, and the results in Tables 4 and 3 confirm the optimality and effectiveness of local steering in the incoherent subspaces defined by spectral co-clustering. We now present the rest of the evidence in turn.

Spectral co-clustering yields meaningful clusters. Table 1 summarizes cluster quality across datasets. The majority of scaffolds appear in only one cluster, and spectral co-clustering produces far fewer scaffold–cluster edges across the bipartite cut than random assignment, indicating non-trivial alignment. Within-cluster Tanimoto similarity and Levenshtein edit distance confirm that clusters group chemically similar molecules (Appendix E, Table 7; Figure 7). Clusters reflect structural patterns, scaffold symmetries, and chemical property likeness, eliciting effective steering.

Table 1: Spectral co-clustering quality statistics per dataset. "Var" is the average fraction of within-cluster variance captured by the dominant eigenvector; "Cluster-Specific Scaf" is the percentage of scaffolds appearing in only one cluster; "Scaf/Cluster" is the average number of unique scaffolds per cluster; "Cut" and "Cut (R)" are edges crossing the bipartite scaffold–cluster cut for spectral co-clustering and random assignment, respectively. Full distributions are in Appendix E.

	Var % ↑	Cluster-Specific Scaf % ↑	Scaf/Cluster	Cut ↓	Cut (R)
chEMBL	42.06	94.5	53	**1777**	14791
moses	65.26	87.6	65	**4556**	14972
qm9	44.04	69.0	32	**8817**	13365
ZINC	43.30	91.7	55	**3258**	14916

We have validated that spectral co-clustering partitions molecular representations into chemically meaningful subspaces. We now develop a framework for property-guided steering. The key insight is that within each cluster, property-relevant directions can be extracted locally and then combined into a single global steering vector that recovers local control with provably bounded interference.

3.3 Local Steering Directions Vary With Chemical Structure

Section 3.1 predicts that local steering directions should diverge systematically as their molecular neighborhoods become chemically dissimilar. We test this directly, and detail our experimental setup in Appendix F. For each dataset and property we compute the pairwise angle between every pair of cluster-local CAA directions (computed in the full dense space, without coordinate restriction) and the cross-cluster average Tanimoto similarity over 20 sampled molecules per cluster. Table 2 reports the results for piecewise linearity of steering directions. The correlation is negative in every dataset–property pair, meaning that chemically similar clusters steer in more similar directions. On QM9 the effect is strongest ($r = -0.31$ to -0.39); on MOSES it is weaker but consistent ($r = -0.08$ to -0.17). The magnitude is modest with $r^2 \approx 0.10$ on the strongest pairs, so chemical similarity accounts for roughly a tenth of the variance in angular divergence. This is expected since Tanimoto

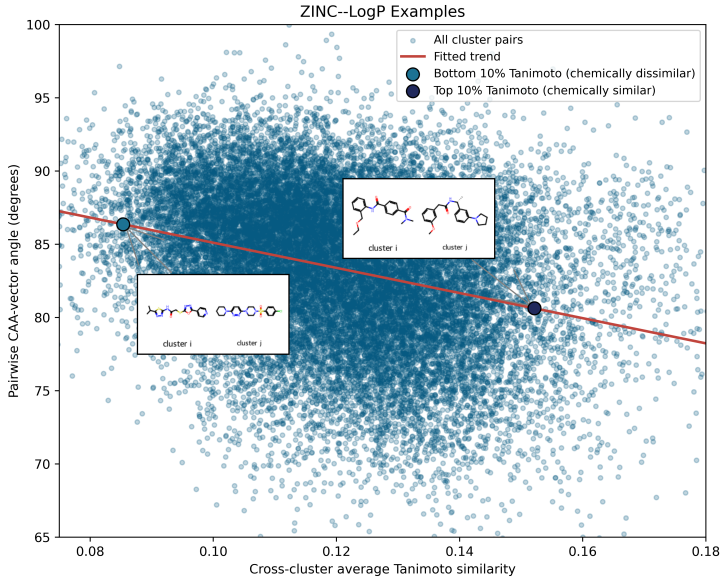


Figure 1: Local steering directions rotate with chemical position, shown for LogP on ZINC. Each point is one of the 23,436 cluster pairs, plotted by the angle between the two clusters' local CAA directions against the average cross-cluster Tanimoto similarity of their molecules. Insets show a representative molecule from each cluster for one pair drawn from the top and bottom 10% of chemical similarity. The fit falls from $88°$ to $74°$ across the observed range of chemical similarity.

Table 2: Local steering directions vary systematically with chemical structure. A negative correlation means chemically similar clusters steer in more similar directions. Mantel p is a permutation test over joint row/column labelings, which accounts for the dependence among pairs sharing a cluster. The final column residualizes both angle and Tanimoto against $\min(N_i, N_j)$, confirming the relationship is not an artifact of noisier directions in smaller clusters.

Dataset	Prop.	Pairs	Mean θ (°)	Pearson r	Mantel p	Partial r
ChEMBL	LogP	30381	80.87	-0.14	< 0.001	-0.28
	QED	28680	84.22	-0.05	< 0.001	-0.15
	SA	17578	85.86	-0.02	0.550	-0.06
MOSES	LogP	16836	81.17	-0.141	< 0.001	-0.16
	QED	17578	85.77	-0.075	0.020	-0.07
	SA	13530	82.36	-0.169	< 0.001	-0.20
QM9	LogP	31626	81.04	-0.390	< 0.001	-0.43
	QED	29161	86.72	-0.321	< 0.001	-0.34
	SA	16836	85.60	-0.311	< 0.001	-0.32
ZINC	LogP	23436	83.34	-0.30	< 0.001	-0.29
	QED	30381	87.03	-0.27	< 0.001	-0.27
	SA	6670	83.99	-0.05	0.390	0.03

similarity is a rough proxy for position on the representation manifold, cluster directions are estimated from tens of molecules in 4096 dimensions, and much of the residual variance is estimation noise. The important finding is the monotonic relationship since estimation noise perturbs directions in random directions and cannot produce this consistent dependence on chemical similarity.

Figure 1 shows the result for LogP on ZINC. Each of the cluster pairs is plotted by the angle between its two local steering directions against the Tanimoto similarity of the molecules. The fit falls from $88°$ for the least similar pairs to $74°$ for the most similar. Clusters containing similar molecules steer in more nearly parallel directions. Table 2 reports the same analysis across all datasets and properties. We also see in this geometrically in Appendix S where we compare the composite ChemSteer

vector against a single global CAA direction computed over the whole dataset. The cosine similarities fall between 0.44 and 0.68 across datasets and properties, which is what a piecewise-tangent approximation should look like relative to a single chord across the same curved surface.

The two null results are, explicably, both on SA. Across datasets, the SA partial correlation tracks scaffold purity (Table 1) monotonically: QM9 (69.0% of scaffolds in a single cluster, $r_{\text{partial}} = -0.32$), MOSES ($87.6\%$, -0.20), ZINC (91.7%, $+0.03$) and ChEMBL (94.5%, -0.06). No such ordering holds for LogP or QED. This is expected because SA is computed from scaffold-level features, so it is close to a function of the scaffold alone, whereas LogP varies with substituents within a fixed scaffold and QED is a composite of six descriptors. A scaffold-coherent cluster is therefore a more SA-homogeneous cluster. Steering remains effective because even small directions in embedding space have nuanced control. Even though here the directions do not vary because chemical similarity remains high, this does not impact the steering capacity of the directions.

4 CHEMSTEER

ChemSteer operates in a post-hoc manner on pretrained models without additional model training or fine-tuning. Figure 2 illustrates our pipeline, and we detail each step in Section 4.2.

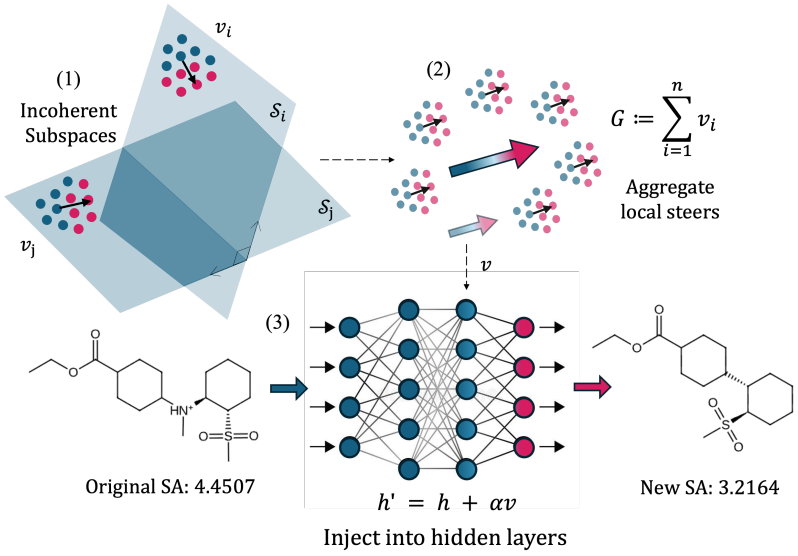


Figure 2: **ChemSteer pipeline.** (1) Local steering vector extraction for each incoherent subspace identified via spectral co-clustering. (2) Superposition of local steering vectors to construct property-specific global steering vector. (3) Steered molecular generation via global steering vector injection into the hidden-state activations of SmileyLlama at layer 24.

4.1 MODEL, DATASETS, AND MOLECULAR REPRESENTATIONS

We evaluate ChemSteer on SmileyLlama, a LLaMA-3 variant pretrained on SMILES strings for molecular generation Cavanagh et al. (2026), and on GPT-2 as a secondary baseline. We use standard chemical benchmarking datasets: ZINC250K Irwin & Shoichet (2004), MOSES Polykovskiy et al. (2020), QM9 Ramakrishnan et al. (2014), and ChEMBL Zdrazil et al. (2024). ChEMBL and MOSES contain diverse, real small molecules with varying properties, designed to span broad chemical space. ZINC250K is large and structurally diverse, whereas QM9 has limited structural diversity and covers only a subset of chemical space. Each molecule is represented as a SMILES string and encoded by SmileyLlama into a 4096-dimensional latent vector $z \in \mathbb{R}^{4096}$, obtained by averaging token-level hidden states from a designated layer. We extract representations from layer 24 since steering is most effective in intermediate transformer layers, which encode rich feature representations, while earlier layers capture low-level information and later layers offer limited control Rogers et al. (2020); Subramani et al. (2022). We provide a layer ablation in Appendix H

4.2 CHEMSTEER PIPELINE

ChemSteer Step (1): Incoherent Subspace Formation We apply spectral co-clustering Dhillon (2001) to the molecular representations to identify incoherent subspaces $\{S_k\}_{k=1}^n \subset \mathbb{R}^d$. Co-clustering captures the bipartite structure between molecules and latent dimensions, yielding localized, chemically coherent regions of latent space. Dimensions irrelevant to each cluster are zeroed out, giving disjoint coordinate supports with exactly $\epsilon = 0$.

Let $\mathcal{S}_1, \ldots, \mathcal{S}_n$ be subspaces of $\mathbb{R}^d$, and let $P_k \in \mathbb{R}^{d\times d}$ denote the orthogonal projector onto $\mathcal{S}_k$. Assume the following *operator-norm overlap* (subspace incoherence) condition: $\exists \epsilon \in [0, 1)$ such that

$$\forall i \neq k, \qquad \|P_k P_i\|_{\text{op}} \leq \epsilon, \text{ where } \|A\|_{\text{op}} = \sup_{\|x\|=1} \|Ax\|. \tag{1}$$

This bounds the maximum energy leakage of projection of a vector from one subspace to another, and is equivalent to requiring large principal angles between subspaces. In high-dimensional spaces, incoherence allows distinct features to coexist in superposition while being recoverable via projection. Spectral co-clustering yields $\varepsilon = 0$ by construction (Section 3.2). As interference between subspaces grows, steering vectors cannot accurately exploit the symmetries captured by incoherent subspaces. At $\epsilon = 0$, the bound in Equation 7 is tight: $P_k G = v_k$ exactly. For $\epsilon > 0$ it characterizes graceful degradation (Appendix L, Figure 8). Moreover, restricting each local vector to its own coordinate block preserves its direction. The cosine between v_k and the component of the corresponding direction in $\mathcal{S}_k$ is 0.93–0.98 across all datasets and both models (Appendix O, Tables 19 and 10), and drops under random cluster assignment (Appendix H.3).

ChemSteer Step (2): Global steering vector from local directions. For each subspace S_k and target property, we extract a local steering vector via contrastive activation difference $v_k := \frac{\mu_k^+ - \mu_k^-}{\|\mu_k^+ - \mu_k^-\|}$, from (Panickssery et al., 2024; Subramani et al., 2022). μ_k^+, μ_k^- are mean representations of the top and bottom quartile molecules (by property value) within S_k. For each $i \in \{1, \ldots, n\}$, pick a unit vector $v_i \in \mathcal{S}_i$, and define

$$G := \sum_{i=1}^{n} v_i, \qquad \hat{G} := \frac{G}{\|G\|}. \tag{2}$$

Fix index k. Consider projection of $\hat{G}$ onto $\mathcal{S}_k$: $w_k := P_k \hat{G}$. Define the *normalized* projected vector $\tilde{w}_k := \frac{w_k}{\|w_k\|} = \frac{P_k \hat{G}}{\|P_k \hat{G}\|} = \frac{P_k G}{\|P_k G\|}$, with disjoint supports, $P_k G = v_k$ exactly. Appendix M bounds the degradation as $sim \geq 1 - O(n-1)^2\epsilon^2$ when supports overlap.

ChemSteer Step (3): Steered Generation

We inject the global steering vector into SmileyLlama's hidden states at layer 24 (and GPT2 layer 12): $h' = h + \alpha v$, where $\alpha > 0$ steers toward higher property values, and $\alpha < 0$ steer toward lower values. Property thresholds and α ablations are in Appendix I, K. The coordinate blocks of G are equal to the corresponding local vectors by construction in Step 2, so the αv_k correspond directly to the coordinates associated with property/cluster k, without requiring an additional projection or dictionary lookup step. Steering vectors are unit-normalized at construction and the composed G is renormalized before applying α (a multiple of the model's activation-magnitude scale at the injection layer; Appendix K), so the perturbation is scaled relative to the residual stream's norm. To encourage structural similarity between the original and steered molecules, we use a prompt-based conditional generation scheme. Given an original molecule's SMILES string, `orig_SMILES`, we construct the prompt: `Original SMILES: {orig_SMILES} \n Modified SMILES:`

Steering is applied only during the generation of the modified portion, ensuring that the intervention affects molecular generation rather than prompt interpretation. This follows the design of SmileyLlama Cavanagh et al. (2026), which frames molecule generation as a prompt-completion task.

5 EXPERIMENTAL RESULTS

We evaluate ChemSteer using SmileyLlama and GPT2 on each dataset. We compare against ChemSpacE from Du et al. (2022), Gradient Flow from Eckmann et al. (2022), and all variants

Table 3: Comparison of top-100 generated molecules across molecular generation methods on ChEMBL and MOSES. ChemSteer improves LogP, QED, and SA, outperforming all prior approaches. Thus, ChemSteer performs well on real-world pharmaceutically-relevant datasets.

