Learned Representations

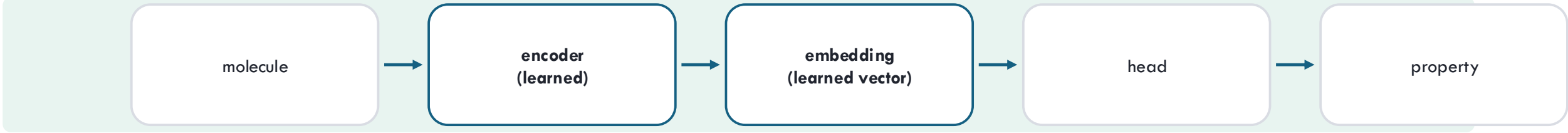
Letting the model choose the features

YOU choose the features



fixed before you see the task

THE MODEL chooses the features



fitted from data — this is the learned representation

J. Schleinitz · Wednesday, August 12, 2026 · RSC 275, Caltech

Plan

- **Why learn a representation?** the label bottleneck (10 min)
- **Unsupervised learning:** clustering, chemical space maps, what they do and do not tell you (20 min)
- **Autoencoders and VAEs:** compression, latent spaces, and an honest look at molecular generation (20 min)
- — break —
- **Self-supervised pre-training:** masked prediction and contrastive learning (15 min)
- **Transfer learning:** frozen, partial, full — and where it genuinely pays (15 min)

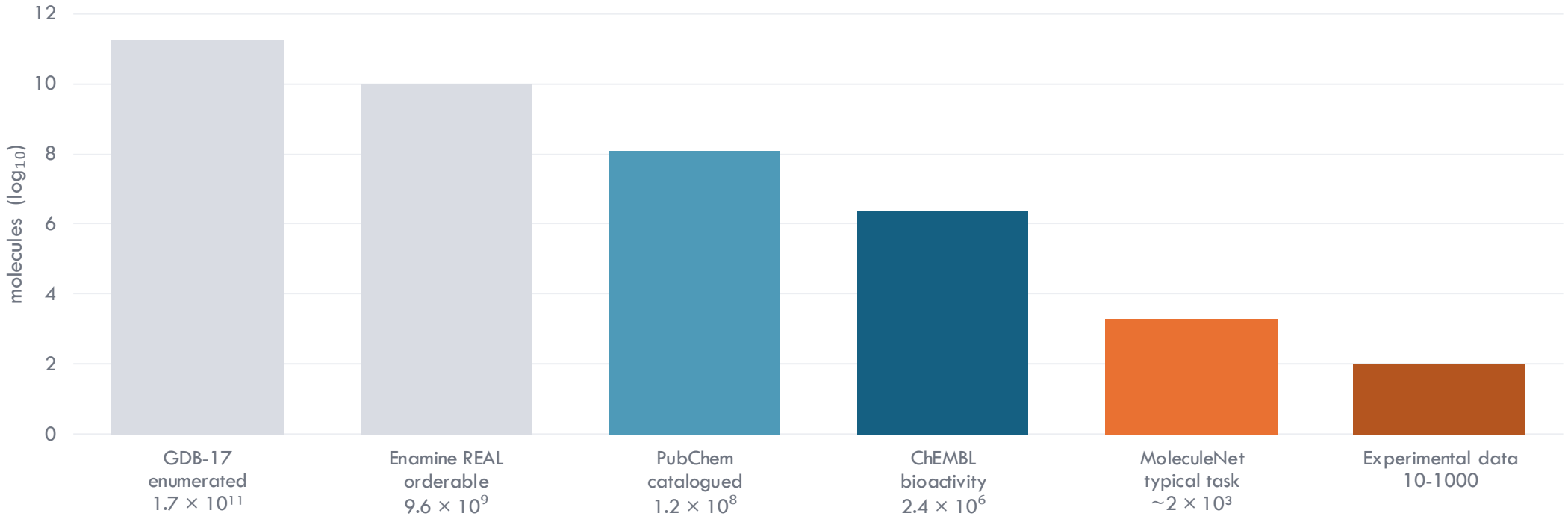
Tutorial at 15:15 — you will build the map and measure the crossover yourself, on real data.

01

Why learn a representation?

- Hand-crafted vs learned features
- The label bottleneck
- Three kinds of supervision

Labels are expensive not inputs



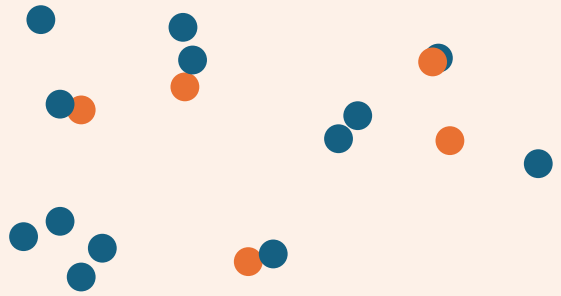
Labels are the scarce resource

GDB-17: Ruddigkeit et al., J. Chem. Inf. Model. 2012, 52, 2864 · Enamine REAL (2026) · PubChem · ChEMBL

Learning modes

SUPERVISED

learn x → y

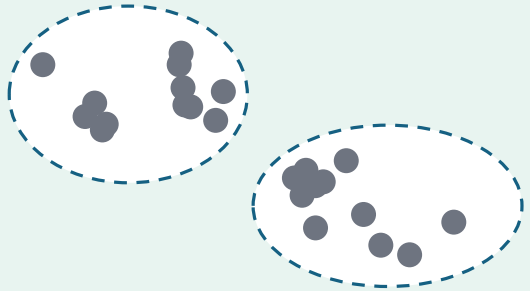


every point carries a label

labels are expensive

UNSUPERVISED

learn the shape of x

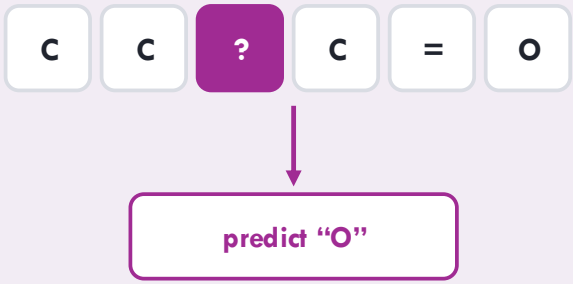


no labels

unlimited data

SELF-SUPERVISED

learn x → part of x



automated labels generation

supervised signal, no annotation

02

Unsupervised learning

- Clustering, dimensionality reduction, density
- Chemical space maps — and how to read them honestly
- VAEs treated separately

Unsupervised learning

1 · CLUSTERING

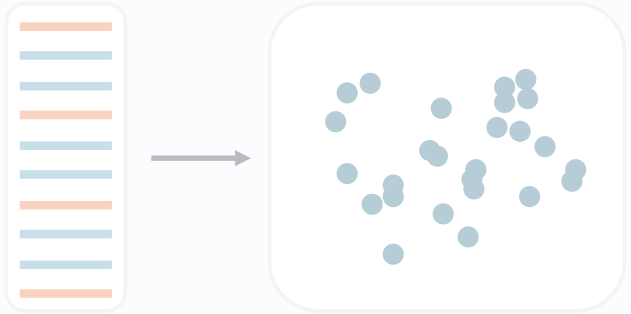
which molecules belong together?



groups, no coordinates

2 · DIMENSIONALITY REDUCTION

can I see 2048 dimensions?



coordinates you can plot

3 · DENSITY & GENERATION

what does a plausible molecule look like?

sample new points from the model

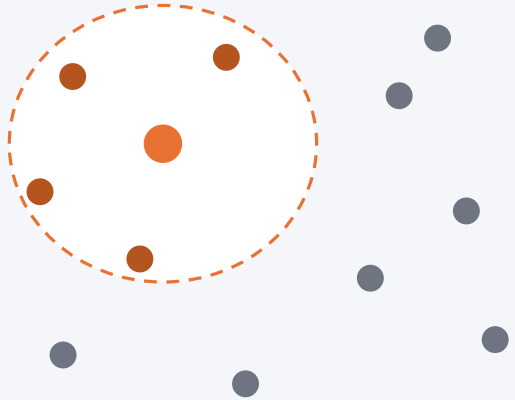
Butina clustering: the cheminformatics default

1 · count neighbours within t



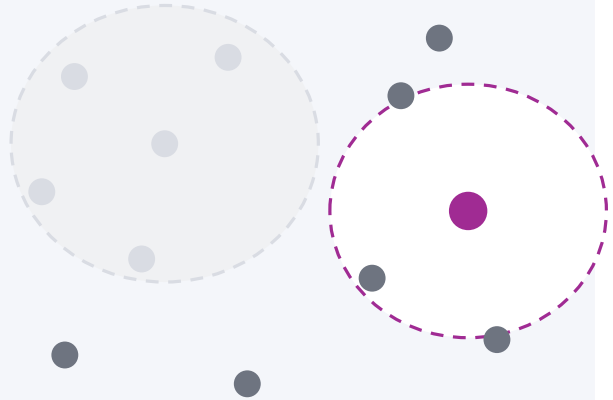
t = Tanimoto cut-off

2 · densest molecule claims its sphere



cluster centre + members

3 · repeat on what is left



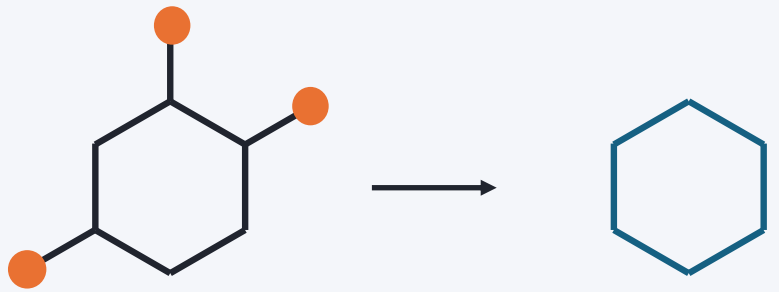
greedy, deterministic, no k to choose

No k to choose, deterministic. Run it at two or three cut-offs and check your conclusion survives.

