AETHERMOORE RESEARCH PACKET

Layered-Lattice Molecular AI: Beatable Targets

A narrow competitive lane for robust molecular reasoning under topology loss, activity cliffs, and explanation pressure.

chemistry molecules lattice activity-cliffs explainability

Abstract. The beatable target is not every molecular AI task. It is robust compound reasoning where topology, descriptors, fragments, spectra, literature, and claim boundaries must remain visible.

Claim boundary. Does not claim benchmark superiority until tested against established chemistry and molecular-ML suites.

Status: Research and product-targeting brief.

Research grade: Research Brief

Review model: Competitive landscape and failure-mode pass; benchmark validation still needed.

PDF | Source note

Source Text

Layered-Lattice Molecular AI: Beatable Targets

Generated: 2026-05-31

Thesis

SCBE should not try to beat the best molecular AI systems everywhere. The beatable target is narrower:

> Robust compound reasoning under topology loss, activity cliffs, scaffold leakage, and explanation requirements.

This matches the SCBE layered-lattice idea: keep multiple graphs/fields underneath the main molecular graph, with weights for atom identity, proton/electron counts, valence, fragments, descriptors, geometry, spectra, literature, and claim boundaries.

What The Best Systems Are Good At

System family	Strength	Why not compete head-on yet
Schrödinger / enterprise modeling suites	Physics-based molecular modeling, enterprise drug/materials workflows.	Huge expert-built stack; beat by bridging/provenance, not by reimplementing.
RDKit/OpenEye/Open Babel/CDK	Cheminformatics parsing, descriptors, fingerprints, file formats, substructures.	Use them as engines; SCBE adds orchestration and explanation.
ChemCrow/Coscientist-style agents	LLM + chemistry tools + planning; Coscientist connects to lab automation.	SCBE should stay computational-only unless expert-reviewed lab integrations exist.
GNN/graph-transformer molecular ML	Learned property prediction and molecular representation.	Strong on some benchmarks, but vulnerable to data splits, activity cliffs, and interpretability issues.

System family	Strength	Why not compete head-on yet
TDC/MoleculeNet/GuacaMol/MOSES	Public benchmark ecosystems.	Use these to measure claims; do not call local fixtures official scores.

Current Weak Spots In The Field

1. **Activity cliffs**

- Small structural changes can cause large activity/property changes. - Smooth learned embeddings often struggle because they assume nearby structures have nearby behavior. - This is a direct fit for a multi-field card system that tracks small local differences separately from global similarity.

1. **Topology loss**

- Atom counts/formulas alone do not preserve molecular identity. - Example: ethanol and dimethyl ether both reduce to C2H6O. - SCBE already has a benchmark showing recomposition after an atom-mud step requires descriptors/fragments, not atom bags alone.

1. **Scaffold leakage and split weakness**

- Scaffold splits can overestimate real generalization because "different" scaffolds may still be chemically similar. - SCBE can add explicit split receipts and similarity checks across multiple fields.

1. **Fingerprints remain hard to beat**

- Classical fingerprints/descriptors are still competitive with deep models on many molecular prediction problems. - SCBE should treat fingerprints as one lattice, not a primitive baseline to ignore.

1. **Explainability**

- Molecular GNN explanations are still an active research problem. - SCBE can win by producing transparent move histories: which cards preserved identity,

which were lost, which decided recomposition.

SCBE Layered-Lattice Model

Represent each compound as several connected fields:

```
main graph: atoms + bonds
  -> atom bag lattice: element counts, isotope/proton/neutron/ele
  -> descriptor lattice: MW, LogP, TPSA, HBD/HBA, ring counts, fr
  -> fingerprint lattice: Morgan/ECFP, MACCS, topology fingerprir
  -> fragment lattice: SMARTS/substructures, functional groups, p
  -> electron field: partial charges, HOMO/LUMO/density proxy whe
  -> geometry field: conformers, distances, angles, shape/electro
  -> literature field: PubChem/ChEMBL/BindingDB/papers/patents
  -> governance/proof field: safety decision, source, hash, clair
```

The value is not one graph. The value is a reversible and auditable weaving of fields.