Model	ChEMBL			MOSES		
	logP ↑	QED ↑	SA ↓	logP ↑	QED ↑	SA ↓
Random	2.69 ± 0.42	0.85 ± 0.02	2.01 ± 0.37	1.29 ± 0.44	0.91 ± 0.02	2.33 ± 0.35
Chemspace	3.37 ± 0.63	0.85 ± 0.02	1.95 ± 0.24	2.50 ± 0.29	0.91 ± 0.01	1.80 ± 0.22
Gradient Flow	3.22 ± 0.59	0.85 ± 0.02	1.98 ± 0.23	2.72 ± 0.39	0.92 ± 0.01	1.84 ± 0.16
Wave (spv)	3.46 ± 0.61	0.84 ± 0.02	1.98 ± 0.23	2.54 ± 0.34	0.91 ± 0.01	1.83 ± 0.21
Wave (unsup)	3.40 ± 0.61	0.85 ± 0.02	1.98 ± 0.23	2.39 ± 0.32	0.91 ± 0.01	1.83 ± 0.21
HJ (spv)	3.55 ± 0.70	0.87 ± 0.02	1.92 ± 0.20	2.43 ± 0.35	0.91 ± 0.01	1.83 ± 0.21
HJ (unsup)	3.40 ± 0.54	0.80 ± 0.03	1.78 ± 0.34	2.42 ± 0.37	0.91 ± 0.01	1.82 ± 0.16
FP	3.28 ± 0.63	0.85 ± 0.02	1.94 ± 0.29	2.66 ± 0.30	0.92 ± 0.01	1.80 ± 0.15
SmileyLlama (no steer)	3.62 ± 0.87	0.87 ± 0.03	1.86 ± 0.21	3.82 ± 0.58	0.88 ± 0.03	1.93 ± 0.13
SmileyLlama (naive steer)	5.02 ± 0.61	0.86 ± 0.03	1.90 ± 0.13	3.95 ± 0.56	0.90 ± 0.02	1.92 ± 0.12
ChemSteer	$\mathbf{5.24 \pm 1.19}$	$\mathbf{0.94 \pm 0.10}$	$\mathbf{1.55 \pm 0.18}$	$\mathbf{6.68 \pm 0.64}$	$\mathbf{0.95 \pm 0.00}$	$\mathbf{1.50 \pm 0.08}$

Table 4: Comparison of top-100 generated molecules from all methods on ZINC250K and QM9. For ZINC250K, ChemSteer improves LogP, QED, and SA, outperforming prior approaches, showing the success of our method on diverse datasets with complex chemical spaces. Even in a low-complexity, non-diverse chemical space of QM9, ChemSteer achieves improvement and effective control.

Model	ZINC250K			QM9		
	logP ↑	QED ↑	SA ↓	logP ↑	QED ↑	SA ↓
Random	1.06 ± 0.42	0.85 ± 0.02	2.53 ± 0.34	1.17 ± 0.34	0.63 ± 0.02	1.63 ± 0.20
Chemspace	1.61 ± 0.44	0.86 ± 0.02	2.15 ± 0.33	2.05 ± 0.40	0.65 ± 0.01	1.61 ± 0.14
Gradient Flow	1.95 ± 0.52	0.86 ± 0.02	2.13 ± 0.28	2.47 ± 0.47	0.65 ± 0.01	1.52 ± 0.15
Wave (spv)	1.61 ± 0.45	0.86 ± 0.02	2.15 ± 0.32	1.88 ± 0.36	0.64 ± 0.02	1.62 ± 0.17
Wave (unsup)	1.55 ± 0.53	0.87 ± 0.02	2.06 ± 0.36	2.02 ± 0.47	0.65 ± 0.01	1.89 ± 0.29
HJ (spv)	1.65 ± 0.54	0.86 ± 0.02	2.15 ± 0.32	1.88 ± 0.36	0.65 ± 0.01	1.58 ± 0.16
HJ (unsup)	1.65 ± 0.44	0.86 ± 0.02	2.19 ± 0.33	1.90 ± 0.40	0.64 ± 0.02	$\mathbf{1.41 \pm 0.18}$
FP	1.92 ± 0.51	0.86 ± 0.02	2.09 ± 0.33	2.40 ± 0.43	0.66 ± 0.01	1.48 ± 0.15
SmileyLlama (no steer)	2.87 ± 0.42	0.88 ± 0.04	2.05 ± 0.27	1.50 ± 0.45	0.81 ± 0.02	2.63 ± 0.31
SmileyLlama (naive steer)	4.34 ± 0.67	0.89 ± 0.03	1.94 ± 0.20	0.54 ± 0.66	0.58 ± 0.04	2.95 ± 0.50
ChemSteer	$\mathbf{6.82 \pm 0.71}$	$\mathbf{0.94 \pm 0.00}$	$\mathbf{1.54 \pm 0.10}$	$\mathbf{2.70 \pm 0.16}$	$\mathbf{0.92 \pm 0.01}$	1.61 ± 0.17

of ChemFlow in (Wei et al., 2024) (Random, Wave (supervised), Wave (unsupervised), Hamilton-Jacobi (HJ) (supervised), HJ (unsupervised), and Fokker Planc (FP)). Performance is measured in terms of the method's ability to increase LogP and QED, and to decrease SA. For each dataset-molecular property pair, we report the mean and standard deviation of the 100 best generated molecules. We report the overall averages for the entire generated set in Appendix J. Table 14 contains the α coefficients for each steering vector and dataset, determined via a sweep on α across datasets and properties. Appendix G elaborates on the performance of ChemSteer over the naive global vector and no steering baselines. Appendix Q shows that ChemSteer outperforms Gemini.

ChemSteer Outperforms on Real-World Drug-Like Molecules. Tables 3 summarizes the results on ChEMBL and MOSES. These are two pharmaceutical datasets with diverse scaffolds and property distributions which attempt to span the vast chemical space, crafted for benchmarking generative models. ChemSteer consistently outperforms all baselines, achieving substantially improved LogP, QED and SA values. These results show that incoherent subspace steering via ChemSteer is the most effective, highest performing framework for real-world molecule generation.

ChemSteer Outperforms on High- and Low-Complexity Datasets Results on ZINC250K and QM9 are in Table 4. The ZINC250K dataset is very structurally diverse. ChemSteer again achieves strong improvements across all metrics. It reduces SA by a large margin, and increases QED and LogP. On QM9, ChemSteer achieves the largest improvement in LogP, while maintaining competitive performance on SA and very effective QED control. The QM9 outcome also constitutes a cross-experiment consistency check on the local-LRH thesis: it exhibits both the largest angular-divergence signal *and* the most catastrophic naive-global-steering failure. Under our framing, these are the same phenomenon: the dataset with the greatest local manifold curvature is the dataset

Table 5: ChemSteer has high generation quality across datasets (%). Full table in Appendix P.

	ChemSteer			ChemSpace			LIMO/ ChemFlow		
Dataset	Valid.	Unique.	Novel.	Valid.	Unique.	Novel.	Valid.	Unique.	Novel.
ChEMBL	**100**	**97.8**	**100**	100	56.6	94.2	100	97.8	100
MOSES	**93.2**	93.9	**100**	82.4	34.8	79.7	100*	**98.5**	100
QM9	**96.4**	**96.3**	**100**	91.6	24.2	88.7	100*	93.4	100
ZINC250K	**87.1**	**99.5**	**100**	71.3	15.8	71.3	100*	98.8	100

where a single global chord approximation to the tangent field fails most severely. More broadly, ChemSteer's piecewise composition recovers meaningful steering across both structurally-diverse (ZINC250K, ChEMBL, MOSES) and constrained (QM9) chemical spaces, whereas naive global steering degrades most severely precisely where local curvature is greatest.

Validity, Uniqueness, and Novelty of Generated Molecules. Table 5 reports validity (RDKit-parsable), uniqueness, and novelty. ChemSteer achieves strong performance across all metrics, with high validity (100.0% on ChEMBL, 96.4% on QM9, 93.2% on MOSES, 87.1% on ZINC250K), while ChemSpace degrades notably (82.4% on MOSES, 71.3% on ZINC250K), indicating instability. ChemFlow variants report uniformly 100% validity by construction via SELFIES decoding, which guarantees syntactically valid SMILES but not chemical plausibility, as invalid or implausible structures may still occur. ChemSteer maintains high uniqueness (97.8% on ChEMBL, 99.5% on ZINC250K), whereas ChemSpace exhibits severe collapse (56.6%, 15.8%), with similar trends on MOSES and QM9. ChemSteer's 100% novelty across datasets matches or exceeds all baselines.

6 Discussion, Limitations and Future Work

Discussion. ChemSteer is a molecular property control framework that performs piecewise linear steering as a superposition of local directions. It outperforms naive global steering and prior latent-space methods (Tables 3 and 4, Appendix G). The steering coefficient has a dataset-dependent optimum (Table 14), which is the distance over which the linear approximation stays on-manifold; and the composite vector diverges from the global one by a consistent margin (Appendix S). The LRH describes chemical property control only within limited neighborhoods due to manifold curvature, and we validate that local steering directions diverge systematically as their molecular neighborhoods become chemically dissimilar. The contribution of ChemSteer is to identify the regions over which it works and to compose the resulting directions without interference, recovering strong control of LogP, QED and SA without retraining, fine-tuning, or auxiliary property predictors.

Theoretical structure of the claim. Theorem 1 is unconditional. It states that given a nonnegative affinity with block structure, spectral co-clustering recovers the blocks. Our empirical validations in Section 3.3 and Appendix C.7 constitute evidence that aligned block LRH holds approximately in CLMs on the four benchmark datasets we study. On domains where alignment fails, Theorem 1's recovery guarantee still holds but the optimality claim may not. Identifying such domains, for instance, properties whose linear-response structure decouples from the representation's clustering structure, is an important direction for future work.

Limitations. ChemSteer assumes the emergence of approximately incoherent subspaces in pre-trained chemical language models, but does not explain their origin. While empirically observed and consistent with the LRH, a principled understanding of data-driven or optimization factors that produce incoherent chemical subspaces remains an open problem. In addition, we do not fully characterize how incoherence varies across layers. For highly entangled or structurally constrained properties, the subspace decomposition assumption may break down, limiting the effectiveness of linear steering. More broadly, ChemSteer relies on the quality of the learned representation space; therefore, weakly expressive or poorly trained models may yield less effective steering.

Future Work. A key direction is multi-objective steering (see Appendix R). Incorporating structure-aware or docking constraints is another extension. More broadly, developing a principled theory for the emergence of symmetry-induced structure, the aligned block LRH, and understanding how transformers learn such structure is an important direction for future research.

AI USE STATEMENT

We used generative AI tools minimally to help with basic tasks which do not involve intellectual contributions. We wrote all original text ourselves and occasionally used AI to edit prose for concision and clarity. We also used AI to help with the implementation of some supporting analysis and figure code. We did not use AI for any implementation of our main method. All experiments were designed and run by the authors, and their outputs were inspected by the authors, All numerical claims were checked against the saved outputs. The authors take full responsibility for the content, claims, code, figures, and artifacts of this work.

ETHICS STATEMENT

The goal of our work is to advance the field of machine learning, particularly in the area of controlling molecular generation. Our framework is intended to improve the performance and efficiency of molecular generation, a substantial positive contribution to applications such as drug discover, chemical research, and materials design. The ability to steer molecular properties has incredibly positive potential, especially for chemical development; however, this naturally means that such methods could be misused to accelerate the discovery of harmful compounds if used irresponsibly. We believe that the ethical and societal implications of this work are consistent with those commonly encountered in machine learning research for scientific discovery, and we do not identify any specific risks that warrant additional discussion beyond standard considerations (i.e. responsible use).

REPRODUCIBILITY STATEMENT

All datasets used are available for download from the original source. All code needed to reproduce experiments and generate all figures within the paper is available in the supplementary material, and experimental setup information is outlined in the README. Follow the guidelines on the supplementary material README pages to reproduce the performance results, as well as the plots presented throughout the paper.

REFERENCES

Joseph M. Cavanagh, Kunyang Sun, Andrew Gritsevskiy, Dorian Bagni, Yingze Wang, Thomas D. Bannister, and Teresa Head-Gordon. Smileyllama: Modifying large language models for directed chemical space exploration, 2026. URL https://arxiv.org/abs/2409.02231.

Seyone Chithrananda, Gabriel Grand, and Bharath Ramsundar. Chemberta: Large-scale self-supervised pretraining for molecular property prediction, 2020. URL https://arxiv.org/abs/2010.09885.

Yifan Deng, Spencer S. Ericksen, and Anthony Gitter. Chemical language model linker: Blending text and molecules with modular adapters. *Journal of Chemical Information and Modeling*, 65 (17):8944–8956, August 2025. ISSN 1549-960X. doi: 10.1021/acs.jcim.5c00853. URL http://dx.doi.org/10.1021/acs.jcim.5c00853.

Inderjit S. Dhillon. Co-clustering documents and words using bipartite spectral graph partitioning. In *Knowledge Discovery and Data Mining*, 2001. URL https://api.semanticscholar.org/CorpusID:11847258.

Aleksandar Dimitrievikj and Jude Wells. Steering vector fields for property-controlled molecular generation with chemical language models. *bioRxiv*, 2025. doi: 10.1101/2025.09.24.678080. URL https://www.biorxiv.org/content/early/2025/09/25/2025.09.24.678080.

Yuanqi Du, Xian Liu, Nilay Shah, Shengchao Liu, Jieyu Zhang, and Bolei Zhou. Chemspace: Interpretable and interactive chemical space exploration. *ChemRxiv*, 2022(1005), 2022. doi: 10.26434/chemrxiv-2022-x49mh-v3. URL https://chemrxiv.org/doi/abs/10.26434/chemrxiv-2022-x49mh-v3.

Peter Eckmann, Kunyang Sun, Bo Zhao, Mudong Feng, Michael Gilson, and Rose Yu. LIMO: Latent inceptionism for targeted molecule generation. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 5777–5792. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/eckmann22a.html.

Carl Edwards, Tuan Lai, Kevin Ros, Garrett Honke, Kyunghyun Cho, and Heng Ji. Translation between molecules and natural language, 2022. URL https://arxiv.org/abs/2204.11817.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, Roger Grosse, Sam McCandlish, Jared Kaplan, Dario Amodei, Martin Wattenberg, and Christopher Olah. Toy models of superposition, 2022. URL https://arxiv.org/abs/2209.10652.

Joshua Engels, Eric J. Michaud, Isaac Liao, Wes Gurnee, and Max Tegmark. Not all language model features are one-dimensionally linear, 2025. URL https://arxiv.org/abs/2405.14860.

Benedek Fabian, Thomas Edlich, Héléna Gaspar, Marwin Segler, Joshua Meyers, Marco Fiscato, and Mohamed Ahmed. Molecular representation learning with language models and domain-relevant auxiliary tasks, 2020. URL https://arxiv.org/abs/2011.13230.

Amit Gangwal and Antonio Lavecchia. Unleashing the power of generative ai in drug discovery. *Drug Discovery Today*, 29(6):103992, 2024. ISSN 1359-6446. doi: https://doi.org/10.1016/j.drudis.2024.103992. URL https://www.sciencedirect.com/science/article/pii/S135964462400117X.

Ragy Haddad, Eleni Litsa, Zhen Liu, Xin Yu, Daniel Burkhardt, and Govinda Bhisetti. Targeted molecular generation with latent reinforcement learning, 12 2024.

Emiel Hoogeboom, Victor Garcia Satorras, Clément Vignac, and Max Welling. Equivariant diffusion for molecule generation in 3d, 2022. URL https://arxiv.org/abs/2203.17003.

John J. Irwin and Brian K. Shoichet. Zinc a free database of commercially available compounds for virtual screening. *Journal of Chemical Information and Modeling*, 45(1):177–182, 12 2004. ISSN 1549-9596. doi: 10.1021/ci049714+. URL https://doi.org/10.1021/ci049714+.

Denis Kleyko, Dmitri A. Rachkovskij, Evgeny Osipov, and Abbas Rahimi. A survey on hyperdimensional computing aka vector symbolic architectures, part i: Models and data transformations. *ACM Computing Surveys*, 55(6):1–40, December 2022. ISSN 1557-7341. doi: 10.1145/3538531. URL http://dx.doi.org/10.1145/3538531.

Amit Kumar and Ravindran Kannan. Clustering with spectral norm and the k-means algorithm. In *Proceedings of the 2010 IEEE 51st Annual Symposium on Foundations of Computer Science*, FOCS '10, pp. 299–308, USA, 2010. IEEE Computer Society. ISBN 9780769542447. doi: 10.1109/FOCS.2010.35. URL https://doi.org/10.1109/FOCS.2010.35.