Butina, J. Chem. Inf. Comput. Sci. 1999, 39, 747

Scaffolds: chemistry is far more clustered than it looks

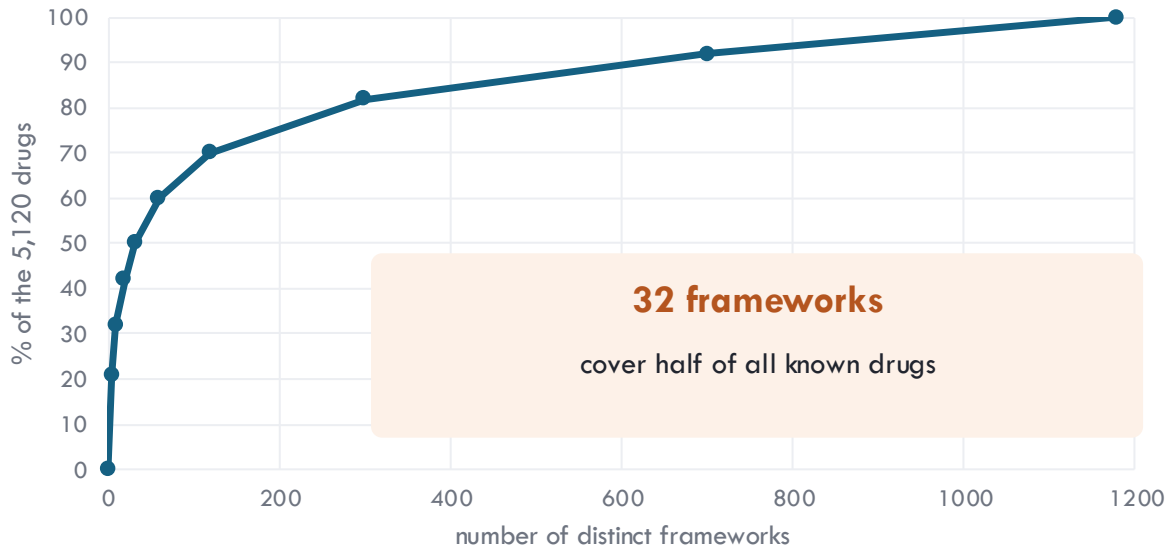
strip side chains → keep the framework



molecule

Bemis–Murcko framework

scaffold split = hold out whole frameworks, not random molecules



1,179 distinct frameworks among 5,120 drugs

32 frameworks cover half of all known drugs.

Bemis & Murcko, J. Med. Chem. 1996, 39, 2887 — 1,179 frameworks among 5,120 drugs

Unsupervised learning

1 · CLUSTERING

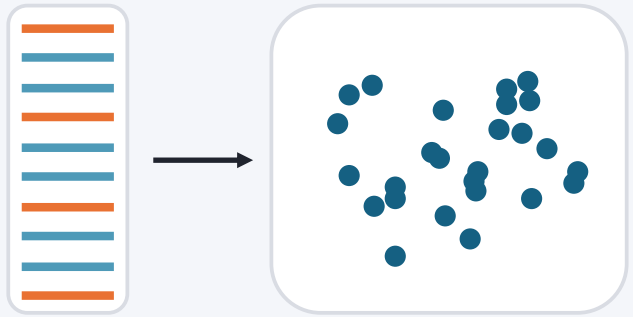
which molecules belong together?



groups, no coordinates

2 · DIMENSIONALITY REDUCTION

can I see 2048 dimensions?



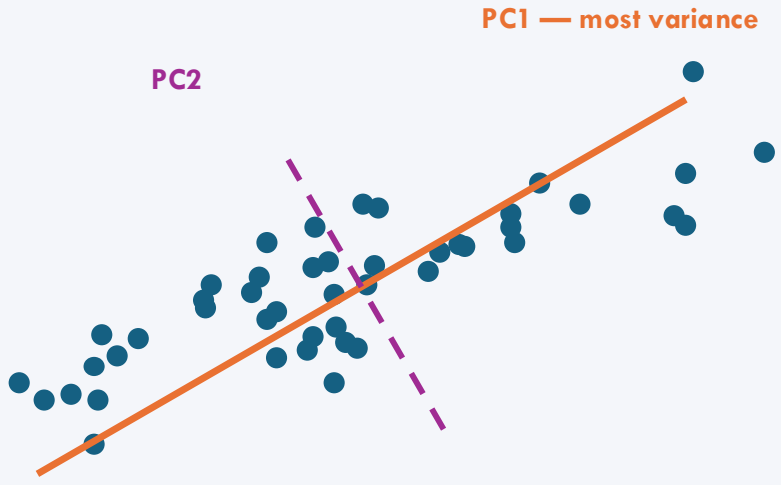
coordinates you can plot

3 · DENSITY & GENERATION

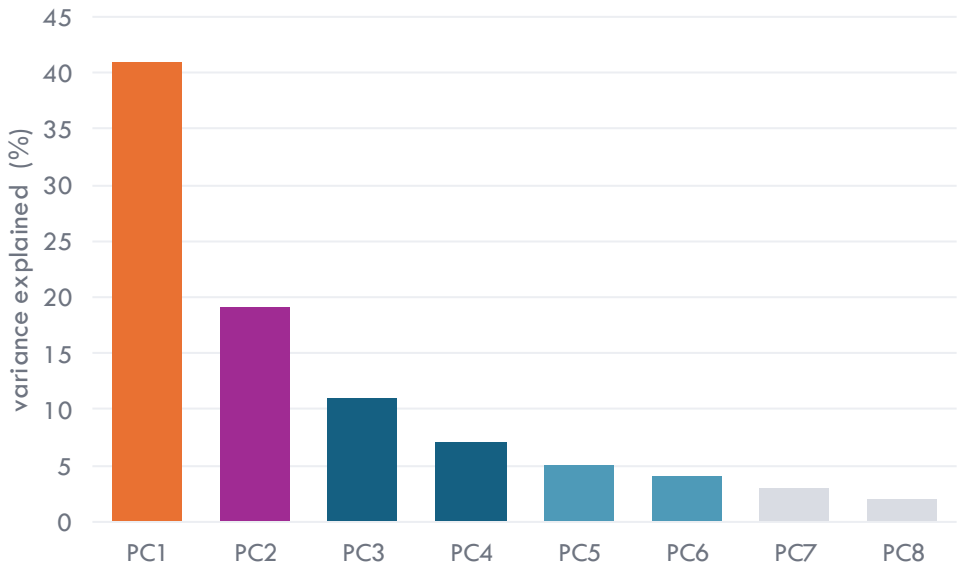
what does a plausible molecule look like?

sample new points from the model

PCA: a rotation, and an honest one



a rotation, nothing more



2 components ≈ 60% of the variance — the rest is not noise, it is chemistry

t-SNE and UMAP: keep the neighbors, lose the distances

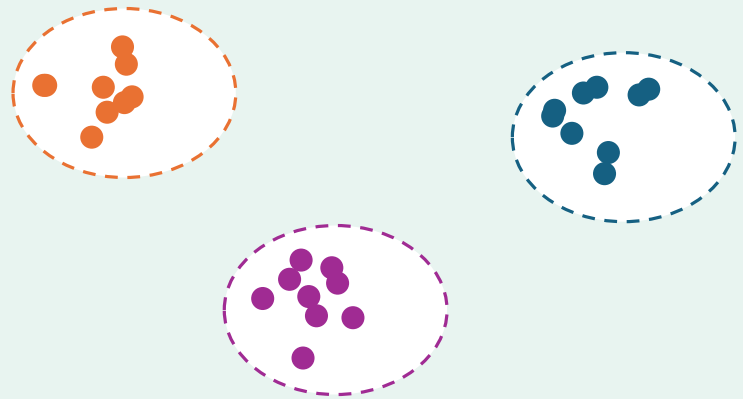
high-dimensional fingerprint space



distances here are mostly meaningless
(everything is far from everything)