First Things We Can Try To Beat

1. Atom-Mud Recomposition

Already implemented:

```
node packages/cli/bin/scbe.js bench compound-decompose --json
```

Goal: expand from 3 cases to 100 known pairs/classes where formulas are ambiguous and identity requires fragments/descriptors.

Win condition:

- atom-only baseline fails or is ambiguous,
- SCBE layered-lattice recomposition recovers the known compound,
- receipt explains which fields mattered.

2. Activity-Cliff Pair Explanation

Build a fixture from known matched molecular pairs:

```
pair A/B -> high structural similarity -> large property/activity
```

Win condition:

- SCBE identifies the local changed fragment,
- explains why global similarity is misleading,
- labels the result as cliff-risk instead of smoothing over it.

3. Scaffold-Leakage Detector

Given train/test molecule sets:

- compute scaffold split,
- compute fingerprint similarity across split,
- flag pairs that are technically split but chemically near.

Win condition:

- SCBE catches leakage-like pairs that a simple scaffold report misses.

4. Multi-Field Candidate Ranking

Given a compound description and candidate list:

- rank by formula, fragments, descriptors, fingerprints, and literature aliases.

Win condition:

- SCBE ranks the known target above formula-matched or graph-near decoys,
- produces field-by-field explanation.

5. Fresh-Agent Reproduction

Give a fresh weak model only:

```
scbe bench compound-decompose --json
```

Win condition:

- it can explain the atom-mud ambiguity and recover the evidence without hidden context.

6. Bijective Reaction Round Trips

SCBE should treat bijective reactions as a proof surface:

```
compound graph -> decomposition cards -> transform cards -> recor
```

A reaction is "bijective" only inside a declared representation. For example:

- canonical SMILES -> RDKit molecule -> canonical SMILES can be bijective enough for identity if stereochemistry/isotopes are preserved;
- molecule -> atom bag is not bijective, because topology is lost;
- molecule -> fingerprint is not bijective, because many molecules can collide or share features;
- reaction path -> products can be non-bijective if leaving groups, solvent, conditions, or stereochemistry are omitted.

Win condition:

- SCBE can state which transformation is reversible, which is lossy, and which extra cards restore identity.
- The benchmark must include both passing reversible paths and intentionally non-bijective paths.

Next Implementation Branches

1. `feat(chem): add activity cliff fixture`

- Use RDKit fingerprints/descriptors. - Synthetic or public known pairs first. - Produce local benchmark, not public leaderboard claim.

1. `feat(chem): expand atom-mud recomposition corpus`

- Add formula-isomer groups from RDKit-valid molecules. - Compare atom-only vs layered-lattice recovery.

1. `feat(chem): add scaffold leakage detector`

- Input two molecule sets. - Output scaffold split report plus cross-field similarity warnings.

1. `feat(chem): pubchem evidence bridge`

- Resolve name -> CID/SMILES/InChIKey. - Attach source URLs and input/output hashes.

1. `feat(chem): bijective reaction harness`

- Round-trip canonical SMILES, formula, fragments, fingerprints, and atom bags. - Score each transformation as `BIJECTIVE`, `LOSSY_RECOVERABLE`, or `LOSSY_AMBIGUOUS`. - Add reaction-style transforms after identity round trips are proven.

Sources

- Schrödinger platform: https://www.schrodinger.com/platform/
- RDKit: https://rdkit.org/
- OpenEye OEChem TK: https://www.eyesopen.com/medchem-tk
- ChemCrow / LLMs with chemistry tools: https://www.nature.com/articles/s42256-024-00832-8
- Coscientist autonomous chemical research: https://www.nature.com/articles/s41586-023-06792-0
- MoleculeNet: https://moleculenet.org/
- TDC: https://tdcommons.ai/start
- MoleculeNet paper: https://pmc.ncbi.nlm.nih.gov/articles/PMC5868307/
- Activity cliffs limitation benchmark: https://pmc.ncbi.nlm.nih.gov/articles/PMC9749029/
- Scaffold validation limitations: https://www.emergentmind.com/topics/scaffold-based-validation

- FunQG quotient graph limitations summary: https://pure.unileoben.ac.at/en/publications/funqg-molecular-representation-learning-via-quotient-graphs

---

Generated from `research/layered_lattice_molecular_ai_targets_2026-05-31.md` by `scripts/research/build_research_library.py`.