Nina Panickssery, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Matt Turner. Steering llama 2 via contrastive activation addition, 2024. URL https://arxiv.org/abs/2312.06681.

Daniil Polykovskiy, Alexander Zhebrak, Benjamin Sanchez-Lengeling, Sergey Golovanov, Oktai Tatanov, Stanislav Belyaev, Rauf Kurbanov, Aleksey Artamonov, Vladimir Aladinskiy, Mark Veselov, Artur Kadurin, Simon Johansson, Hongming Chen, Sergey Nikolenko, Alan Aspuru-Guzik, and Alex Zhavoronkov. Molecular sets (moses): A benchmarking platform for molecular generation models, 2020. URL https://arxiv.org/abs/1811.12823.

Raghunathan Ramakrishnan, Pavlo O. Dral, Matthias Rupp, and O. Anatole von Lilienfeld. Quantum chemistry structures and properties of 134 kilo molecules. *Scientific Data*, 1, 2014. URL https://api.semanticscholar.org/CorpusID:15367821.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. A primer in bertology: What we know about how bert works, 2020. URL https://arxiv.org/abs/2002.12327.

Jerret Ross, Brian Belgodere, Vijil Chenthamarakshan, Inkit Padhi, Youssef Mroueh, and Payel Das. Large-scale chemical language representations capture molecular structure and properties, 2022. URL https://arxiv.org/abs/2106.09553.

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. ISSN 0092-8674. doi: https://doi.org/10.1016/j.cell.2020.01.021. URL https://www.sciencedirect.com/science/article/pii/S0092867420301021.

Nishant Subramani, Nivedita Suresh, and Matthew Peters. Extracting latent steering vectors from pretrained language models. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 566–581, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-acl.48. URL https://aclanthology.org/2022.findings-acl.48/.

Daniel Tan, David Chanin, Aengus Lynch, Dimitrios Kanoulas, Brooks Paige, Adria Garriga-Alonso, and Robert Kirk. Analyzing the generalization and reliability of steering vectors, 2025. URL https://arxiv.org/abs/2407.12404.

Joel A. Tropp. Greed is good: algorithmic results for sparse approximation. *IEEE Transactions on Information Theory*, 50:2231–2242, 2004. URL https://api.semanticscholar.org/CorpusID:675692.

Per-Åke Wedin. Perturbation bounds in connection with singular value decomposition. *BIT Numerical Mathematics*, 12:99–111, 1972. URL https://api.semanticscholar.org/CorpusID:122277492.

Guanghao Wei, Yining Huang, Chenru Duan, Yue Song, and Yuanqi Du. Navigating chemical space with latent flows, 2024. URL https://arxiv.org/abs/2405.03987.

Yaodong Yu, Sam Buchanan, Druv Pai, Tianzhe Chu, Ziyang Wu, Shengbang Tong, Benjamin D. Haeffele, and Yi Ma. White-box transformers via sparse rate reduction, 2023. URL https://arxiv.org/abs/2306.01129.

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, 01 2024. ISSN 0305-1048. doi: 10.1093/nar/gkad1004. URL https://doi.org/10.1093/nar/gkad1004.

TABLE OF CONTENTS

A Brief Description of Molecular Properties 13
B Spectral Co-clustering Overview 13
C Optimal Subspace Selection for Property-Linear Steering 14
D Synthetic Validation of Theorem 1 20
E Spectral Co-clustering Subspace Analysis 21
F Section 3.3 Experimental Setup 24
G Comparison with Naive Global Steering 25
H GPT2 and Layer Experiments 27

I Steering Vector Parameters 27
J Overall Performance Averages 28
K Alpha Ablation 29
L General Incoherent Subspace Considerations 29
M Bounded the leakage into S_k 30
N Contrastive Activation Difference Equation 31
O Steering Angles and Model Alignment 31
P Full Results on Validity, Uniqueness, and Novelty 32
Q Gemini 3 Case Study 32
R Multi-Objective Steering 34
S ChemSteer Captures Locally-Structured Geometry 36
T Computer Resources 36

A Brief Description of Molecular Properties

We use several key molecular properties that are crucial metrics in drug discovery, and are commonly used to evaluate chemical datasets and guide generative models. These properties capture various aspects of molecular structure, drug-likeness, and synthetic feasibility. We use the RDKit to directly calculate these metrics from the SMILES string representation of the molecules Below we provide a brief description of each:

- **LogP**: The octanol–water partition coefficient, which measures the hydrophobicity of a molecule. Higher values indicate greater lipophilicity, while lower values correspond to more hydrophilic molecules. LogP is widely used to estimate drug absorption and permeability.
- **Quantitative Estimate of Drug-likeness (QED)**: A composite score between 0 and 1 that quantifies the "drug-likeness" of a molecule based on multiple physicochemical properties, including molecular weight, LogP, hydrogen bond donors/acceptors, and structural alerts. Higher QED values indicate molecules more likely to be pharmacologically viable.
- **Synthetic Accessibility (SA)**: A score estimating the ease of chemical synthesis, typically ranging from 0 (easy to synthesize) to 10 (difficult). It combines factors such as molecular complexity, ring systems, and fragment rarity. Lower SA scores are preferred for practical synthesis of novel molecules.

B Spectral Co-clustering Overview

In our implementation, we directly use the SpectralCoclustering functionality in sci-kit learn. In order to understand the mathematics behind the clustering algorithm, and to fully understand the meaningfulness of the resultant clusters, we provide an overview of the algorithm in this section.

Spectral co-clustering is a graph-based bi-clustering method that jointly clusters rows and columns of a data matrix by leveraging the spectral structure of an associated bipartite graph. In our setting, the rows correspond to individual molecules (samples) and the columns correspond to latent representation dimensions (features).

B.1 Bipartite Graph Construction

Let $A \in \mathbb{R}^{m \times d}$ denote the non-negative data matrix whose ith row $A_{i,:}$ is the d-dimensional steering vector for sample i. To diminish the influence of vector magnitude, we row-normalize.

From A we construct a weighted bipartite graph $B = (U, V, E)$ where

- $U = \{u_1, \ldots, u_m\}$ represents the samples,

- $V = \{v_1, \ldots, v_d\}$ represents the features,
- E contains an edge between sample node u_i and feature node v_j with weight A_{ij}.

This bipartite graph encodes co-occurrence relationships where strong connections between a sample and a feature reflect high activation of that dimension for that sample.

B.2 Spectral Decomposition and Embedding

Spectral co-clustering uses the spectral properties of the graph to reveal cluster structure. The normalized bipartite adjacency is typically represented as:

$$\tilde{B} = D_u^{-1/2} B D_v^{-1/2},$$

where D_u and D_v are diagonal degree matrices for sample and feature nodes, respectively. A singular value decomposition (SVD) of $\tilde{B}$ yields:

$$\tilde{B} = U \Sigma V^\top,$$

where the singular vectors in U (rows) and V (columns) capture global connection patterns between samples and features.

The leading singular vectors provide low-dimensional spectral embeddings for both samples and features. These embeddings locate nodes in the "co-cluster space" where proximity reflects association in the original bipartite graph.

B.3 Clustering in Spectral Space

Once spectral embeddings are obtained, the k-means clustering algorithm is applied to the embeddings of both samples and features. The result is a set of cluster assignments for nodes in U and V, which correspond to sample clusters and feature clusters, respectively. These assignments define subspaces $\{S_k\}_{k=1}^n$ with each subspace supported on a subset of latent dimensions.

A key theoretical property of spectral co-clustering is that the basis vectors defining these subspaces are orthogonal by construction since the spectral embeddings are derived from an orthogonal decomposition (SVD) of the normalized bipartite graph. As a result, when subspace overlap is measured via the operator-norm overlap metric $\epsilon = \|P_i P_j\|$ (see Equation 1), the worst-case overlap can be near zero regardless of the dataset.

C Optimal Subspace Selection for Property-Linear Steering

Rather than assuming that the incoherent subspaces emerge as a geometric property of the representation space, we pose the clustering step in ChemSteer as an optimization problem with the goal of finding a joint partition of the molecules and latent dimensions under which a target chemical property is modeled in a piecewise linear manner. In this way, the subspace partitioning is an instance of a general combinatorial optimization problem, with spectral co-clustering emerging (under assumptions) as an exact solver of the unsupervised case.

C.1 Problem Formalization

Let $H \in \mathbb{R}^{N \times d}$ be the representation matrix, with row $h_i^\top = H_{i,:}$ the d-dimensional latent representation of molecule $i \in [N]$, and let $y \in \mathbb{R}^N$ be the property vector (LogP, QED, or SA) with y_i the target value for molecule i. Fix the number of subspaces $k \in \mathbb{N}$ as a hyperparameter. For $D \subseteq [d]$ and $h \in \mathbb{R}^d$, write $h|_D \in \mathbb{R}^{|D|}$ for the coordinate restriction of h to D.

For a molecule-to-subspace assignment $\pi_C : [N] \to [k]$ (inducing row partition $C_j = \pi_C^{-1}(j)$) and a dimension-to-subspace assignment $\pi_D : [d] \to [k]$ (inducing column partition $D_j = \pi_D^{-1}(j)$), let $S_j = \mathrm{span}\{e_l : l \in D_j\}$; disjoint D_j give $S_j \cap S_{j'} = \{0\}$ and hence subspace incoherence $\epsilon = \|P_{S_j} P_{S_{j'}}\|_{op} = 0$, exactly as in the ChemSteer construction.

Local objective. Within cluster j, let $C_j^+, C_j^- \subseteq C_j$ be the top- and bottom-quartile molecules by property value, with class means $\mu_j^+ = \frac{1}{|C_j^+|}\sum_{i\in C_j^+} h_i|_{D_j}$ and $\mu_j^- = \frac{1}{|C_j^-|}\sum_{i\in C_j^-} h_i|_{D_j}$. For a unit vector $v \in S_j$, the projected class-mean gap is $v^\top(\mu_j^+ - \mu_j^-)$. By Cauchy–Schwarz this is maximized by

$$v_j^\star \;=\; \frac{\mu_j^+ - \mu_j^-}{\|\mu_j^+ - \mu_j^-\|}, \qquad \text{attaining} \qquad v_j^{\star\top}(\mu_j^+ - \mu_j^-) \;=\; \big\|\mu_j^+ - \mu_j^-\big\|_2, \tag{3}$$

which is exactly the CAA direction of Panickssery et al. (2024).

Optimization problem. Substituting the closed form of Equation 3 into the per-cluster objective and summing gives a purely combinatorial outer problem over partitions:

$$\max_{\pi_C,\,\pi_D}\; \mathrm{Sep}(\pi_C,\pi_D) := \sum_{j=1}^{k} \big\|(\mu_j^+ - \mu_j^-)\big\|_2, \tag{4}$$

subject to $\biguplus_j C_j = [N]$, $\biguplus_j D_j = [d]$ (disjoint partitions), and non-degeneracy $|C_j| \geq c_{\min}$, $|D_j| \geq d_{\min}$ (with C_j^+, C_j^- each required to meet a minimum class size $c_{\min}^{\pm}$ for the local objective to be well-defined). The optimization problem in Equation equation 4 is to partition molecules and latent dimensions such that the maximal total local class-separation is obtained by steering directions. This is NP-hard in general as it is a supervised, doubly-partitioned generalization of bipartite k-clustering.

Interpretation for ChemSteer. The optimization problem posed in Equation 4 is to find the partition of the molecules and latent dimensions such that we obtain minimum aggregate error for linear property predictors within each subspace. This is a localized version of the Linear Representation Hypothesis, where the property is a linear function of the coordinates within a given subspace. ChemSteer uses spectral co-clustering partitions the representation space using only the representation H (via a nonnegative affinity $A = g(H)$), and it never observes y. Therefore, it is an unsupervised heuristic for this problem. In general, the supervised case would result in a different partition, but Sections C.2 gives the precise condition under which the supervised and unsupervised methods align.

C.2 Spectral Co-Clustering as an Optimal Unsupervised Solver

Let $A \in \mathbb{R}_{\geq 0}^{N\times d}$ be the nonnegative affinity for spectral co-clustering (in our implementation, this is an element-wise nonnegative transform of H). Note that this is distinct from the signed H used to compute μ_j^+, μ_j^-. Under the following assumptions, we show that spectral co-clustering solves the optimization problem in Equation 4.

The procedure. Given $A \in \mathbb{R}_{\geq 0}^{N\times d}$ and k, k-dimensional bipartite spectral co-clustering performs:

(S1) Form the degree matrices $D_r = \mathrm{diag}(A\mathbf{1}_d)$ and $D_c = \mathrm{diag}(A^\top \mathbf{1}_N)$ and the normalized matrix $B = D_r^{-1/2} A D_c^{-1/2}$.

(S2) Compute a top-k paired singular basis (U, V) of B, with $U \in \mathbb{R}^{N\times k}$ and $V \in \mathbb{R}^{d\times k}$ having orthonormal columns spanning the leading k left and right singular subspaces.

(S3) Form the joint degree-scaled embedding $Z = \begin{pmatrix} D_r^{-1/2}U \\ D_c^{-1/2}V \end{pmatrix} \in \mathbb{R}^{(N+d)\times k}$, whose rows index the N molecules followed by the d latent coordinates.

(S4) Run k-means on the $N + d$ rows of Z. The cluster containing molecule i defines $\hat{\pi}_C(i)$; the cluster containing coordinate l defines $\hat{\pi}_D(l)$.

Assumption 1 (Block structure of the affinity)**.** *There exist partitions $(C_1^\star, \ldots, C_k^\star)$ of $[N]$ and $(D_1^\star, \ldots, D_k^\star)$ of $[d]$ under which A is exactly block-diagonal after the corresponding row and column permutations.*

Assumption 2 (Connected blocks, positive degree)**.** *For every j, the bipartite graph induced by block A_j is connected, and every row and column of A_j has strictly positive degree.*

Assumption 3 (Non-degenerate partition)**.** $|C_j^\star| \geq c_{\min}$ *and* $|D_j^\star| \geq d_{\min}$ *for all* j*.*

Theorem 1 (Exact recovery)**.** *Under Assumptions 1–3, the procedure (S1)–(S4) exactly recovers the paired block partition $\pi^\star = (\pi_C^\star, \pi_D^\star)$ induced by A's block structure, up to a global relabeling of cluster indices. The recovered partition is feasible for Problem equation 4.*

The proof (Appendix C.3) establishes that any global k-means minimizer recovers the true components by invoking the Perron–Frobenius theorem on each connected block, leveraging the exact multiplicity-k singular structure of the block-diagonal normalized matrix, and showing that the joint degree-scaled embedding collapses each true component to a single point. The joint embedding of rows and columns is essential for the required (C_j, D_j) pairing.

Theorem 1 is silent about the property y. It concerns the affinity alone, and connecting it to steering requires the additional following condition.

Definition 1 (Aligned Block LRH)**.** *Let A satisfy Assumptions 1–3, so that the paired block partition $\pi^\star$ is well-defined. The setup exhibits* aligned block LRH *if*

$$\pi^\star \ \in \ \operatorname*{arg\,max}_{\pi_C, \pi_D} \ \mathrm{Sep}(\pi_C, \pi_D),$$

that is, if the affinity's structural partition coincides with the partition maximizing property-steering separation.

Corollary 1 (CAA optimality under aligned block LRH)**.** *If, in addition to Assumptions 1–3, the setup exhibits aligned block LRH (Definition 1), then the CAA directions extracted within the recovered clusters jointly achieve* $\mathrm{Sep}(\pi^\star)$*, which by Definition 1 equals the maximum* Sep *attainable by any feasible partition.*

Aligned block LRH is a property of the joint object (A, H, y), not of the algorithm, and the corollary follows from Theorem 1 by substitution. Recovery is unconditional, while Sep-optimality holds only when the representation's clustering structure and the property's linear-response structure agree on the same partition.