t-SNE
UMAP

2-D neighbour-preserving map



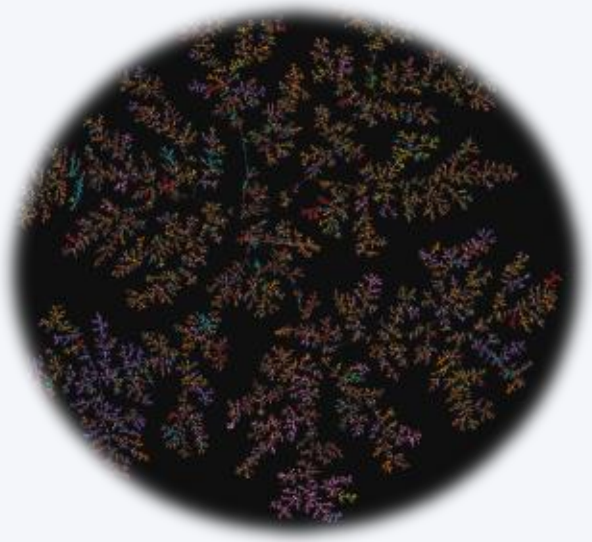
neighbours stay neighbours ✓

gap widths and cluster sizes are not meaningful ✗

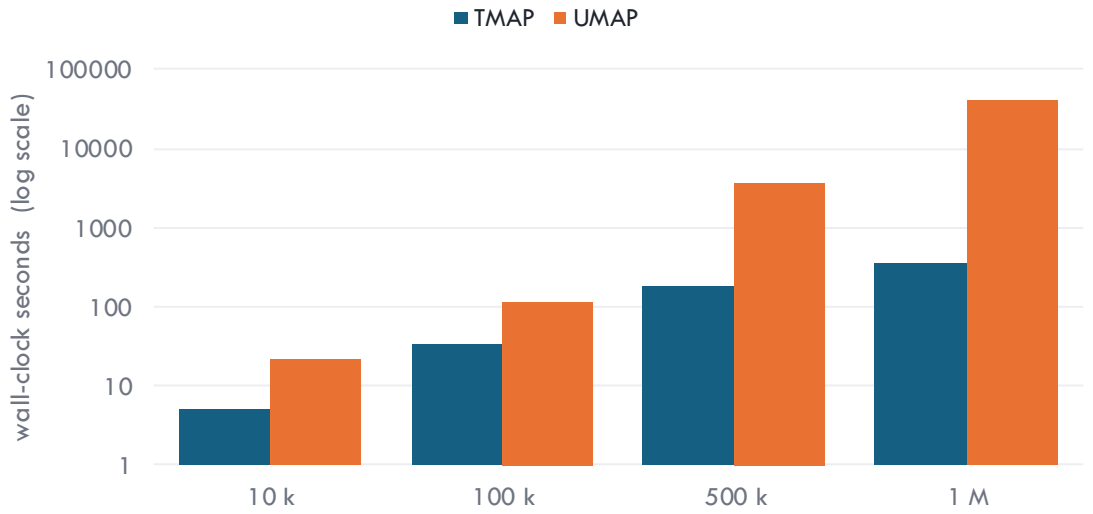
The trade is deliberate: **neighborhoods are preserved, distances are not.**

Algorithm matters at large scale

TMAP: lay out a minimum spanning tree, not a scatter



branches = series of analogues · no overplotting at 10^6 molecules



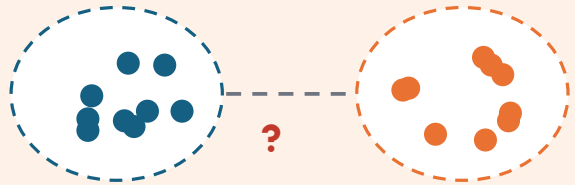
1 M molecules: TMAP 6 min vs UMAP 11.5 h (~117×)

TMAP: ~117× faster than UMAP at 10^6 molecules, and reproducible.

Probst & Reymond, J. Cheminform. 2020, 12, 12; Flores Sepúlveda & Reymond, J. Chem. Inf. Model. 2026, 66, 5595; https://tmap.gdb.tools/src/chembl/chembl.html

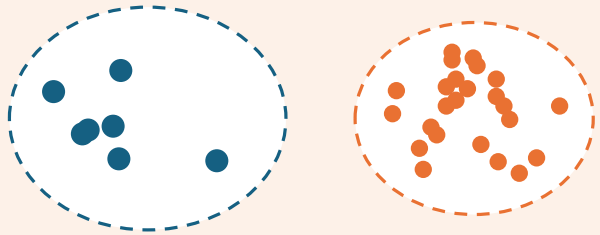
Three things your map is not telling you

✗ gap width means nothing



perplexity changes it

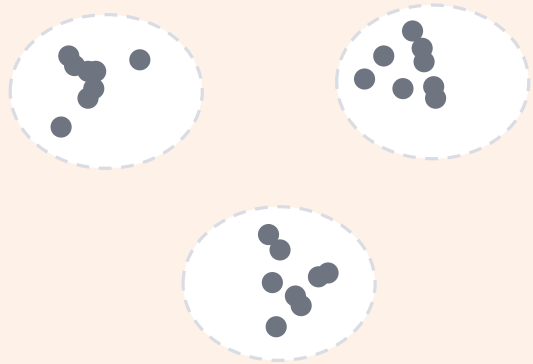
✗ cluster size means nothing



7 molecules **240 molecules**

drawn almost the same size

✗ clusters appear in pure noise



input was uniform random

Never read clusters off a map. Cluster explicitly in the original space, then use the map only to display the result.

Unsupervised learning

1 · CLUSTERING

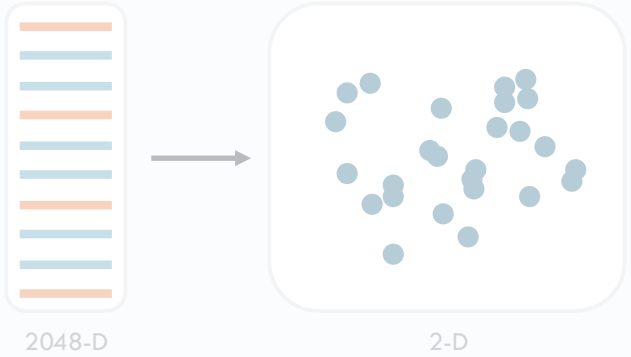
which molecules belong together?



groups, no coordinates

2 · DIMENSIONALITY REDUCTION

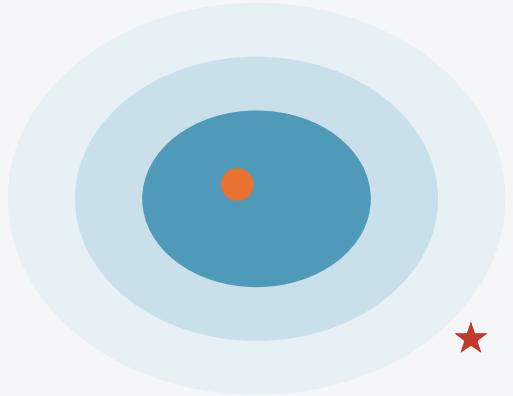
can I see 2048 dimensions?



coordinates you can plot

3 · DENSITY & GENERATION

what does a plausible molecule look like?



sample new points from the model

Case study: the periodic table, from text alone

no chemistry was taught

"… $LiFePO_4$ cathodes show …"

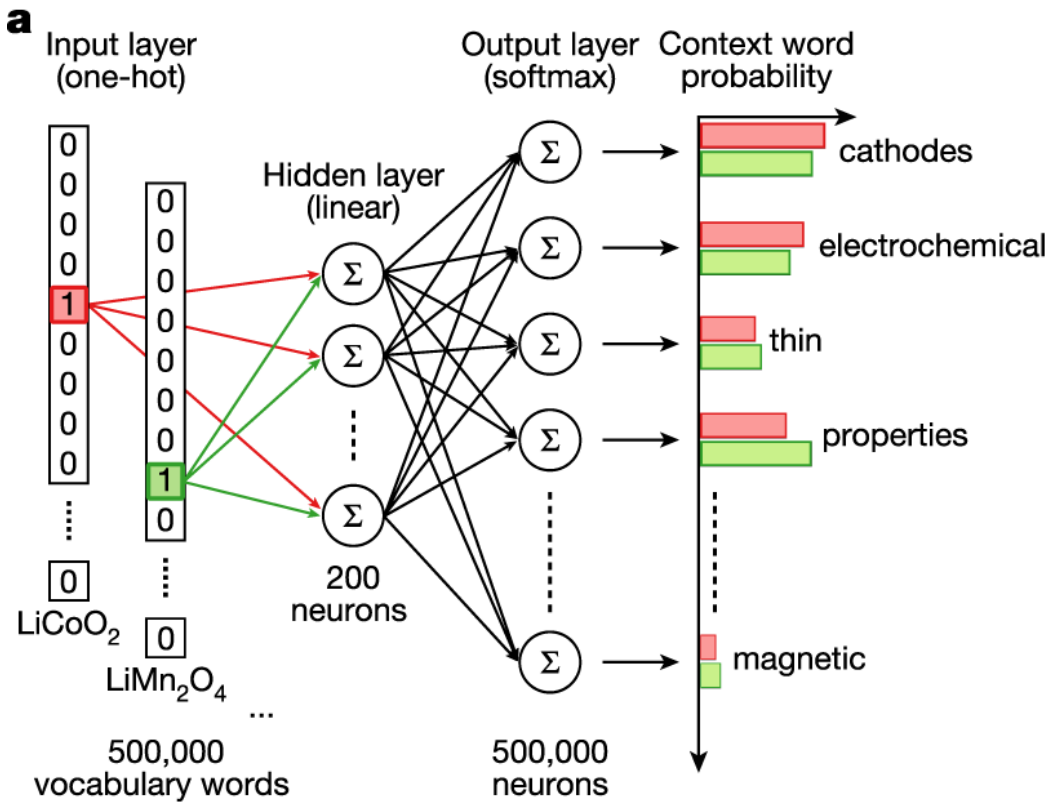
"… we anneal the $CoSb_3$ skutterudite …"

"… Bi_2Te_3 thermoelectric performance …"

3.3 M materials-science abstracts

1.5 M used · word2vec · 200-D

the only signal is which words co-occur



Unsupervised learning on ~1 M abstracts produce a representation.