The proof establishes that any global k-means minimizer recovers the true components by invoking the Perron–Frobenius theorem on each connected block, leveraging the exact multiplicity-k singular structure of the block-diagonal normalized matrix, and demonstrating that the joint degree-scaled embedding collapses each true component to a single point. The joint embedding of rows and columns is essential for the required (C_j, D_j) pairing.

Disjointness, orthogonality, and $\varepsilon = 0$ are guaranteed directly from spectral co-clustering. The recovery result (Theorem 1) is unconditional in the sense that it holds under Assumptions 1–3 regardless of what the property y looks like. The optimality result (Corollary 1) is conditional on aligned block LRH holding for the specific tuple (A, H, y) under study. This alignment is a claim about the joint object being clustered and the property being predicted; it is not a claim about spectral co-clustering.

Alignment can fail in two identifiable ways. First, A's block structure may not correspond to any partition under which y is well-approximated linearly, making it such that the representation's clustering geometry is unrelated to the property's response geometry. Second, A's block partition may support local linearity but some other partition supports it better. In this case $\pi^\star$ is only a local Sep-optimum, not a global one. In both cases, spectral co-clustering still recovers A's block partition, but the guarantee that recovery equals Sep-maximization is what may be lost.

Whether the deployed ChemSteer pipeline exhibits aligned block LRH is therefore an empirical question about chemical language models specifically. We examine two operational signatures of alignment: (i) systematic angular divergence between local CAA directions with chemical position (Section 3.3), which is the observable signature of the property responding smoothly to representation position within blocks; and (ii) the deployed partition's held-out separation exceeding alternatives (Appendix C.7). Together they are consistent evidence that alignment holds approximately for the domains and datasets studied. Additionally, we also validate the chemical meaningfulness of the resultant clusters in Appendix B.

C.3 Proof of Theorem 1

Proof of Theorem 1. Permuting rows and columns relabels the vertices of the bipartite graph and commutes with every step of (S1)–(S4), so by Assumption 1 we may assume without loss of generality that A is already block diagonal, $A = \mathrm{blkdiag}(A_1, \dots, A_k)$ with $A_j \in \mathbb{R}_{\geq 0}^{|C_j^\star| \times |D_j^\star|}$.

Because off-block entries vanish, the degree of a row $i \in C_j^\star$ depends only on A_j, and likewise for columns; write $D_{r,j}, D_{c,j}$ for the corresponding diagonal blocks. By Assumption 2 every diagonal entry is strictly positive, so $D_r^{-1/2}$ and $D_c^{-1/2}$ exist and $B = D_r^{-1/2} A D_c^{-1/2} = \mathrm{blkdiag}(B_1, \dots, B_k)$ with $B_j = D_{r,j}^{-1/2} A_j D_{c,j}^{-1/2}$.

Step 1: each B_j has largest singular value exactly 1, with multiplicity one.

Using $A_j \mathbf{1} = D_{r,j}\mathbf{1}$ and $A_j^\top \mathbf{1} = D_{c,j}\mathbf{1}$,

$$B_j\big(D_{c,j}^{1/2}\mathbf{1}\big) = D_{r,j}^{-1/2} A_j \mathbf{1} = D_{r,j}^{1/2}\mathbf{1}, \qquad B_j^\top\big(D_{r,j}^{1/2}\mathbf{1}\big) = D_{c,j}^{-1/2} A_j^\top \mathbf{1} = D_{c,j}^{1/2}\mathbf{1},$$

so $\big(D_{r,j}^{1/2}\mathbf{1},\ D_{c,j}^{1/2}\mathbf{1}\big)$ is a paired singular vector pair of B_j with singular value 1, and both vectors are strictly positive by Assumption 2.

To see that 1 is the *largest* singular value, form the symmetrized matrix

$$\mathcal{B}_j = \begin{pmatrix} 0 & B_j \\ B_j^\top & 0 \end{pmatrix} = \mathcal{D}_j^{-1/2} \mathcal{M}_j \mathcal{D}_j^{-1/2}, \qquad \mathcal{M}_j = \begin{pmatrix} 0 & A_j \\ A_j^\top & 0 \end{pmatrix}, \quad \mathcal{D}_j = \mathrm{diag}(D_{r,j}, D_{c,j}),$$

which is the normalized adjacency matrix of the bipartite graph induced by A_j. Its eigenvalues are $\pm\sigma_i(B_j)$. Since $\mathcal{B}_j$ is similar to $\mathcal{D}_j^{-1}\mathcal{M}_j$, a row-stochastic nonnegative matrix, its spectral radius is exactly 1; hence $\sigma_i(B_j) \leq 1$ for all i.

By Assumption 2 the graph is connected, so $\mathcal{M}_j$ is irreducible and the Perron–Frobenius theorem gives that the eigenvalue 1 of $\mathcal{B}_j$ is simple. Consequently $\sigma_1(B_j) = 1$ with multiplicity one and $\sigma_l(B_j) < 1$ for all $l \geq 2$.

Step 2: the top-k singular subspaces of B are spanned by the block indicators.

Block diagonality means the singular values of B are the union of those of the blocks. By Step 1, the value 1 occurs exactly k times, once per block, and every remaining singular value is strictly less than 1. The leading k-dimensional left and right singular subspaces $\mathcal{U}, \mathcal{V}$ are therefore uniquely determined.

Let ι_j denote zero-padding of a block-j vector into $\mathbb{R}^N$ (for rows) or $\mathbb{R}^d$ (for columns), and set

$$\hat{u}_j = \frac{\iota_j\big(D_{r,j}^{1/2}\mathbf{1}\big)}{\big\|D_{r,j}^{1/2}\mathbf{1}\big\|}, \qquad \hat{v}_j = \frac{\iota_j\big(D_{c,j}^{1/2}\mathbf{1}\big)}{\big\|D_{c,j}^{1/2}\mathbf{1}\big\|}.$$

The $\hat{u}_j$ have disjoint supports and are therefore orthonormal, as are the $\hat{v}_j$, and Step 1 gives $B\hat{v}_j = \hat{u}_j$ once the two normalizations agree. They do: writing $\mathrm{vol}(A_j) = \sum_{i \in C_j^\star} \sum_{l \in D_j^\star} A_{il}$ for the total edge weight of block j,

$$\big\|D_{r,j}^{1/2}\mathbf{1}\big\|^2 = \sum_{i \in C_j^\star} (D_r)_{ii} = \mathrm{vol}(A_j) = \sum_{l \in D_j^\star} (D_c)_{ll} = \big\|D_{c,j}^{1/2}\mathbf{1}\big\|^2,$$

since both sums count every edge of block j exactly once. Hence $U^\star = [\hat{u}_1, \dots, \hat{u}_k]$ and $V^\star = [\hat{v}_1, \dots, \hat{v}_k]$ form an admissible paired basis with $BV^\star = U^\star$. Any other orthonormal basis of $\mathcal{V}$ has the form $V = V^\star R$ for some orthogonal $R \in O(k)$, and then $U = BV = BV^\star R = U^\star R$: admissible paired bases differ only by a common orthogonal rotation.

Step 3: the joint embedding collapses each component to a single point.

Under the canonical basis, a row $i \in C_j^\star$ maps to

$$Z_{i,:} = (D_r)_{ii}^{-1/2} \cdot \hat{u}_j(i)\, e_j^\top = (D_r)_{ii}^{-1/2} \cdot \frac{(D_r)_{ii}^{1/2}}{\sqrt{\mathrm{vol}(A_j)}}\, e_j^\top = \frac{e_j^\top}{\sqrt{\mathrm{vol}(A_j)}},$$

and the identical computation for a column $l \in D_j^\star$ gives $Z_{l,:} = e_j^\top / \sqrt{\mathrm{vol}(A_j)}$, using the equality of normalizations established in Step 2. So all $|C_j^\star| + |D_j^\star|$ vertices of component j, rows and columns alike, map to the single point $p_j := e_j / \sqrt{\mathrm{vol}(A_j)} \in \mathbb{R}^k$. For $j \neq m$ the points p_j and p_m are supported on distinct coordinates and are both nonzero, hence distinct. Under any other admissible basis the embedding becomes ZR, a rigid rotation, which leaves all pairwise distances and therefore the k-means objective unchanged.

Step 4: k-means recovers the components uniquely.

The $N+d$ embedded vertices occupy exactly k distinct locations $p_1, \ldots, p_k$. Assigning each location its own cluster achieves zero within-cluster sum of squares. Conversely, suppose some assignment achieves zero cost. Then no cluster contains two distinct locations, since a cluster spanning two distinct points has positive within-cluster sum of squares. If the vertices at some location p_j were split across two or more clusters, then at most $k-2$ clusters would remain for the $k-1$ other locations, so by the pigeonhole principle some cluster would contain two distinct locations, a contradiction. Hence every zero-cost assignment is a bijection between clusters and locations, and the component assignment is the unique global k-means minimizer up to relabeling. Step (S4) selects it.

Because rows and columns are embedded and clustered *jointly* on $[N] \sqcup [d]$, the row cluster and the column cluster carrying a given label are exactly those belonging to the same component, so the pairing $(\hat{C}_j, \hat{D}_j)$ is consistent by construction. Thus $\hat{\pi}^{\text{SCC-}k} = \pi^\star$ up to a global relabeling, and by Assumption 3 the recovered partition satisfies $|\hat{C}_j| \geq c_{\min}$ and $|\hat{D}_j| \geq d_{\min}$ for every j, so it is feasible for Problem equation 4. □

Proof of Corollary 1. Theorem 1 yields $\hat{\pi}^{\text{SCC-}k} = \pi^\star$ up to relabeling, and Definition 1 states $\pi^\star \in \arg\max \mathrm{Sep}$. Combining, $\mathrm{Sep}(\hat{\pi}^{\text{SCC-}k}) = \max_{\pi_C, \pi_D} \mathrm{Sep}(\pi_C, \pi_D)$. □

C.4 Theorem 1 Discussion

The theorem is stated for a nonnegative affinity matrix A supplied to spectral co-clustering, distinct from the possibly-signed representation matrix H. In deployed ChemSteer, A may be derived from H by a nonnegative transformation $A = g(H)$ (e.g. element-wise absolute value or a co-activation similarity kernel). The theorem does not require H itself to be block diagonal, only A. Alignment between A's block partition and the Sep-maximizing partition is required by Corollary 1, not by Theorem 1.

The joint degree-scaled embedding Z is essential. If rows and columns were clustered separately (k-means on $D_r^{-1/2}U$ for rows and independently on $D_c^{-1/2}V$ for columns), the two clusterings could assign the same true component to different cluster labels (e.g. true component j receives row-label 2 but column-label 5). Because the downstream property-steering and CAA objectives require the paired (C_j, D_j) association to be consistent, independent labelings can produce arbitrary degradation even when the underlying partitions are structurally correct. Joint clustering on $[N] \sqcup [d]$ enforces label consistency by construction.

C.5 Approximation

Denote by $\bar{A}$ the exact block-diagonal reference affinity satisfying Assumptions 1–3 (so $\bar{A} = \mathrm{blkdiag}(\bar{A}_1, \ldots, \bar{A}_k)$ with each $\bar{A}_j$ inducing a connected bipartite graph with positive degrees), and let

$$\bar{B} = \bar{D}_r^{-1/2} \bar{A} \bar{D}_c^{-1/2}$$

be its normalization. By the exact theorem, $\sigma_1(\bar{B}) = \cdots = \sigma_k(\bar{B}) = 1$. Define the reference spectral separation

$$\delta_0 \;:=\; 1 - \sigma_{k+1}(\bar{B}) \;>\; 0$$

(with the convention $\sigma_{k+1}(\bar{B}) = 0$ when $k = \min(N, d)$). The exact theorem admits an approximate analog under perturbed versions of the assumptions:

- off-diagonal blocks of the observed A small in Frobenius norm rather than exactly zero: $\|A - \bar{A}\|_F$ small,

- row and column degrees of A bounded uniformly away from zero,
- approximate rather than maximal alignment between the affinity block structure and the separation objective,
- bounded feature magnitudes and representations (needed to control perturbation bounds under approximate assignments),
- the reference separation $\delta_0 = 1 - \sigma_{k+1}(\bar{B})$ bounded below and larger than the induced perturbation of the normalized matrix.

Under such conditions, Davis–Kahan / Wedin perturbation bounds Wedin (1972) yield an estimate of the form

$$\left\|\sin\Theta(\hat{\mathcal{U}},\mathcal{U}_\star)\right\| \;\lesssim\; \frac{\|B-\bar{B}\|}{\delta_0},$$

where $\hat{\mathcal{U}}$ is the top-k singular subspace of the perturbed normalized matrix B and $\mathcal{U}_\star$ is the reference singular subspace of $\bar{B}$, valid when the numerator is smaller than the denominator. Positive-degree lower bounds control the sensitivity of the degree-scaling step $A \mapsto B$; cluster-volume separation conditions convert spectral-subspace error into k-means clustering error via a Kumar–Kannan-style analysis Kumar & Kannan (2010).

C.6 Supervised Case

Theorem 1 concerns the *unsupervised* case: spectral co-clustering operates on the affinity matrix A alone and does not optimize Sep directly. Under aligned separation, this unsupervised procedure recovers the optimal partition $\pi^\star$ because the block structure of A and the maximum separation structure agree. When alignment fails, SCC recovers A's block structure but not necessarily the partition that maximizes separation.

Better algorithms are possible in principle: any procedure that incorporates the target property or separation objective directly can achieve a separation score no lower than the unsupervised optimum. Candidate approaches include:

- Two-stage supervised refinement. Initialize with $\hat{\pi}^{\text{SCC-}k}$, then run a small number of alternating-optimization steps on the separation objective.
- Continuous relaxation. Relax Z_C, Z_D to $[0,1]$ row-stochastic matrices under Gumbel-softmax parameterizations and solve as a differentiable program.

C.7 Empirical Test of Aligned Block LRH

Theorem 1 is exact under (1)–(3). For synthetic validation, please refer to Appendix D. We assess how aligned the representations that ChemSteer operates on are with all evaluations on held-out test set molecules. For a partition and its extracted CAA steering directions, define per cluster the held-out separation gap $v_j^\top(\bar{\mu}_j^{+,\text{test}} - \bar{\mu}_j^{-,\text{test}})$ (test-set class means, same fixed train-derived quartile thresholds), the held-out AUC of classifying test top- vs. bottom-quartile molecules by $v_j^\top h$, and the held-out Pearson correlation of $v_j^\top h_i$ with y_i over all test molecules routed to cluster j. We report (1) $\mathrm{Sep}^\star(1)$, a single global CAA direction (no partitioning); (2) $\widetilde{\mathrm{Sep}}^\star(k)$, row-piecewise CAA, full dimensionality per cluster (isolates the value of row-wise local specialization alone); (3) $\mathrm{Sep}^\star(k)$, the coordinate-restricted optimum of equation 4, approximated via alternating optimization initialized from the deployed SCC partition; (4) $\mathrm{Sep}(\pi^{\mathrm{ChemSteer}})$, the actual CAA vectors, evaluated for held-out generalization; and (5) $\mathrm{Sep}(\pi^{\mathrm{rand}})$, a random-partition null.

Table 6 shows that the ChemSteer partitioning performs the best, with the exception of $\mathrm{Sep}^\star(1)$. However, as previously discussed and empirically validated (Tables 4 and 3), the naive global direction which is tested by $\mathrm{Sep}^\star(1)$ is not effective for steering. $\mathrm{Sep}^\star(1)$ achieves the highest values in Table 6 merely due to the fact that the positive and negative sets are defined across the entire dataset, making this more separable by construction, however, we have clearly shown that this separating direction is not effective for steering. Moreover, $\mathrm{Sep}^\star(1)$ is not a feasible partition. Equation 4 optimizes over partitions into exactly k blocks subject to $|C_j| \geq c_{min}$, and $|D_j| \geq d_{min}$. $\mathrm{Sep}^\star(1)$ is a single unpartitioned cluster, hence, violates this condition.

Table 6: Separation AUCs across datasets and molecular properties.

Dataset	Property	$\mathrm{Sep}^\star(1)$	$\widetilde{\mathrm{Sep}}^\star(k)$	$\mathrm{Sep}^\star(k)$	$\mathrm{Sep}(\pi^{\mathrm{ChemSteer}})$	$\mathrm{Sep}(\pi^{\mathrm{rand}})$
ChEMBL	LogP	0.9181	0.5815	0.7188	0.7390	0.6503
ChEMBL	QED	0.9106	0.5644	0.7668	0.7982	0.7312
ChEMBL	SA	0.9616	0.6069	0.8153	0.8996	0.6002
ZINC	LogP	0.9418	0.6364	0.6695	0.7544	0.6194
ZINC	QED	0.7680	0.5269	0.6358	0.6846	0.5379
ZINC	SA	0.9934	0.7610	0.9264	0.9269	0.7127
QM9	LogP	0.9431	0.5900	0.7792	0.8867	0.6318
QM9	QED	0.8409	0.6482	0.6903	0.8229	0.5833
QM9	SA	0.9635	0.7605	0.7575	0.8766	0.6942
MOSES	LogP	0.8762	0.6301	0.6599	0.7406	0.5672
MOSES	QED	0.8181	0.5666	0.5978	0.7551	0.5318
MOSES	SA	0.9051	0.6131	0.8909	0.8426	0.5842

Therefore, we can disregard this column of the table. As compared with the other partitioning strategies and the random null, ChemSteer's partitioning ($\mathrm{Sep}(\pi^{\mathrm{ChemSteer}})$) proves to have the highest AUC among all datasets and molecular properties, the only exception being MOSES SA. This empirically confirms that the conditions of Corollary 1 (aligned block LRH) hold approximately in the chemical setting. The results show that the deployed spectral co-clustering partition achieves higher held-out separation than alternatives. Under the aligned block LRH, Corollary 1 predicts this ordering, so the empirical result is consistent with the operational claim of the corollary in this domain. Intuitive reasoning behind this follows from the idea that the local subspaces need to be chemically similar enough to have meaningful steering directions, but the spaces cannot be too similar as to not provide a useful direction of contrast in positive and negative sets.

D Synthetic Validation of Theorem 1

Section C.7 tests whether real molecular representations approximately satisfy Theorem 1's assumptions. It validates that these assumptions actually hold up in the chemical domain. Here, we confirm that when we operate in the strictly synthetic setting where all assumptions are satisfied by construction, we see the exact behavior as outlined in 1. In this synthetic validation, we generate the synthetic data, so the true optimal partition and its exact $\mathrm{Sep}(\pi^\star)$ value are known quantities, making this an exact check.

Construction. We generate $N = k \cdot n_r$ molecules and $d = k \cdot n_c$ latent dimensions with a planted k-block structure. Each molecule i, assigned to true row-block j, draws a latent score $z_i \sim \mathrm{Uniform}(-1, 1)$; the target property is $y_i = z_i$. For columns in i's own true block ($m = j$), $H_{i,l} = \beta z_i + \eta_{i,l}$ with $\eta_{i,l} \sim \mathcal{N}(0, \sigma^2)$ (a real, planted class-mean gap along a per-coordinate signal, scaled by β); for columns in any other block ($m \neq j$), $H_{i,l} = 0$ exactly (in the base construction) or $\eta_{i,l} \sim \mathcal{N}(0, \gamma^2)$ for a controlled leakage level γ (Check 4 below). Rows and columns are then randomly permuted so the true blocks are not trivially contiguous. The affinity A passed to spectral co-clustering is computed by the identical transform used in the deployed pipeline: $A = |H|$, row-normalized.

We have $\gamma = 0$, so the off-block entries of H are exactly zero for every molecule regardless of its true cluster, so (A1)–(A2) hold exactly by construction (block-diagonal A; within-block entries are all strictly positive continuous random variables, giving trivial connectivity). Any partition reassigning a dimension away from its true home block strictly decreases that block's contribution to Sep while contributing exactly zero to the receiving cluster (since the receiving cluster's own rows are exactly zero there).

Check 1+2: exact recovery and random-partition margin. With $k = 5$, $n_r = 60$, $n_c = 15$, $\beta = 3$, $\sigma = 1$, the spectral co-clustering recovers the true row and column partitions exactly (Adjusted

Rand Index $= 1.000000$ for both), and the recovered partition's Sep value equals the true partition's to machine precision (87.0819 in both cases). Against 500 uniformly random partitions at the same k, the true partition's Sep value exceeds every single draw (mean random $\mathrm{Sep} = 18.58$, max $= 21.86$, true $= 87.08$, which is roughly $4\times$ the largest random draw, 100th percentile).

Check 3: exhaustive verification over column-block permutations. Holding the row partition at ground truth, we enumerate all $k! = 24$ ways of assigning the $k = 4$ pre-existing dimension blocks to the 4 clusters ($n_r = 80$ which is large enough that each cluster's own quartile groups reliably clear the minimum-class-size floor) and compute Sep for every assignment. The true (identity) assignment achieves the strict, unique maximum (40.73 vs. 20.38 for the next highest).

Check 4: graceful degradation under controlled violation of (A1). Sweeping the leakage level γ (relative to σ) at $k = 5$, $n_r = 60$, $n_c = 15$:

γ/σ	0.0	0.1	0.3	0.6	1.0	1.5	2.0
Row ARI	1.000	1.000	0.932	0.561	0.334	0.186	0.037
Col ARI	1.000	1.000	1.000	1.000	1.000	1.000	0.466

Recovery degrades smoothly rather than catastrophically as the block structure is progressively violated. Column recovery is substantially more stable despite this leakage than row recovery in this construction. The algorithm's dimension partitioning appears to tolerate representation noise better than its molecule partitioning at comparable leakage levels, at least under this synthetic planted-signal validation.

Relation to the real-data experiments. This appendix establishes that the recovery mechanism works exactly as proven, when its assumptions hold. Validation that our real molecular representations satisfy the assumptions is given in Appendix C.

E Spectral Co-clustering Subspace Analysis

We provide a full analysis of the resultant clusters from spectral co-clustering on each dataset. We confirm for each dataset that spectral co-clustering yields semantically meaningful clusters (Figure 3 for ZINC, Figure 4 for QM9, Figure 5 for MOSES, and Figure 6 for ChEMBL).

E.1 Interpreting Co-Clustering Results

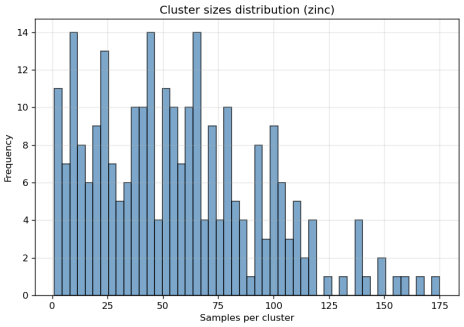
The orthogonality and low overlap provided by spectral co-clustering do not by themselves guarantee that the resulting clusters are interpretable or meaningful with respect to the underlying chemistry. In other words, low overlap ϵ is a necessary but not a sufficient condition for meaningful structure. Whether clusters correspond to chemically coherent or scaffold-enriched groups depends on the data distribution and the latent structure learned by the model.

We validate the meaningfulness of the spectral co-clustering output through several empirical diagnostics reported in Section 3.2.

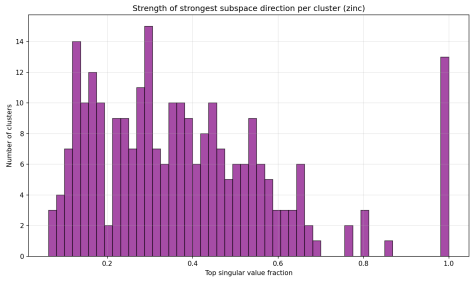
Cluster scaffold uniqueness: We measure the fraction of scaffolds that are unique to each cluster and the distribution of scaffolds per cluster. Higher scaffold specificity suggests that clusters capture chemically distinct families rather than arbitrary partitions.

Subspace strength: Within each cluster, we compute the singular value decomposition and quantify the fraction of variance captured by the dominant direction. Clusters with a strong dominant mode indicate coherent representation structure within the subspace.

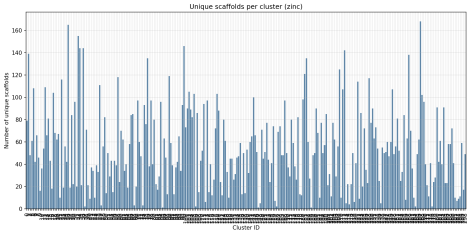
Chemical similarity: We compute average within-cluster Tanimoto similarity and SMILES edit distance to assess whether molecules grouped by clustering are chemically and structurally similar beyond merely sharing latent activation patterns.



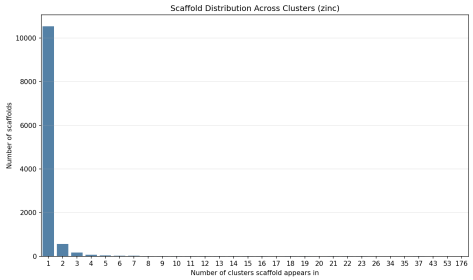
(a) Distribution of cluster sizes for spectral co-clustering on ZINC with 300 clusters; the average cluster size is 56 samples per cluster, with an average rank of 10.



(b) Distribution of the strength of the strongest subspace direction per cluster for spectral co-clustering on ZINC with 300 clusters; shows that the top singular value captures a significant amount of the energy for most clusters, averaging 43.30%.



(c) Barplot of the number of unique scaffold within each of the 300 clusters from spectral co-clustering on ZINC. Each bar (x-axis) represents a specific cluster. The average number of unique scaffolds per cluster is 55.



(d) Barplot showing the scaffold distribution across clusters for spectral co-clustering on ZINC with 300 clusters; 91.7% of scaffolds only occur in a singular cluster. The relationship between scaffolds and clusters is many-to-one.

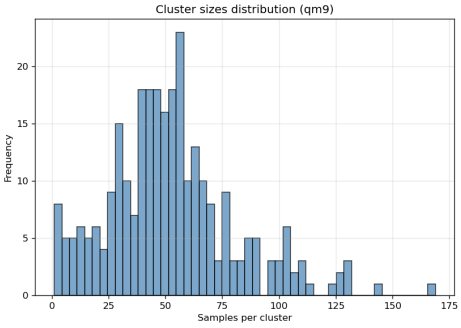
Figure 3: Analysis of spectral co-clustering with 300 clusters on the ZINC dataset. The resultant clusters are semantically meaningful for ZINC.

Bipartite cut statistics: By interpreting the scaffold–cluster assignment as a cut on the bipartite graph between scaffolds and clusters, we quantify the number of edges crossing the cut relative to randomized assignments. Fewer cut edges than a random baseline indicate alignment between chemical motifs and cluster structure.

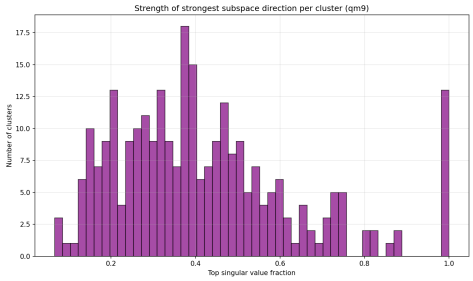
Together, these diagnostics demonstrate that spectral co-clustering not only yields mathematically orthogonal subspaces, but also uncovers clusters that reflect meaningful chemical and structural patterns in the data. Additional cluster analysis visualizations for each dataset are presented below.

E.2 Cluster Statistics

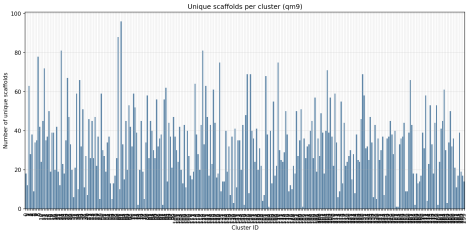
We assess chemical similarity within clusters. We compute the average pairwise Tanimoto similarity and SMILES Levenshtein edit distance among molecules assigned to the same cluster and report the results in Table 7. Spectral co-clustering yields substantially higher within-cluster Tanimoto similarity and lower edit distance compared to random assignments (R), indicating that clusters group chemically and structurally related molecules rather than arbitrary collections. These data confirm that the spectral co-clustering method produces semantically meaningful clusters for molecular data.



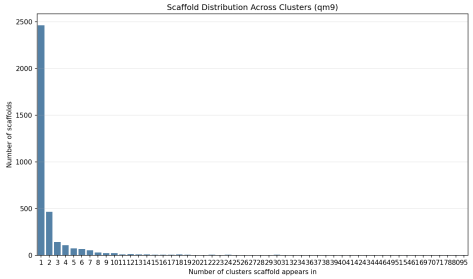
(a) Distribution of cluster sizes for spectral co-clustering on QM9 with 300 clusters; the average cluster size is 52 samples per cluster, with an average rank of 10.



(b) Distribution of the strength of the strongest subspace direction per cluster for spectral co-clustering on QM9 with 300 clusters; shows that the top singular value captures a significant amount of the energy for most clusters, averaging 44.04%.



(c) Barplot of the number of unique scaffold within each of the 300 clusters from spectral co-clustering on QM9. Each bar (x-axis) represents a specific cluster. The average number of unique scaffolds per cluster is 32.



(d) Barplot showing the scaffold distribution across clusters for spectral co-clustering on QM9 with 300 clusters; 69% of scaffolds only occur in a singular cluster. The relationship between scaffolds and clusters is many-to-one.

Figure 4: Analysis of spectral co-clustering with 300 clusters on the QM9 dataset. The resultant clusters are semantically meaningful for QM9.

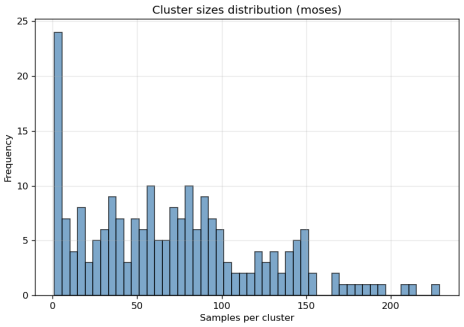
Table 7: Within-cluster chemical and structural similarity metrics for spectral co-clustering. We report the average pairwise Tanimoto similarity (TS) (fingerprint-based) and average Levenshtein edit distance among SMILES in each cluster. Higher Tanimoto similarity and lower edit distance indicate that clusters contain chemically and structurally related molecules. Randomly assigned clusters (designated with (R)) are included for comparison.

	TS ↑	Edit ↓	TS (R)	Edit (R)
chEMBL	**0.144**	**46.61**	0.111	54.28
moses	**0.180**	**21.49**	0.133	25.03
qm9	**0.162**	**7.10**	0.078	9.78
ZINC	**0.156**	**30.28**	0.119	34.35

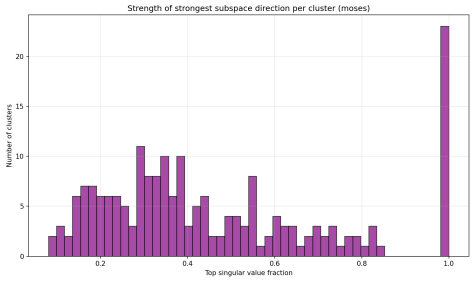
E.3 Within Cluster Molecular Similarity

The improvement in Tanimoto similarity is highly statistically significant (paired t-test, $p = 2.6 \times 10^{-20}$; Wilcoxon signed-rank test, $p = 3.9 \times 10^{-18}$), with a large effect size (Cohen's $d = 0.93$). The bootstrap confidence interval $[0.040, 0.057]$ excludes zero, indicating a consistent and practically meaningful improvement.