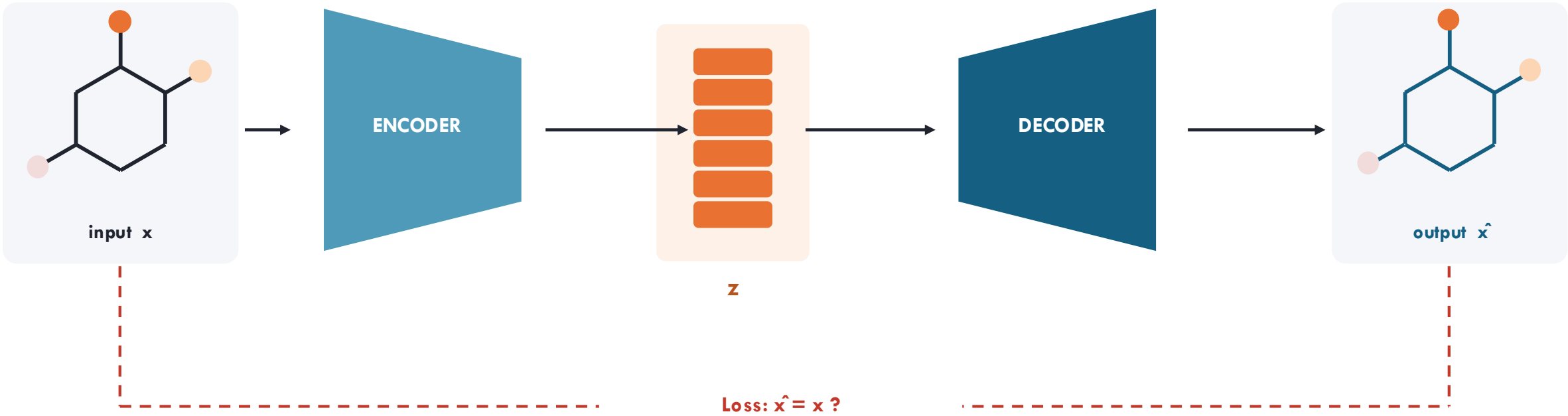
Tshitoyan et al., Nature 2019, 571, 95

03

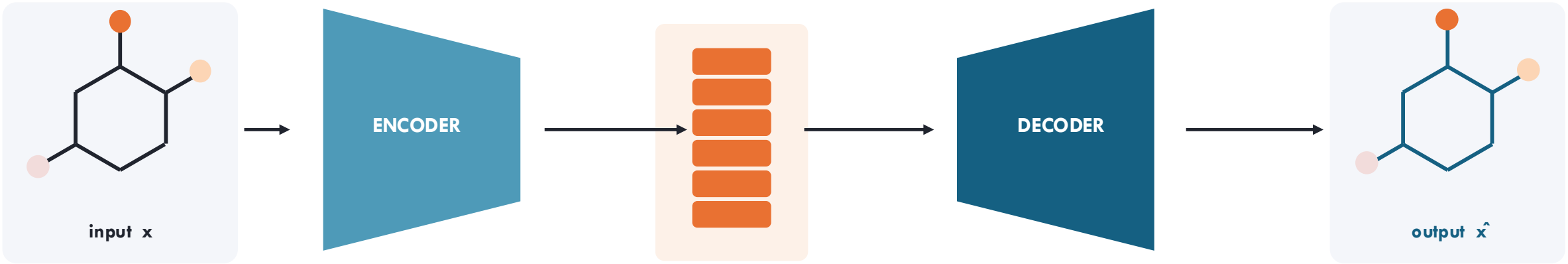
Autoencoders and VAEs

- Compression as representation learning
- Why a plain autoencoder is not enough
- What VAEs are — and are not — good for

The autoencoder: learn by rebuilding the input

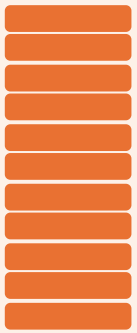


Compressing the representation



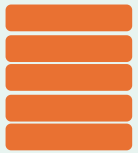
bottleneck too wide

z is as big as x



bottleneck just right

$z \ll x$



bottleneck too narrow

$z \approx 2$



The problem: a plain autoencoder learns islands

Autoencoder latent space



decode

what comes out

C1=CC=CC=C1C(=O)N(C)(C)(C)C	✗
c1ccc2cc(	✗
CC(C)(C)(C)(C)C	✗
N#N#N=O	✗
C1CC1CC1CC1CC…	✗

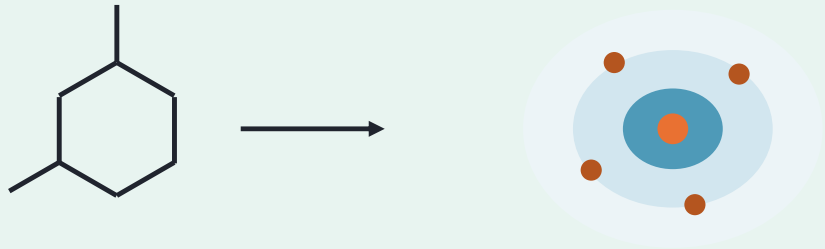
invalid, unparseable, or chemically absurd

The VAE in one idea: encode a cloud, not a point

Autoencoder (AE): x → one point



VAE: x → a small cloud



mean + spread

Two forces, and a dial between them

RECONSTRUCTION

"keep every molecule distinguishable"



pushes clouds APART

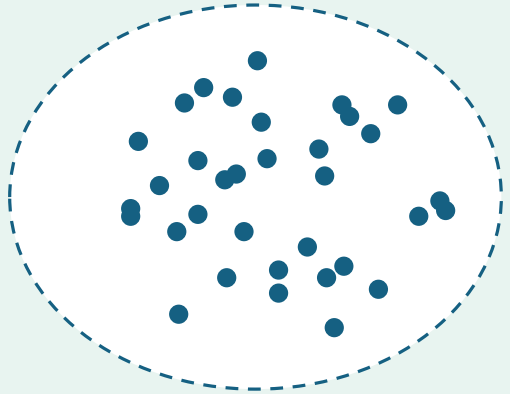
→ islands and holes come back

the VAE loss

reconstruct + β × stay-organised

REGULARISATION

"all clouds should look like one blob"