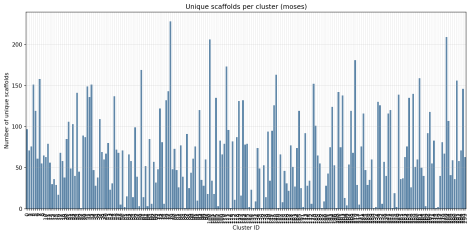
It is important to note, however, that these metrics have limitations. Tanimoto similarity captures fingerprint-level similarity, which may miss slight scaffold variations, stereochemistry, or functional



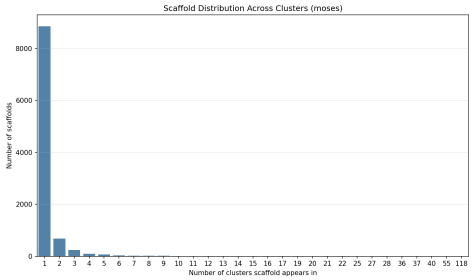
(a) Distribution of cluster sizes for spectral co-clustering on MOSES with 300 clusters; the average cluster size is 71 samples per cluster, with an average rank of 7.



(b) Distribution of the strength of the strongest subspace direction per cluster for spectral co-clustering on MOSES with 300 clusters; shows that the top singular value captures a significant amount of the energy for most clusters, averaging 65.26%.



(c) Barplot of the number of unique scaffold within each of the 300 clusters from spectral co-clustering on MOSES. Each bar (x-axis) represents a specific cluster. The average number of unique scaffolds per cluster is 65.



(d) Barplot showing the scaffold distribution across clusters for spectral co-clustering on MOSES with 300 clusters; 87.6% of scaffolds only occur in a singular cluster. The relationship between scaffolds and clusters is many-to-one.

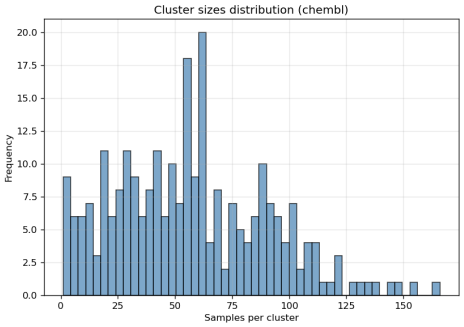
Figure 5: Analysis of spectral co-clustering with 300 clusters on the MOSES dataset. The resultant clusters are semantically meaningful for MOSES.

group rearrangements. Similarly, Levenshtein edit distance on SMILES strings is sensitive to the particular linearization of the molecule and does not necessarily reflect true topological or scaffold similarity. There also may be substrings or molecular substructures (such as carbon rings) which repeat within a scaffold, causing the edit distance to be arbitrarily large even for structurally similar molecules. Overall, these results do indeed suggest that spectral co-clustering can group molecules with shared chemical features more effectively than random assignment.

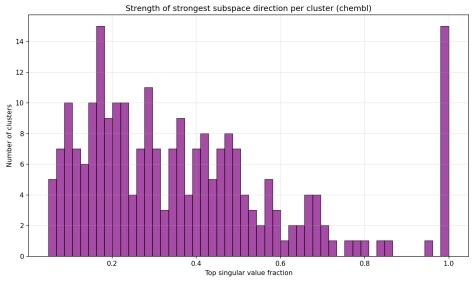
We illustrate the aforementioned shortcoming of similarity measures in Figure 7. Figure 7a shows an example of the scaffolds that occur in a cluster resulting from spectral co-clustering, and 7b shows an example of the scaffolds in an example cluster from random assignment. We selected two random clusters with the same number of scaffolds from each cluster assignment method for comparison. All scaffolds within a are unique; they may only differ by the placement or value of a single character. Slight differences in SMILES highlight that standard similarity measures can underestimate scaffold-level relationships.

F Section 3.3 Experimental Setup

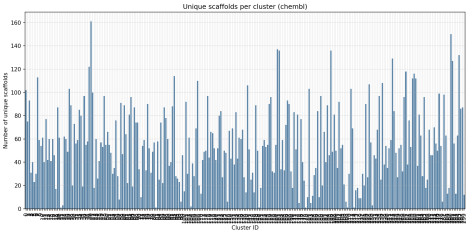
Setup. For each dataset and property we compute a contrastive activation direction within every cluster, using the full 4096 dimensions rather than the coordinate-restricted subspace, so that the measured angles reflect direction rather than support. We then form two matrices over cluster pairs: the pairwise angle θ_{ij} between local directions, and the cross-cluster average Tanimoto similarity



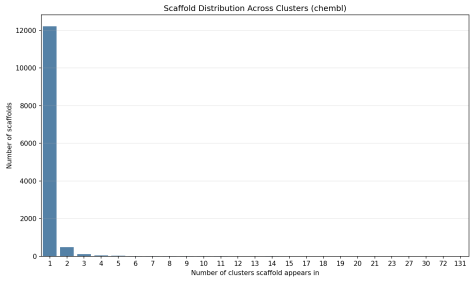
(a) Distribution of cluster sizes for spectral co-clustering on ChEMBL with 300 clusters; the average cluster size is 57 samples per cluster, with an average rank of 10.



(b) Distribution of the strength of the strongest subspace direction per cluster for spectral co-clustering on ChEMBL with 300 clusters; shows that the top singular value captures a significant amount of the energy for most clusters, averaging 42.06%.



(c) Barplot of the number of unique scaffold within each of the 300 clusters from spectral co-clustering on ChEMBL. Each bar (x-axis) represents a specific cluster. The average number of unique scaffolds per cluster is 53.



(d) Barplot showing the scaffold distribution across clusters for spectral co-clustering on ChEMBL with 300 clusters; 94.5% scaffolds only occur in a singular cluster. The relationship between scaffolds and clusters is many-to-one.

Figure 6: Analysis of spectral co-clustering with 300 clusters on the ChEMBL dataset. The resultant clusters are semantically meaningful for ChEMBL.

T_{ij} over Morgan fingerprints of 20 molecules sampled per cluster. Under the curvature account, θ_{ij} should decrease with T_{ij}.

Statistics. Cluster pairs are not independent: pairs sharing a cluster are coupled. We therefore assess significance with a Mantel permutation test, which permutes the joint row/column indices of one matrix ($n_{\text{perm}} = 10,000$) rather than treating pairs as exchangeable. Smaller clusters yield noisier direction estimates and are also more chemically homogeneous, which could manufacture the correlation; we control for this by residualizing both θ_{ij} and T_{ij} against $\min(N_i, N_j)$ before correlating.

G Comparison with Naive Global Steering

In this section we provide baselines against no steering and naive global steering. Our claim is not only that the clustered vector from ChemSteer outperforms the naive one, but that the incoherent-subspace decomposition is a better model of how the LLM itself represents molecular structure. That is, the representation space is not organized as a single coherent latent manifold but as disjoint feature subspaces in superposition, consistent with the superposition hypothesis we build on in the Introduction. A single global CAA vector computed over the whole dataset is then an average that mixes signals from every subspace. A locally-computed vector, by contrast, is extracted only from

```
Cluster 1: 20 unique scaffolds
    N=c1ccc(S(=O)(=O)c2ccccc2)c[nH]1
    O=C(c1ccccc1S(=O)(=O)c1ccccc1)N1CCC1
    O=C(c1cccnc1S(=O)(=O)c1ccccc1)N1CCCC1
    O=C(c1cnccc1S(=O)(=O)c1ccccc1)N1CCC1
    O=C(c1ncccc1S(=O)(=O)c1ccccc1)N1CCC1
    O=S(=O)(c1ccccc1)c1cc[nH]c1
    O=S(=O)(c1ccccc1)c1ccc[nH]1
    O=S(=O)(c1ccccc1)c1ccccc1
    O=S(=O)(c1ccccc1)c1ccccn1
    O=S(=O)(c1ccccc1)c1cccnc1
    O=S(=O)(c1ccccc1)c1ccco1
    O=S(=O)(c1ccccc1)c1ccncc1
    O=S(=O)(c1ccccc1)c1ccncn1
    O=S(=O)(c1ccccc1)c1ccnnc1
    O=S(=O)(c1ccccc1)c1cnccn1
    O=S(=O)(c1ccccc1)n1cccc1
    O=S(=O)(c1ccccn1)c1ccccn1
    O=S(=O)(c1cccnc1)c1ccccn1
    O=c1ccc(S(=O)(=O)c2ccccc2)c[nH]1
    O=c1cccc(S(=O)(=O)c2ccccc2)[nH]1
```

(a) Unique scaffolds from cluster 1 (arbitrary) of the normal spectral co-clustering cluster assignment.

```
Cluster 56: 20 unique scaffolds
    O=C(CCNC(=O)c1cc2scnc2[nH]1)NC1CCCCC1
    O=C(CCn1cnnn1)N1CCCC(CCc2ccccc2)C1
    O=C(CN1C(=O)CCOc2ccccc21)Nc1c[nH]ccc1=O
    O=C(CN1CCCNC1=O)NC1CCCOc2ccccc21
    O=C(CN1CCOC(Cn2cncn2)C1)NCCc1ccsc1
    O=C(COCc1nc2ccccc2[nH]1)N1CCCOCC1
    O=C(CS(=O)(=O)c1nncn1C1CC1)NCC1CCCCC1
    O=C(CSc1cn[nH]n1)N1c2ccccc2CCc2ccccc21
    O=C(CSc1cn[nH]n1)Nc1ccccc1
    O=C(Cc1noc(CCSc2ccccn2)n1)NC1CCCC1
    O=C(NCCCc1ccsc1)c1ccccc1
    O=C(NCc1ccc(Cc2cccs2)s1)c1cnccn1
    O=C(Nc1c[nH]ccc1=O)c1cn(Cc2cccs2)nn1
    O=C(Nc1cn[nH]c1)N1CCCC2CCCC21
    O=C(OC(Cn1cncn1)c1ccccc1)c1cccnc1
    O=C(OCc1ccccc1)N1CCCC1
    O=C(c1cc2sccc2[nH]1)N1CC2CCCCC2C1
    O=C(c1cccnc1)N1CC2CCCC2C1
    O=C(c1ccco1)N1CCCC2(CC(Nc3ncccn3)CO2)C1
    O=C(c1ccsc1)N1CC2COCC2(COCC2CC2)C1
```

(b) Unique scaffolds from cluster 56 (arbitrary) of the randomized spectral co-clustering cluster assignment.

Figure 7: Comparison of unique scaffolds in a single cluster under (a) spectral co-clustering (b) versus random cluster assignment. Subtle differences in SMILES strings illustrate that simple similarity metrics may not fully capture scaffold-level relationships.

Table 8: Top-100 molecular property scores (mean $\pm$ standard deviation) for no steering, naive global steering, and the original ChemSteer method.

Method	Dataset	LogP	QED	SA
No steer	ChEMBL	3.62 ± 0.87	0.87 ± 0.03	1.86 ± 0.21
	MOSES	3.82 ± 0.58	0.88 ± 0.03	1.93 ± 0.13
	QM9	1.50 ± 0.45	0.81 ± 0.02	2.63 ± 0.31
	ZINC	2.87 ± 0.42	0.88 ± 0.04	2.05 ± 0.27
Naive global steer	ChEMBL	5.02 ± 0.61	0.86 ± 0.03	1.90 ± 0.13
	MOSES	3.95 ± 0.56	0.90 ± 0.02	1.92 ± 0.12
	QM9	0.54 ± 0.66	0.58 ± 0.04	2.95 ± 0.50
	ZINC	4.34 ± 0.67	0.89 ± 0.03	1.94 ± 0.20
Original ChemSteer	ChEMBL	5.24 ± 1.19	0.94 ± 0.10	1.55 ± 0.18
	MOSES	6.68 ± 0.64	0.95 ± 0.00	1.50 ± 0.08
	QM9	2.70 ± 0.16	0.92 ± 0.01	1.61 ± 0.17
	ZINC	6.82 ± 0.71	0.94 ± 0.00	1.54 ± 0.10

the coordinates relevant to its own subspace and contains less noise. This is also consistent with what we already observe geometrically in Appendix O and with our discussion of Table 19 therein.

We ran the following ablation experiment using the naive single global vector and a no steering setting. In both cases, ChemSteer remains the most effective method, beating both of these ablations. This is evidence that the steering is effective over unsteered generation, and that the ChemSteer composite vector outperforms the naive global one.

Table 9: Top-100 steering performance comparison between ChemSteer models using SmileyLlama and GPT-2 backbones across datasets and molecular properties.

	ChemSteer (SmileyLlama)			ChemSteer (GPT-2)		
Dataset	LogP	QED	SA	LogP	QED	SA
ChEMBL	5.24 ± 1.19	0.94 ± 0.10	1.55 ± 0.18	2.25 ± 0.35	0.94 ± 0.00	1.92 ± 0.10
MOSES	6.68 ± 0.64	0.95 ± 0.00	1.50 ± 0.08	4.34 ± 0.46	0.94 ± 0.00	1.91 ± 0.09
QM9	2.70 ± 0.16	0.92 ± 0.01	1.61 ± 0.17	2.19 ± 0.50	0.94 ± 0.01	1.92 ± 0.16
ZINC250K	6.82 ± 0.71	0.94 ± 0.00	1.54 ± 0.10	3.22 ± 0.39	0.94 ± 0.00	1.91 ± 0.09

H GPT2 AND LAYER EXPERIMENTS

H.1 GPT2

We also show performance using GPT2 as the backbone for ChemSteer in place of SmileyLlama in Table 9.

H.2 LAYER ABLATION

To further explore the effect of the layer from which we extract our representations, we perform a study using a 5000 molecule subset of the ZINC dataset used in our other experiments. We extract latent representations from layers 1, 12, 20, 24, 28 and 31 of SmileyLlama, and calculate the cosine similarities between the sparse and dense local vectors, using 100 clusters due to the smaller dataset size. We see similar values between layers (Table 11) and pick layer 24 for the rest of our analysis, aligned with insights from Rogers et al. (2020); Subramani et al. (2022).

Table 10: Cosine similarity between sparse and dense local steering vectors using GPT2 (computed over non-zero sparse features). Higher is better.

Dataset	LogP	QED	SA
ChEMBL	0.981	0.981	0.980
MOSES	0.979	0.981	0.979
QM9	0.982	0.980	0.983
ZINC	0.979	0.978	0.979

H.3 CHEMSTEER WITH RANDOM CLUSTERING

In order to evaluate the effectiveness of spectral co-clustering in capturing property signals, we compare the results in Table 19 to the alignment between sparse and dense steering vectors calculated from random clusters. We construct these random clusters by first randomly assigning 15,000 samples to 300 clusters. For each cluster, we then identify and retain the top 20 dimensions when sparsifying, based on feature importance within the cluster. This mirrors the role of spectral co-clustering, which similarly identifies a small set of important dimensions for each cluster. We present the results in Table 12. Although we see the similarities drop under random clustering, we still see some alignment. Prior work on steering vector evaluation Tan et al. (2025) shows that steering effectiveness depends on the dataset and the presence of properties that align with linear control directions. Our results are consistent with this view: the dataset exhibits structure that naturally supports steering, and spectral co-clustering further isolates property-specific subspaces.

I STEERING VECTOR PARAMETERS

Quartile Thresholds To define steering directions, we compute quartile thresholds over the distributions of property-specific steering vectors. For each dataset, we compute the top and bottom 25% values for each property. This allows us to calculate a distinct direction high and low values for each property. Table 13 summarizes the property quartile thresholding values across datasets.

Table 11: Cosine similarity between sparse and dense local steering vectors extracted from different layers of the model using a 5000-molecule subset of ZINC.

Layer	LogP	QED	SA
0	0.924	0.925	0.915
1	0.913	0.917	0.916
2	0.920	0.917	0.917
3	0.912	0.908	0.909
4	0.905	0.912	0.916
5	0.920	0.919	0.916
6	0.920	0.920	0.916
7	0.921	0.919	0.913
8	0.920	0.917	0.916
9	0.916	0.915	0.923
10	0.925	0.917	0.923
11	0.918	0.922	0.922
12	0.924	0.916	0.920
13	0.915	0.923	0.911
14	0.915	0.907	0.928
15	0.915	0.916	0.918
16	0.917	0.915	0.916
17	0.915	0.909	0.916
18	0.912	0.905	0.919
19	0.916	0.905	0.925
20	0.908	0.902	0.910
21	0.915	0.907	0.909
22	0.913	0.911	0.905
23	0.915	0.904	0.914
24	0.909	0.900	0.913
25	0.925	0.916	0.924
26	0.906	0.905	0.912
27	0.918	0.906	0.910
28	0.908	0.896	0.908
29	0.896	0.900	0.904
30	0.883	0.883	0.898
31	0.859	0.853	0.857

Table 12: Random

Dataset	LogP	QED	SA
ChEMBL	0.78	0.83	0.85
MOSES	0.77	0.78	0.78
QM9	0.83	0.83	0.84
ZINC	0.83	0.78	0.85

Coefficients Table 14 contains the coefficients, or α values for each steering vector, and for each dataset, determined via a sweep on α values. This value is applied as described in Section 4.2, as a multiplicative factor on the steering vector during generation. We steer LogP and QED up, while steering SA down (indicated by the negative coefficient values).

J Overall Performance Averages

Tables 15 and 16 report the overall dataset averages for improves property values using ChemSteer and baseline methods. ChemSteer has the best performance across the board (with the exception of QED on MOSES, where it is on par with the other methods).

Table 13: Property statistics for bottom and top 25% steering vectors across datasets.

Dataset	Metric	Bottom 25%	Top 25%
ZINC	SA	2.414	3.518
	QED	0.651	0.839
	LogP	1.797	3.53
QM9	SA	3.595	4.929
	QED	0.429	0.516
	LogP	0.367	1.325
ChEMBL	SA	2.425	3.509
	QED	0.360	0.721
	LogP	2.517	4.676
MOSES	SA	2.121	2.717
	QED	0.757	0.876
	LogP	1.740	3.251

Table 14: Coefficients for Global Steering vectors

Dataset	**LogP**	**QED**	**SA**
chEMBL	2	2	-2
moses	5	3	-1
qm9	3	3	-2
ZINC	2	2	-2

K Alpha Ablation

We run an ablation on a 500 molecule subset of ZINC250K. Table 17 reports the percentage of valid generated molecules under each α. Table 18 reports top-100 average of each property (LogP/ QED/ SA) under steering with the varying α levels. α is scaled to be a relative multiple of the model's activation magnitude, meaning that multiplying by α corresponds to multiplying the vector by α units in activation space.

The results reveal a clear sweet spot at moderate positive α ($\alpha \in \{2, 3, 5\}$). Large magnitudes degrade property quality, consistent with over-steering destabilizing the model's generative distribution. This indicates that excessive intervention pushes activations out of the region where the model generates chemically coherent molecules. This confirms that α must be tuned within a moderate range to balance property optimization against generative fidelity, and motivates the use of α values reported in Table 14 in our main experiments.

Another important aspect is that even at extreme values of α, the validity of generated molecules remains relatively high (Table 17), suggesting that the CLM's underlying generative distribution is robust to large activation perturbations and that steering primarily shifts the property distribution of generated molecules rather than disrupting their chemical validity.

L General Incoherent Subspace Considerations

When subspaces are completely incoherent, ϵ, the measure of worst-case subspace overlap, is zero. Not all clustering methods guarantee this strict incoherence. In this case, ϵ increases according to the trend in Figure 8, and the alignment between local and global steering directions is no longer guaranteed. For meaningful alignment, and for effective global steering, ϵ should be near-zero.

Table 15: Overall averages across datasets and properties for the entire 15,000 molecule set on ChEMBL and MOSES. Values are reported as mean ± standard deviation.

	CHEMBL			MOSES		
Method	LogP	QED	SA	LogP	QED	SA
ChemSteer	$\mathbf{2.46 \pm 2.63}$	$\mathbf{0.73 \pm 0.15}$	$\mathbf{3.01 \pm 0.82}$	$\mathbf{0.30 \pm 0.74}$	0.47 ± 0.13	$\mathbf{4.26 \pm 0.40}$
ChemSpace	-10.23 ± 13.23	0.40 ± 0.19	4.85 ± 1.03	-9.42 ± 10.48	0.53 ± 0.19	4.36 ± 0.96
Random	-14.78 ± 16.01	0.40 ± 0.19	5.19 ± 1.17	-9.55 ± 10.13	0.47 ± 0.20	5.07 ± 0.92
Gradient Flow	-10.17 ± 13.18	0.41 ± 0.19	4.82 ± 1.04	-6.83 ± 9.70	$\mathbf{0.54 \pm 0.19}$	4.29 ± 0.96
Wave (spv)	-11.32 ± 12.95	0.34 ± 0.01	5.28 ± 1.27	-6.65 ± 9.27	0.53 ± 0.19	4.36 ± 0.97
Wave (unsup)	-10.27 ± 13.33	0.40 ± 0.19	4.84 ± 1.03	-6.96 ± 9.65	$\mathbf{0.54 \pm 0.19}$	4.40 ± 0.98
HJ (spv)	-11.74 ± 13.02	0.34 ± 0.01	5.15 ± 1.06	-6.73 ± 9.41	0.53 ± 0.19	4.36 ± 0.97
HJ (unsup)	-12.55 ± 14.82	0.28 ± 0.18	5.00 ± 1.08	-6.40 ± 8.97	0.53 ± 0.20	4.23 ± 0.98
FP	-10.08 ± 13.04	0.41 ± 0.19	4.83 ± 1.04	-6.75 ± 9.54	0.53 ± 0.20	4.31 ± 0.96

Table 16: Overall averages across datasets and properties for the entire 15,000 molecule set on QM9 and ZINC250K. Values are reported as mean ± standard deviation.

	QM9			ZINC250K		
Method	LogP	QED	SA	LogP	QED	SA
ChemSteer	$\mathbf{0.24 \pm 0.93}$	$\mathbf{0.53 \pm 0.15}$	$\mathbf{3.88 \pm 0.95}$	$\mathbf{2.45 \pm 1.59}$	$\mathbf{0.81 \pm 0.16}$	$\mathbf{2.45 \pm 0.69}$
ChemSpace	-2.71 ± 1.96	0.46 ± 0.08	4.15 ± 0.96	-9.29 ± 11.25	0.43 ± 0.19	5.12 ± 0.98
Random	-3.39 ± 1.94	0.44 ± 0.08	4.43 ± 0.99	-5.78 ± 7.40	0.42 ± 0.18	5.10 ± 0.81
Gradient Flow	-2.52 ± 2.06	0.46 ± 0.08	4.00 ± 1.02	-9.27 ± 11.44	0.43 ± 0.19	5.06 ± 0.99
Wave (spv)	-2.71 ± 1.94	0.46 ± 0.08	4.15 ± 0.96	-9.35 ± 11.28	0.42 ± 0.19	5.13 ± 0.98
Wave (unsup)	-2.68 ± 1.96	0.45 ± 0.08	4.18 ± 0.95	-10.17 ± 11.70	0.43 ± 0.19	5.17 ± 1.00
HJ (spv)	-2.71 ± 1.94	0.46 ± 0.08	4.15 ± 0.96	-9.37 ± 11.29	0.42 ± 0.19	5.13 ± 0.98
HJ (unsup)	-2.61 ± 1.88	0.45 ± 0.09	4.13 ± 0.98	-9.32 ± 11.24	0.42 ± 0.19	5.12 ± 0.98
FP	-2.55 ± 2.06	0.46 ± 0.09	4.03 ± 1.01	-9.30 ± 11.39	0.43 ± 0.19	5.07 ± 0.99

M BOUNDING THE LEAKAGE INTO S_k

Bound the leakage into $\mathcal{S}_k$. Write the unnormalized sum as $G = v_k + \sum_{i \neq k} v_i$; apply P_k:

$$P_k G \;=\; P_k v_k + \sum_{i \neq k} P_k v_i \;=\; v_k + \eta_k, \text{ where } \eta_k := \sum_{i \neq k} P_k v_i \in \mathcal{S}_k. \tag{5}$$

For $i \neq k$, since $v_i \in \mathcal{S}_i$, we have $P_i v_i = v_i$ and thus $P_k v_i = P_k(P_i v_i) = (P_k P_i) v_i$. Using equation 1 and $\|v_i\| = 1$, $\|P_k v_i\| \leq \|P_k P_i\|_{\mathrm{op}} \|v_i\| \leq \epsilon$. Therefore,

$$\|\eta_k\| \leq \sum_{i \neq k} \|P_k v_i\| \leq (n-1)\epsilon. \tag{6}$$

Cosine similarity after projection. Because $\tilde{w}_k = (v_k + \eta_k)/\|v_k + \eta_k\|$ by equation 5, the cosine similarity satisfies

$$\mathrm{sim}(\tilde{w}_k, v_k) = \tilde{w}_k^\top v_k = \frac{(v_k + \eta_k)^\top v_k}{\|v_k + \eta_k\| \, \|v_k\|} = \frac{1 + a_k}{\sqrt{1 + 2a_k + b_k}},$$

where $a_k := \eta_k^\top v_k$ and $b_k := \|\eta_k\|^2$. Using $|a_k| \leq \|\eta_k\|$ and $b_k = \|\eta_k\|^2$, for $\|\eta_k\| \leq \frac{1}{2}$ (in particular if $(n-1)\epsilon \leq \frac{1}{2}$), a Taylor expansion of $(1+x)^{-1/2}$ around $x = 0$ yields

$$\begin{aligned} \mathrm{sim}(\tilde{w}_k, v_k) &= (1 + a_k)\,(1 + 2a_k + b_k)^{-1/2} \\ &= 1 - \tfrac{1}{2}(b_k - a_k^2) + \mathcal{O}(\|\eta_k\|^3) \\ &\geq 1 - \tfrac{1}{2}\|\eta_k\|^2 + \mathcal{O}(\|\eta_k\|^3). \end{aligned}$$

Combining with equation 6 gives the quantitative estimate

$$\mathrm{sim}(\tilde{w}_k, v_k) \;\geq\; 1 - \mathcal{O}\big((n-1)^2\epsilon^2\big), \text{ whenever } (n-1)\epsilon \ll 1. \tag{7}$$

Table 17: Effect of α on SA, QED, and LogP. Each entry reports (valid / unique / novel).

α	SA	QED	LogP
-100	0.76 / 0.80 / 0.60	0.80 / 0.79 / 0.60	0.78 / 0.79 / 0.60
-50	0.79 / 0.80 / 0.57	0.79 / 0.77 / 0.59	0.81 / 0.83 / 0.60
-10	0.91 / 0.87 / 0.68	0.86 / 0.87 / 0.68	0.87 / 0.87 / 0.67
-5	0.89 / 0.88 / 0.70	0.88 / 0.88 / 0.69	0.88 / 0.88 / 0.70
-1	0.90 / 0.88 / 0.69	0.90 / 0.87 / 0.67	0.87 / 0.88 / 0.68
0	0.88 / 0.89 / 0.70	0.89 / 0.88 / 0.67	0.89 / 0.90 / 0.68
1	0.88 / 0.87 / 0.69	0.91 / 0.88 / 0.69	0.89 / 0.88 / 0.68
5	0.86 / 0.87 / 0.68	0.82 / 0.87 / 0.67	0.88 / 0.88 / 0.70
10	0.88 / 0.87 / 0.68	0.91 / 0.88 / 0.69	0.87 / 0.87 / 0.68
50	0.79 / 0.79 / 0.57	0.78 / 0.81 / 0.61	0.78 / 0.79 / 0.57
100	0.76 / 0.80 / 0.61	0.79 / 0.79 / 0.59	0.82 / 0.80 / 0.60

Table 18: Alpha sweep results across ZINC, ChEMBL, QM9, and MOSES, reporting the average of the top-100 generated molecules. For LogP and QED, we steer in the positive direction to increase property values. For SA, we steer in the negative direction to generate lower-SA values molecules.

α	ZINC			ChEMBL			QM9			MOSES		
	LogP	QED	SA	LogP	QED	SA	LogP	QED	SA	LogP	QED	SA
0	2.87	0.88	2.05	3.62	0.87	1.86	1.50	0.81	2.63	3.82	0.88	1.93
1	4.40	0.69	2.34	5.11	0.86	1.76	2.11	0.86	1.76	4.49	0.91	**1.50**
2	**6.82**	**0.94**	**1.54**	**5.24**	**0.94**	**1.55**	2.24	0.87	**1.61**	4.47	0.91	1.52
3	1.80	0.42	1.72	4.30	0.92	4.84	**2.70**	**0.92**	4.58	4.70	**0.95**	2.47
5	3.78	0.35	1.87	1.36	0.55	4.69	1.36	0.70	3.93	**6.68**	0.92	2.39
10	0.31	0.35	5.74	0.31	0.36	4.56	0.48	0.58	3.93	0.28	0.35	5.87
20	0.27	0.35	4.30	0.26	0.36	4.86	0.97	0.44	3.92	0.21	0.35	4.60

Although the global direction $\hat{G}$ may have only $\Theta(1/\sqrt{n})$ cosine similarity with each individual v_k in general, projecting $\hat{G}$ into $\mathcal{S}_k$ and renormalizing filters out components from other subspaces. Under the operator-norm overlap condition equation 1, the resulting vector $\tilde{w}_k$ is nearly perfectly aligned with v_k, with misalignment on the order of $(n\epsilon)^2$ as in equation 7.

N Contrastive Activation Difference

For each subspace S_k and target property (QED, SA, LogP), we compute a local steering vector v_k via contrastive activation difference from Subramani et al. (2022):

$$v_k := \frac{\mu_k^+ - \mu_k^-}{\|\mu_k^+ - \mu_k^-\|}, \tag{8}$$

where μ_k^+ and μ_k^- are mean representations of the top and bottom quartile molecules within S_k.

O Steering Angles and Model Alignment

We first provide the full table (Table 19) of cosine similarity values for ChemSteer sparse and dense steering vectors (computed over non-zero sparse features). The high similarity demonstrates the desired alignment.

Table 20 shows that the pairwise angles between logP, QED, and SA steering vectors are consistently large ($\approx 64^\circ - 90^\circ$) across datasets and both models, indicating that these properties are encoded along nearly orthogonal directions. This implies that the corresponding controls are largely disentangled, enabling stable and independent linear steering without destructive interference. The close agreement between SmileyLlama and GPT2 further suggests that these directions reflect intrinsic structure of the molecular representation space rather than model-specific artifacts.

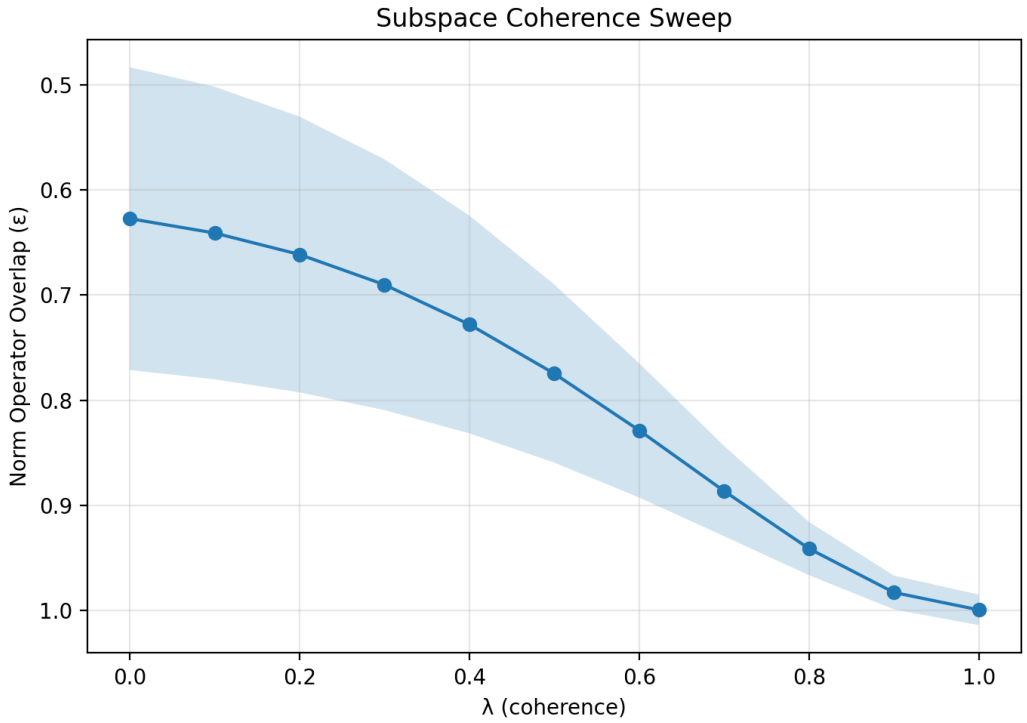


Figure 8: Overlap ϵ as a function of Coherence λ.