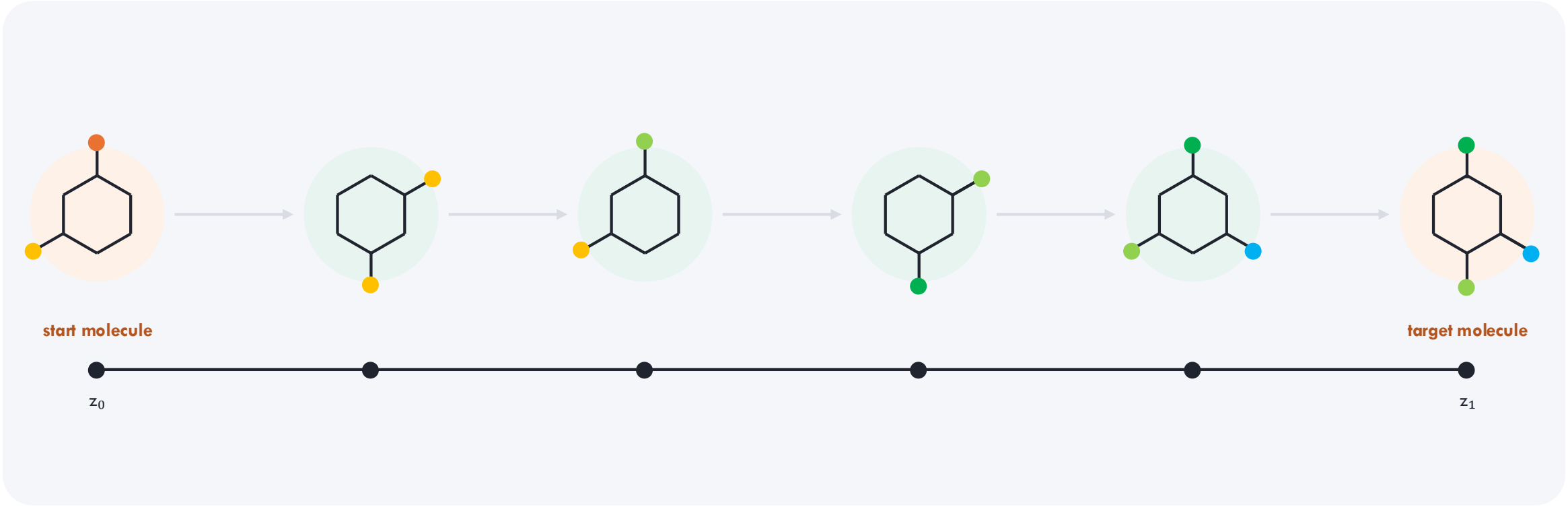


pulls clouds TOGETHER

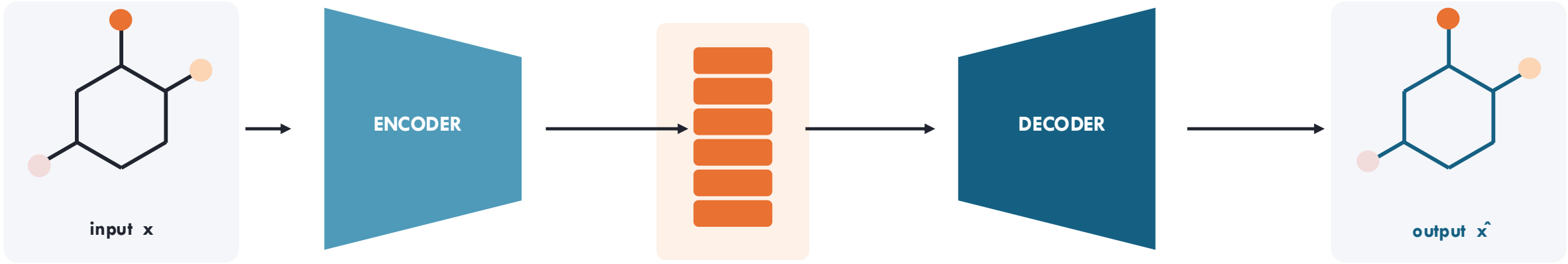
→ one smooth space you can sample

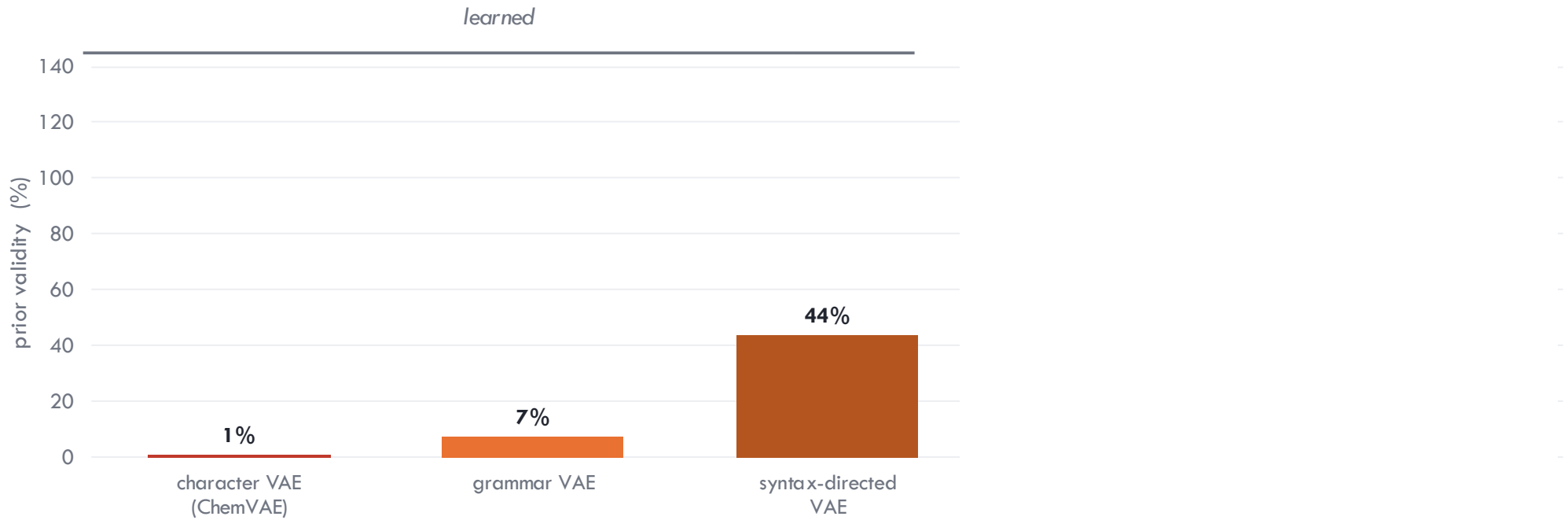
β is the whole design decision. Too small and you are back to islands; too large and the decoder ignores z entirely.

Attempt to build a ”continuous” molecular space



Solving the Validity Issue

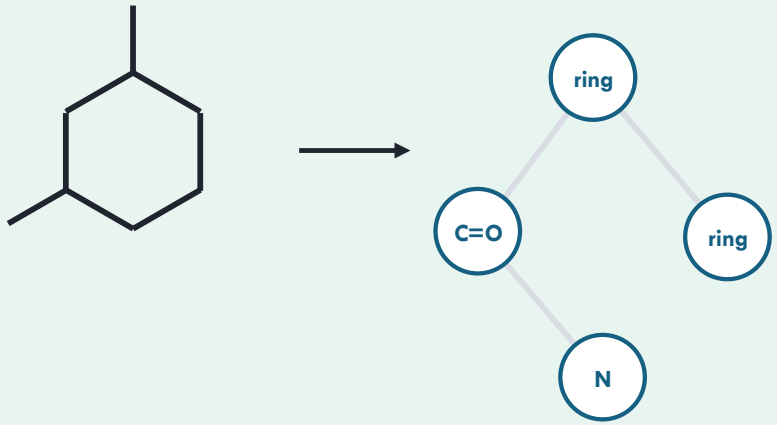




Jin, Barzilay & Jaakkola, ICML 2018, Table 1 · Krenn et al., Mach. Learn.: Sci. Technol. 2020, 1, 045024

Enforcing Validity

JT-VAE — assemble from a vocabulary of valid pieces



junction tree of substructures

780-piece vocabulary · illegal joins are masked out during decoding

100 % valid

SELFIES — a grammar that cannot fail

[F]	[=C]	[=C]	[#N]
X_0	X_1	X_3	X_2
F	F–C	F–C=C	F–C=C=N

asked for a triple bond, valence allowed only a double → clamped, never rejected

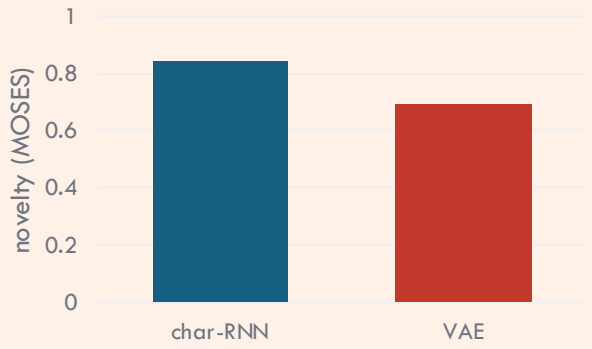
100 % valid · but only local valence is enforced —
not aromaticity, not ring strain, not stability

1 random mutation: SMILES 9.9 % → SELFIES 100 % valid

Both guarantee **syntactic** validity only. A 100%-valid molecule can still be one no chemist would draw.

Validity is not necessarily chemically relevant

it memorises



31 % of what the VAE "designs"
is already in the training set

MOSES benchmark

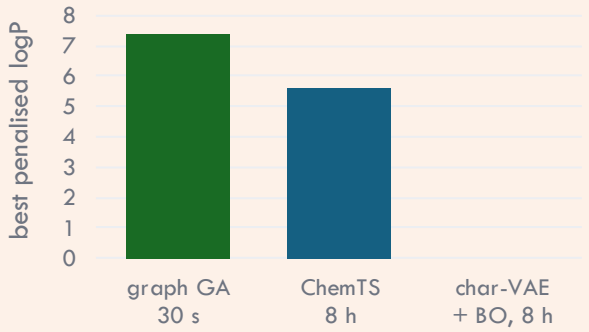
100 % valid ≠ good

model	valid %	FCD ↓
Molecule Chef	99.0	0.73
CGVAE	100.0	11.73
char-VAE	12.0	37.65

the model with perfect validity
matches the real distribution worst

FCD = distance to real chemistry

a genetic algorithm wins

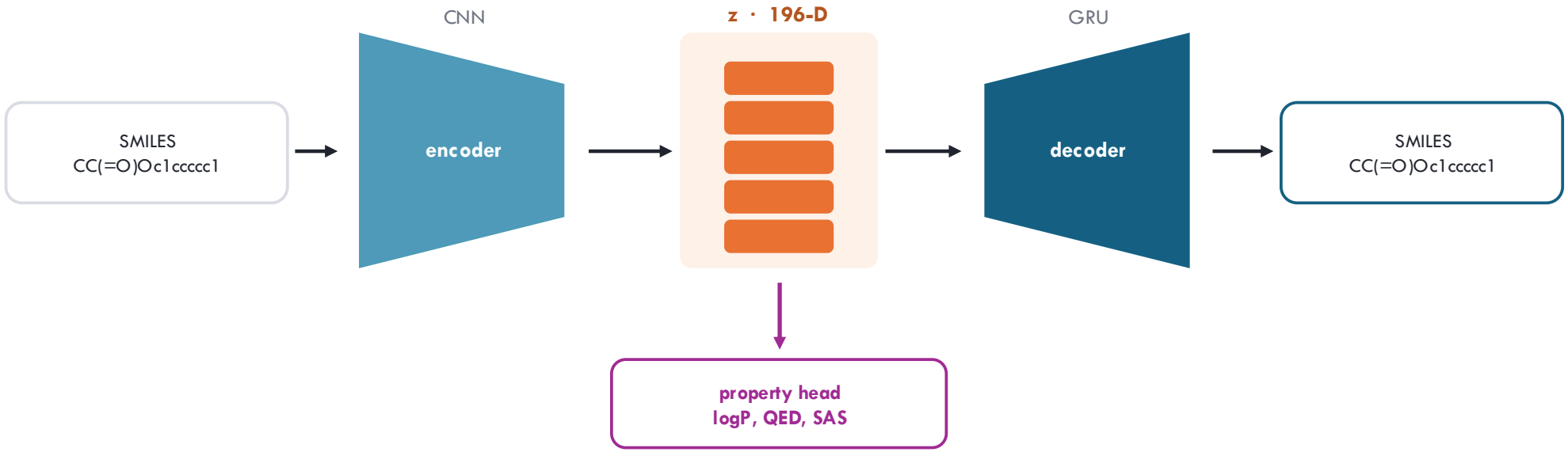


30 seconds on a laptop beats
8 hours of VAE + Bayesian opt.

Jensen, Chem. Sci. 2019

MOSES: Polykovskiy et al., Front. Pharmacol. 2020 · Molecule Chef: Bradshaw et al., NeurIPS 2019 · Jensen, Chem. Sci. 2019, 10, 3567

ChemVAE, 2018



250,000
ZINC molecules, no labels

196
latent dimensions

~4 %
of random z decode to a valid molecule

Joint property training is the idea

Gómez-Bombarelli et al., ACS Cent. Sci. 2018, 4, 268 — validity figures are version- and definition-dependent; see notes

Keep the encoder, throw away the decoder

as a molecule generator

- ✗ beaten by a graph genetic algorithm
- ✗ beaten by an ordinary SMILES LSTM
- ✗ beaten by random screening of ZINC
- ✗ only ~30 % of proposals are synthesisable

the field moved to diffusion, GAs and LLMs
(→ Thursday's lecture)

as a representation learner

- ✓ a fixed-length vector for any molecule
- ✓ trained without a single label
- ✓ a smooth space you can optimise inside
- ✓ the encoder is reusable

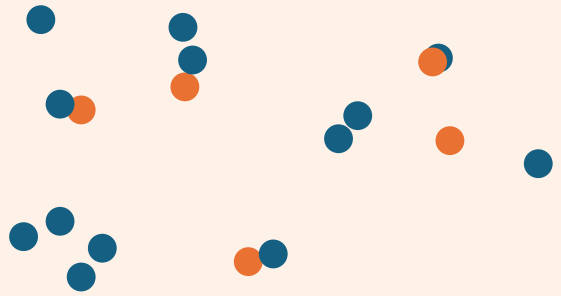
Break

10 minutes — back for pre-training and transfer learning

Learning modes

SUPERVISED

learn x → y

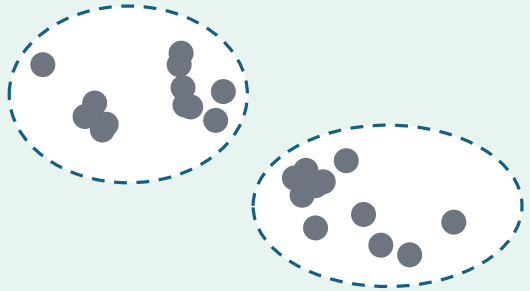


every point carries a label

labels are expensive

UNSUPERVISED

learn the shape of x

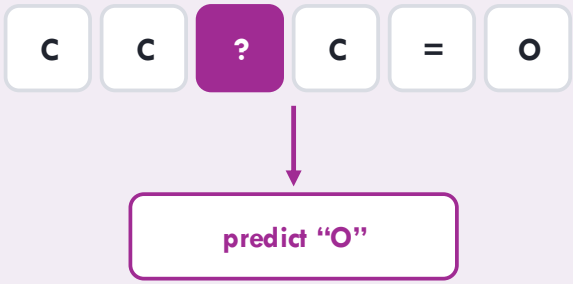


no labels

unlimited data

SELF-SUPERVISED

learn x → part of x



automated labels generation

supervised signal, no annotation

04

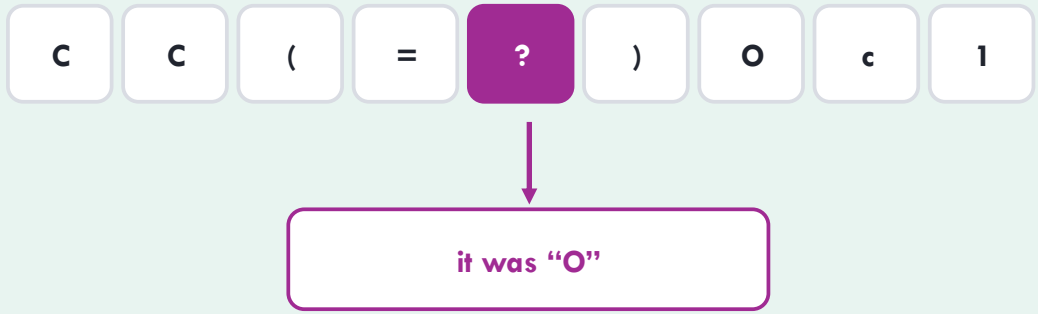
Self-supervised pre-training

- Masked prediction and contrastive learning
- Chemical language models at scale
- Pre-training graph neural networks

Two ways to make the data label itself

A · MASKED PREDICTION

hide part of the molecule, predict it back

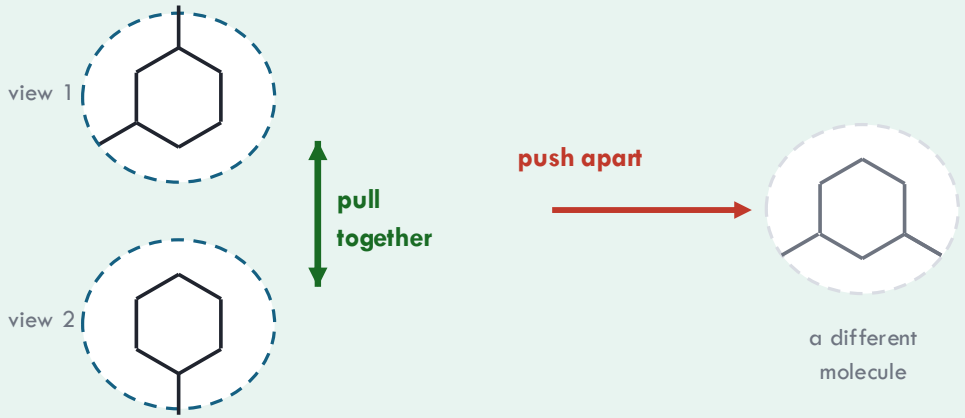


to fill the gap the model must learn valence,
ring closure and functional-group grammar

ChemBERTa · MoLFormer · attribute masking on graphs

B · CONTRASTIVE

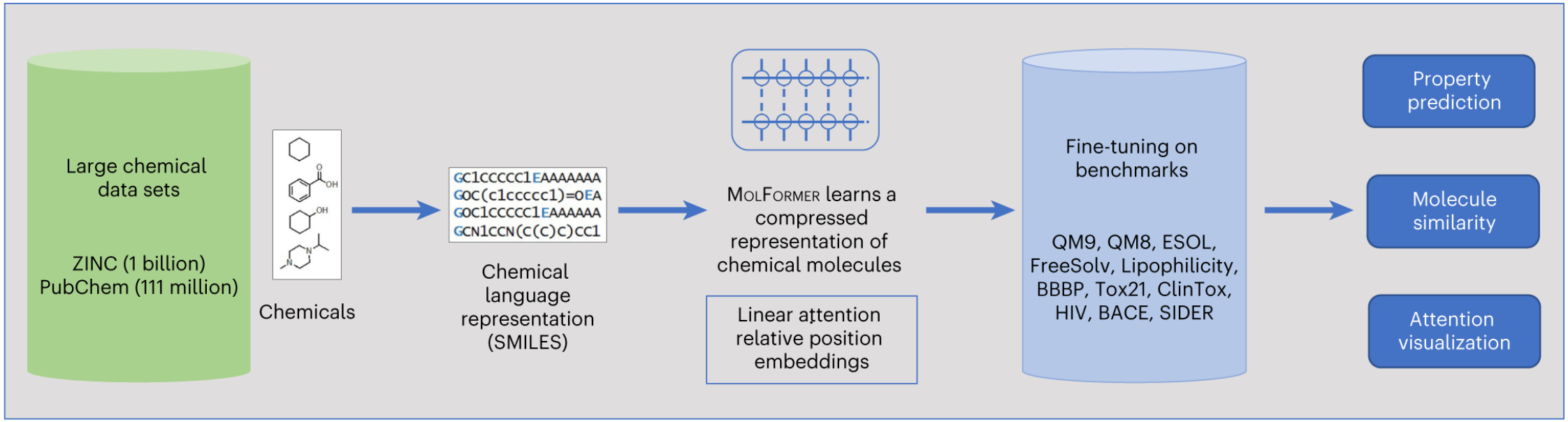
two damaged views of one molecule must agree



MolCLR · GraphCL · InfoMax

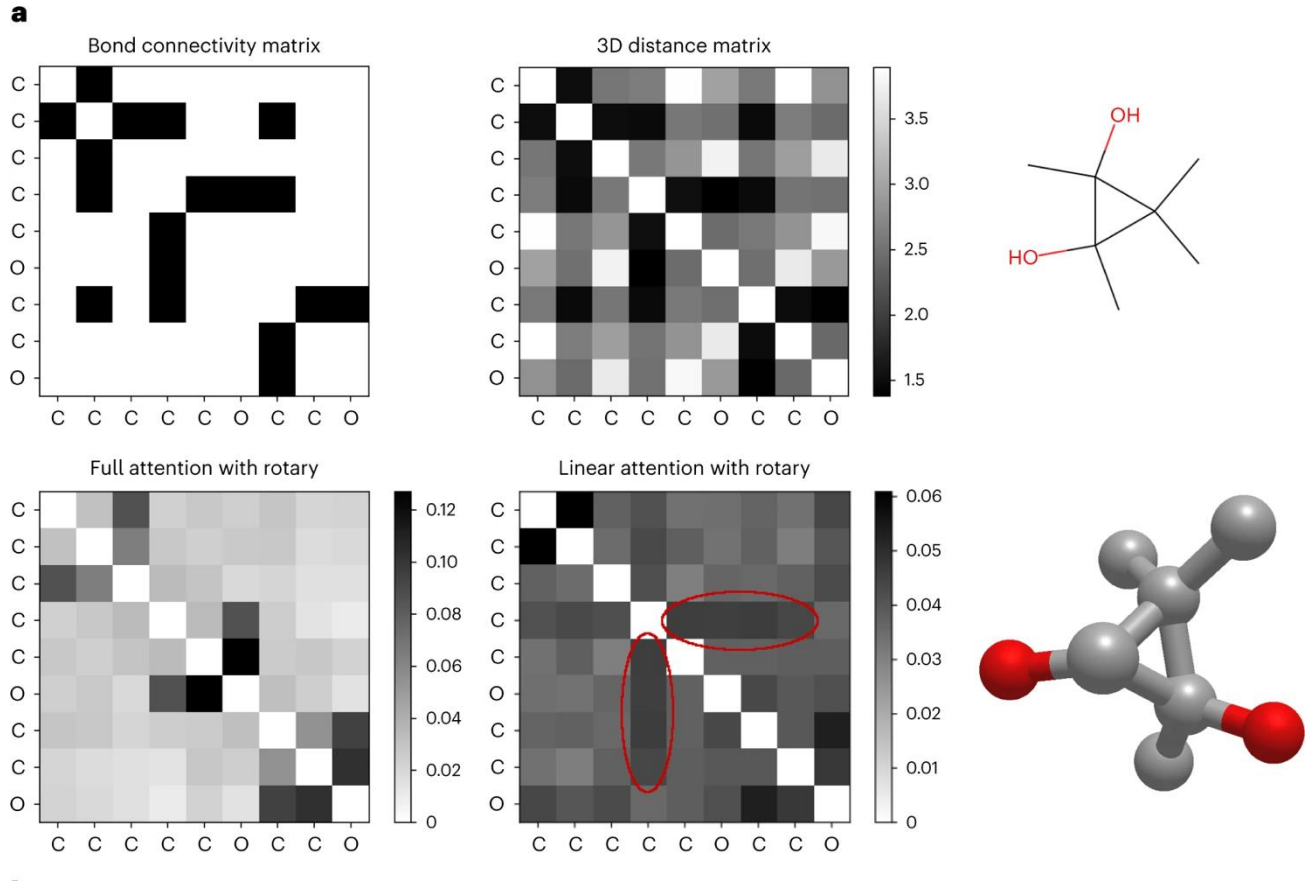
Masked prediction teaches **local chemistry**; contrastive teaches **what to ignore**. Both need only unlabelled molecules.