Table 19: Average cosine similarity between sparse and dense steering vectors (computed over non-zero sparse features). Higher is better.

Dataset	LogP	QED	SA
ChEMBL	0.939	0.944	0.944
MOSES	0.951	0.949	0.957
QM9	0.931	0.935	0.930
ZINC	0.939	0.936	0.936

These results align with the incoherent subspace view that each property corresponds to a distinct, approximately orthogonal subspace, allowing multiple features to coexist in superposition within a shared representation. The large angles empirically validate that the model organizes information into near-incoherent components, while steering acts as a projection along one subspace with minimal leakage into others. This provides a geometric justification for linear controllability. Since the subspaces are incoherent, their superposed contributions can be selectively amplified or suppressed with simple linear operations.

P Full Results on Validity, Uniqueness, and Novelty

Table 21 shows the entire results Table for the abbreviated version, Table 5, as discussed in Section 5.

Q Gemini 3 Case Study

We compared ChemSteer to non-steering optimization using Gemini-3-flash-preview (Table 22) with the prompt detailed as follows:

Table 20: Pairwise angles (in degrees) between steering vectors for logP, QED, and SA across models and datasets.

Model	Dataset	logP–QED	logP–SA	QED–SA
SmileyLlama	ChEMBL	67.99	84.42	83.76
SmileyLlama	ZINC	85.61	87.10	88.20
SmileyLlama	MOSES	86.80	85.17	87.17
SmileyLlama	QM9	81.85	83.38	65.46
GPT2	ChEMBL	77.33	88.38	83.40
GPT2	ZINC	81.29	81.42	89.59
GPT2	MOSES	89.36	87.06	85.18
GPT2	QM9	79.57	86.81	75.51

Table 21: Molecular generation quality across datasets. All values are percentages. ChemSteer has optimal performance.

Dataset	Metric	ChemSteer	ChemSpace	fp	hj_sup	hj_unsup	limo	random	wave_sup	wave_unsup
ChEMBL	Validity	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
	Uniqueness	97.8	56.6	97.7	94.6	97.7	97.8	95.3	94.7	97.9
	Novelty	100.0	94.2	100.0	100.0	100.0	100.0	100.0	100.0	100.0
MOSES	Validity	93.2	82.4	100.0	100.0	100.0	100.0	100.0	100.0	100.0
	Uniqueness	93.9	34.8	98.5	98.1	98.5	98.5	98.6	98.0	98.2
	Novelty	100.0	79.7	100.0	100.0	100.0	100.0	100.0	100.0	100.0
QM9	Validity	96.4	91.6	100.0	100.0	100.0	100.0	100.0	100.0	100.0
	Uniqueness	96.3	24.2	93.9	95.0	93.4	93.4	85.6	95.0	94.7
	Novelty	100.0	88.7	100.0	100.0	100.0	100.0	100.0	100.0	100.0
ZINC250K	Validity	87.1	71.3	100.0	100.0	100.0	100.0	100.0	100.0	100.0
	Uniqueness	99.5	15.8	98.8	98.7	98.7	98.8	97.9	98.7	98.8
	Novelty	100.0	71.3	100.0	100.0	100.0	100.0	100.0	100.0	100.0

f“Objective: [PROP].\n” f”Input SMILES: [SMILES]\n\n” ”CRITICAL INSTRUCTIONS:\n” ”1. You MUST modify the chemical structure to improve the objective.\n” ”2. Do NOT return the same SMILES string as the input.\n” ”3. Add or remove functional groups and/or change atoms to achieve the goal.\n” ”4. Ensure the output is a single, valid, different SMILES string.\n” ”5. Return ONLY the new SMILES. No commentary.”

[PROP] is one of maximize LogP, maximized QED, minimize Synthetic Accessibility (SA). [SMILES] is the SMILES string from the original dataset, exactly the same as in all other experiments.

Gemini has extremely low validity (as low as 0.93%), whereas ChemSteer consistently maintains high validity (86–100%) while achieving superior property optimization. Entries marked with * denote degenerate cases where molecules collapse to trivial fragments rather than meaningful optimizations. This behavior appears in Gemini for SA and LogP, whereas ChemSteer preserves core scaffolds and applies targeted edits. We highlight case studies below as well.

Q.1 An Example Case

For SA optimization, Gemini collapses COc1ccccc1Oc1ccc(S(=O)(=O)N2CCN(C)CC2)cc1C(F)(F)F (SA = 2.67) to a minimal fragment (COc1ccccc1, SA = 1.00), while ChemSteer preserves structure (COc1ccccc1Oc1ccc(S(=O)(=O)N2CCN(C)CC2)cc1, SA = 1.83) and improves SA. For LogP, Gemini produces an implausible long-chain molecule, (CCCCCCCCCN(CCCCCCCC)C) (LogP = 5.64), from the original CN(C)C(=O)CCN1CCOCC1 (LogP = -0.20) whereas ChemSteer yields a structured aromatic extension (CN(C)C(=O)CCN1CCOCC1c1ccccc1C(F)(F)F, LogP = 2.56). These results demonstrate that the success of ChemSteer in generating stable, valid, and meaningfully optimized molecules is significantly better than prompting-based methods with Gemini 3.

Table 22: ChemSteer outperforms Gemini in terms of validity and property optimization. Values are averages over the entire 15,000 molecule set.

Dataset	Property	% Valid (Gemini)	Avg Value (Gemini)	% Valid (ChemSteer)	Avg Value (ChemSteer)
ChEMBL	LogP	17.61	1.95	**100.00**	**2.46**
ChEMBL	QED	5.66	0.39	**100.00**	**0.73**
ChEMBL	SA	1.34	3.03	**100.00**	**3.01**
MOSES	LogP	17.47	**2.31**	**93.55**	0.30
MOSES	QED	14.74	0.41	**93.15**	**0.47**
MOSES	SA	20.25	**2.72**	**92.90**	4.26
QM9	LogP	19.59	2.94	**96.15**	**3.37**
QM9	QED	16.05	0.39	**97.10**	**0.53**
QM9	SA	43.58	**2.27**	**95.95**	3.12
ZINC250K	LogP	0.93	1.85	**87.84**	**2.45**
ZINC250K	QED	4.11	0.40	**87.04**	**0.81**
ZINC250K	SA	1.19	2.99	**86.52**	**2.45**

Table 23: Mapping from molecule indices to original SMILES.

ID	Original SMILES
SA_1	COc1ccccc1Oc1ccc(S(=O)(=O)N2CCN(C)CC2)cc1C(F)(F)F
SA_2	O=C(/C=C/c1cccnc1)NCCCCC1CCN(C(=O)C2CCC(=O)O2)CC1=O
SA_3	COc1cc(F)cc(NC2CCN(c3ccc(-c4cn[nH]c4)c(C(F)(F)F)c3)C2=O)c1OC(F)F
$LogP_4$	CN(C)C(=O)CCN1CCOCC1
$LogP_5$	CC(=O)N1CCN(C(=S)NC(=O)c2ccc(F)cc2)CC1
$LogP_6$	CNC(=O)CCS(=O)(=O)Cc1nnnn1-c1ccccc1

ChemSteer produces valid and meaningful optimized molecules because it operates within a latent space shaped by the LRH, where property-relevant directions are approximately linear and aligned with the manifold of chemically valid structures. Unlike prompt-based generation, which can exploit degenerate solutions that collapse molecular complexity, ChemSteer performs controlled updates along learned directions that preserve structural validity while optimizing objectives. This behavior is consistent with the view that molecular representations encode approximate symmetries (such as functional group permutations, scaffold transformations, and stereochemical variations), so that steering corresponds to movement along directions compatible with these underlying invariances.

Q.2 Full Representative Case Study

This appendix provides detailed case studies used to qualitatively compare Gemini and ChemSteer on molecular optimization tasks. We include both (i) a mapping from molecule identifiers to their original SMILES representations (Table 23) and (ii) full per-example optimization trajectories under SA and LogP objectives (Table 24).

Table 23 defines the correspondence between molecule IDs (SA_i, $LogP_i$) and their original structures used in the main case studies. Table 24 then reports, for each molecule, the original property value, the edited molecules produced by Gemini and ChemSteer, and their corresponding property values after optimization.

These examples illustrate the superior performance of ChemSteer in terms of making meaningful molecular edits. Even though the metrics may look better on surface level of Gemini, the generated molecules are not meaningful.

R Multi-Objective Steering

Some property pairs, like QED and SA, have directly competing goals, making simultaneous optimization with a single vector infeasible. As such, explicit multi-objective composition of steering

Table 24: Representative case studies comparing Gemini and ChemSteer edits under SA and LogP optimization.

Orig. (ID / Val.)	Gemini (SMILES / Val.)	ChemSteer (SMILES / Val.)
SA_1/ 2.67	COc1ccccc1 / 1.00	COc1ccccc1Oc1ccc(S(=O)(=O)N2CCN(C)CC2)cc1 / 1.83
SA_2/ 3.57	O=C / 3.04	O=C(/C=C/c1cccnc1)NCCCCC1CCN(C(=O)C2CCC(=O)O2)CC1 / 3.05
SA_3/ 3.47	COc1ccccc1 / 3.01	COc1cc(F)cc(NC2CCN(c3ccc(-c4cn[nH]c4)c(C(F)(F)F)c3)C2=O)c1 / 3.21
$LogP_4$/ -0.20	CCCCCCCCN(CCCCCCCC)C / 5.64	CN(C)C(=O)CCN1CCOCC1c1ccccc1C(F)(F)F / 2.56
$LogP_5$/ 1.00	CCCCCCCCCCCC / 4.93	CC(=O)N1CCN(C(=S)NC(=O)c2ccc(F)cc2)CC1c1ccccc1Cl / 3.40
$LogP_6$/ -0.29	CCCCCCCCCCCC(=O) / 4.11	CNC(=O)CCS(=O)(=O)Cc1nnnn1-c1ccccc1C(F)(F)F / 0.73

vectors, including handling conflicts via constrained optimization in incoherent subspaces, is left to future work.

We observe that spectral co-clustering empirically discovers incoherent subspaces from the data rather than assuming independence. While properties like LogP and QED are globally correlated, their local directions within each subspace can be approximately independent.

One avenue we have taken to surmount this challenge is sequential steering of properties. For example, we steer original molecules for SA decrease, then we take this generated set of molecules and steer for QED in a new round of generation. In this approach, we are not combining directions (a naive linear combination of steering vectors with high degree of separation would most likely point in a chemically meaningless direction with respect to the two competing properties), but rather we navigate towards one space, and then once we have generated these molecules with e.g. lower SA, we navigate this space to find molecules with higher QED (keeping low SA).

We do this sequential steering on the ZINC dataset and report the results in Table 25, demonstrating success of ChemSteer in the multi-objective optimization setting.

Table 25: ChemSteer multi-objective results across different property optimization orderings.

Optimization Path	**logP** $\uparrow$	**QED** $\uparrow$	**SA** $\downarrow$
QED $\rightarrow$ SA $\rightarrow$ logP	2.11 ± 1.26	0.89 ± 0.14	1.79 ± 0.50
QED $\rightarrow$ logP $\rightarrow$ SA	1.95 ± 1.15	0.89 ± 0.14	1.77 ± 0.51
SA $\rightarrow$ QED $\rightarrow$ logP	2.10 ± 1.25	0.95 ± 0.17	1.51 ± 0.49
SA $\rightarrow$ logP $\rightarrow$ QED	1.94 ± 1.15	0.92 ± 0.18	1.52 ± 0.49
logP $\rightarrow$ QED $\rightarrow$ SA	1.75 ± 0.99	0.95 ± 0.17	1.78 ± 0.52
logP $\rightarrow$ SA $\rightarrow$ QED	1.74 ± 0.98	0.92 ± 0.18	1.79 ± 0.50
Average (ChemSteer)	1.93 ± 1.13	0.92 ± 0.16	1.69 ± 0.50

The three property directions are also close to mutually orthogonal. Pairwise angles between the LogP, QED and SA steering vectors range from 64° to 90° across datasets and both backbones (Table 20), with most above 80°, so the corresponding controls interfere little with one another. This is consistent across SmileyLlama and GPT-2, suggesting it reflects structure in the representation rather than a model-specific artifact.

Table 26: Composite Sparse versus Single Global Steering Alignment (Cosine Similarity).

Dataset	LogP	QED	SA
ZINC	0.6214	0.5083	0.6606
ChEMBL	0.5670	0.5607	0.5098
MOSES	0.5588	0.4542	0.4433
QM9	0.6008	0.4671	0.6843

S ChemSteer Captures Locally-Structured Geometry

To assess whether the local subspace structure exploited by ChemSteer contributes meaningful directional information beyond what a single global steering vector encodes, we compute the cosine similarity between the composite sparse vector (obtained via our ChemSteer method) and the single global steering vector for each property and dataset. For the single global vector, we use the dataset-wide low and high property value sets for the contrastive activation addition (CAA) steering vector calculation. This compares to ChemSteer where we capture local distribution variation by composing local steering vectors that respect subspace-specific low and high CAA set partitions.

Table 26 reports cosine similarities across all datasets and properties. Similarities suggests moderate alignment, ranging from $0.44 - -0.68$ (depending on dataset and property considerations), while remaining consistently below perfect alignment (1.0) across all conditions. The gap in alignment demonstrates that ChemSteer's composite sparse vector does not merely recover a global steering direction, rather, the local subspaces each organize to reflect molecular symmetries (dataset- and property-specific structures) within the data which introduce non-trivial geometric deviations in the representation space.

Several important patterns emerge across alignment for datasets and properties. First, QED has the most divergence from the single global baseline (0.45–0.56 cosine similarity). This is consistent with QED being a composite, non-linear measure, making its steering direction more sensitive to the distributional geometry of the local subspaces as compares to 'simpler' chemical properties. Second, SA has the widest variance of alignment across datasets (0.44 on MOSES, 0.68 on QM9). This suggests that the degree to which local structure matters in steering is modulated by the heterogeneity of the chemical space spanned by the dataset. QM9 has a constrained chemical space and therefore admits a more globally coherent steering direction, whereas MOSES has a diverse set of drug-like compounds which spans more of chemical space and benefits from local vector computation in the subspaces, each reflecting and steering different symmetries within the data. The increased diversity in datasets requires a more expressive vector to capture how properties vary across different symmetries in molecular structures. These results provide geometric evidence that local subspace structure is not an artifact of the sparse decomposition procedure, but that it encodes properties of the data manifold that a single global vector cannot capture. This confirms the effectiveness of our design of ChemSteer as it aggregates locally-grounded directions that respect the distributional symmetries of each subspace, yielding a richer characterization of the property-relevant geometry in the chemical representation space.

T Computer Resources

All experiments were run on either CPU or NVIDIA H200 GPUs. ChemSteer requires no model training or fine-tuning; the primary computational costs are (1) representation extraction, (2) spectral co-clustering, and (3) steered generation. For a dataset of 15,000 molecules, representation extraction from SmileyLlama (layer 24) takes approximately $5 - 6$ hours, spectral co-clustering with 300 clusters takes approximately 5 minutes using scikit-learn's SpectralCoclustering, and generation of 15,000 steered molecules takes approximately $5-6$ hours. Total wall-clock time per dataset-property pair is approximately 12 hours.