Pretraining with masked predictions



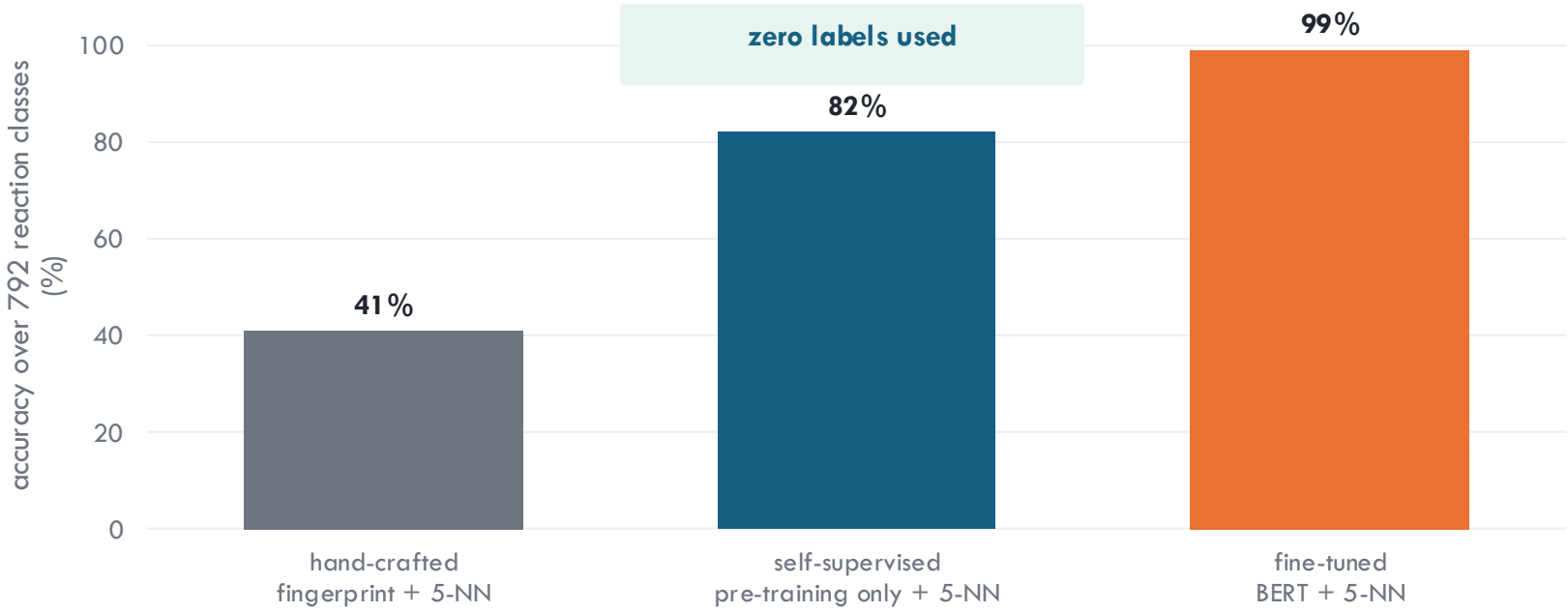
Ross, J., Belgodere, B., Chenthamarakshan, V. et al. Large-scale chemical language representations capture molecular structure and properties. Nat Mach Intell 4, 1256–1264 (2022)

What does the pretraining learns?



Ross, J., Belgodere, B., Chenthamarakshan, V. et al. Large-scale chemical language representations capture molecular structure and properties. Nat Mach Intell 4, 1256–1264 (2022)

Self supervised learning for reaction

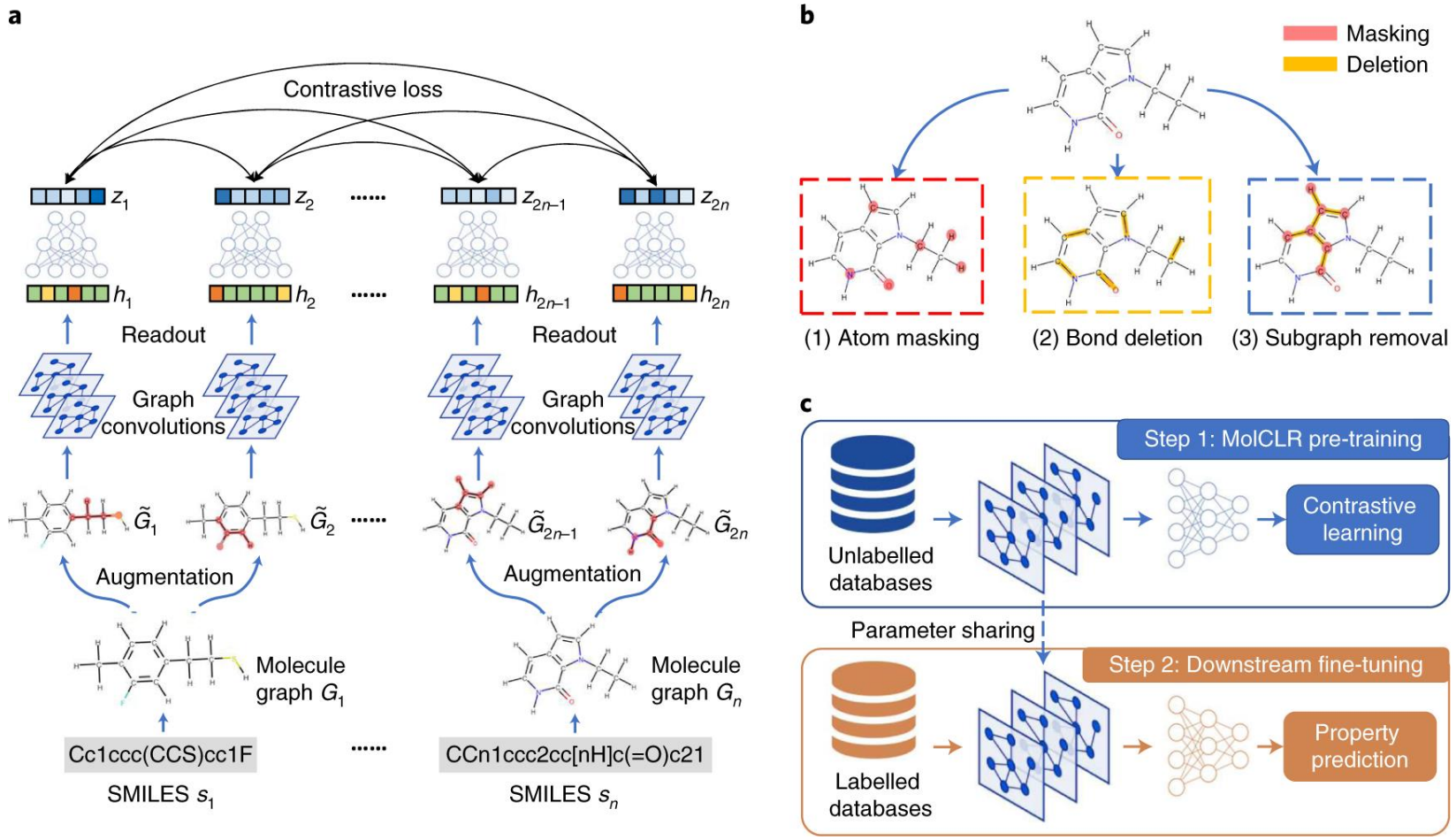


2.6 M unlabelled reactions in · reaction classes come out

Masked-language pre-training alone recovers most of the reaction taxonomy — 81.9% vs 41.0% for a hand-crafted fingerprint, with zero labels.

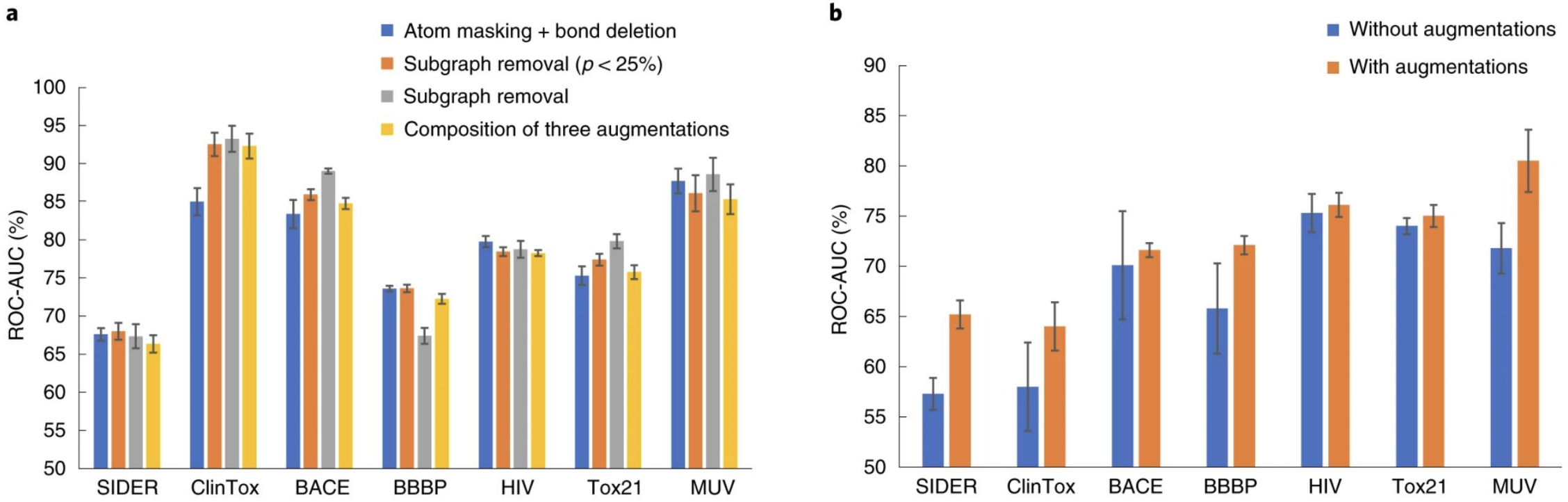
Schwaller et al., Nat. Mach. Intell. 2021, 3, 144 · counterpoint: Probst et al., Digital Discovery 2022, 1, 91 (DRFP)

Choosing augmentations for contrastive learning



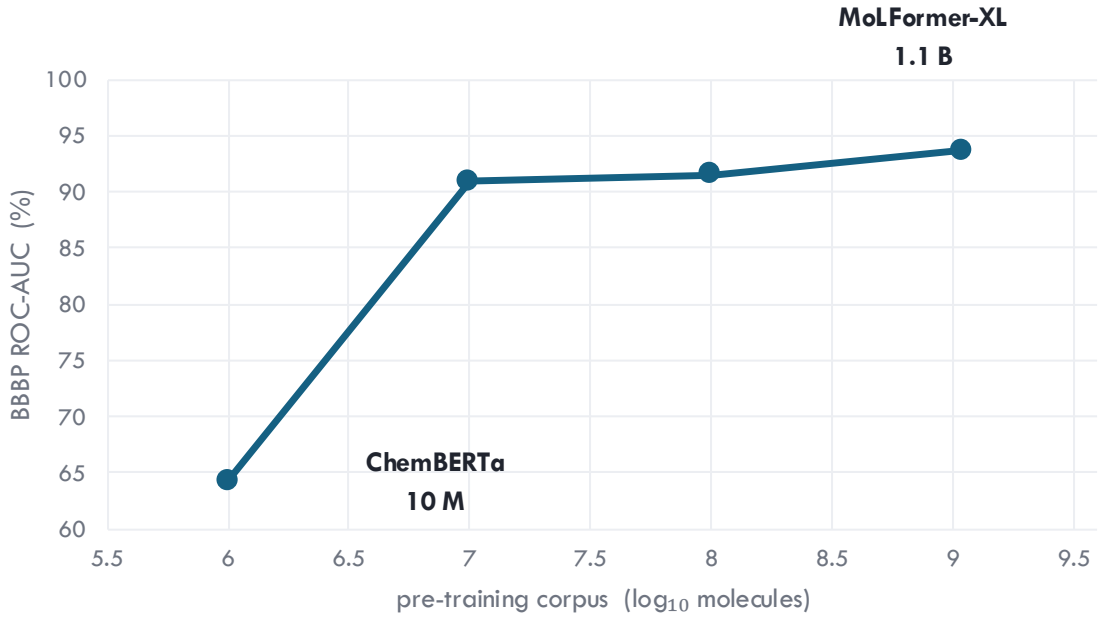
Wang, Y., Wang, J., Cao, Z. et al. Molecular contrastive learning of representations via graph neural networks. Nat Mach Intell 4, 279–287 (2022).

Choosing augmentations for contrastive learning

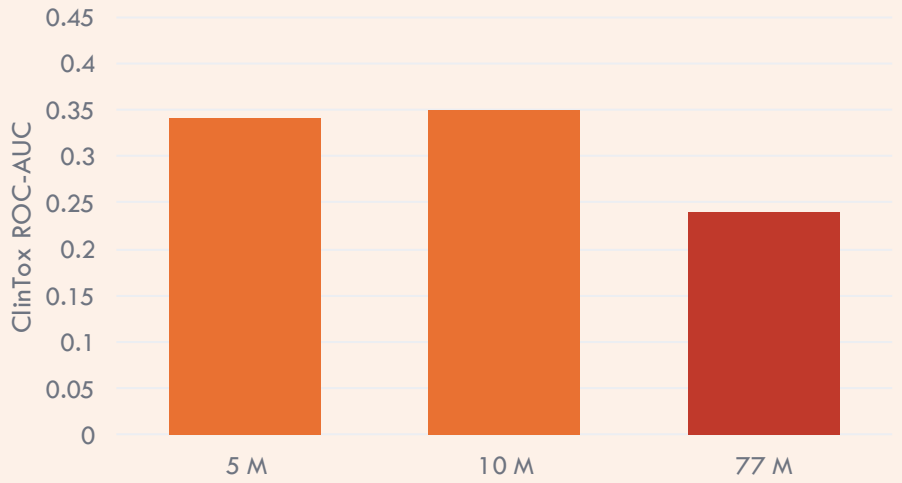


Wang, Y., Wang, J., Cao, Z. et al. Molecular contrastive learning of representations via graph neural networks. Nat Mach Intell 4, 279–287 (2022).

Scale and Diversity Impact Pre-training Performance



1.1 B SMILES on 16 GPUs — linear attention made it affordable



ChemBERTa-2, masked-LM: a 15× bigger corpus made this task worse

corpus composition matters more than corpus size

Corpus composition beats corpus size. A 15× bigger corpus made ClinTox substantially worse; a smaller, more diverse one beat a bigger, narrower one.

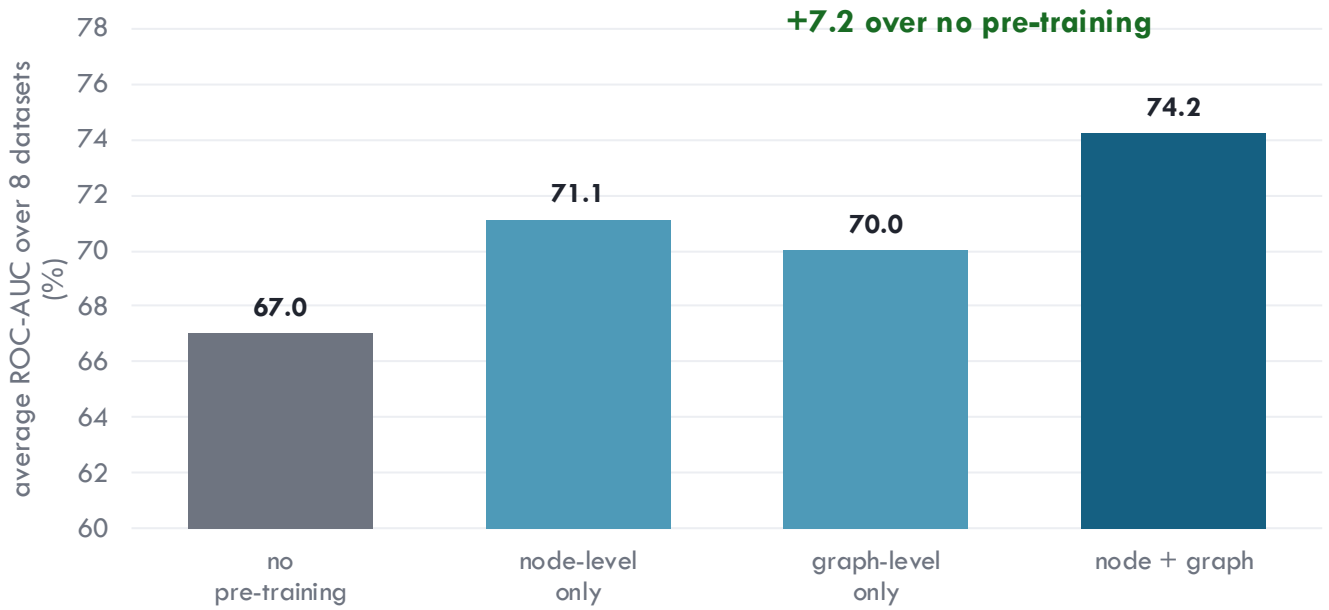
Ross et al., Nat. Mach. Intell. 2022, 4, 1256 · Ahmad et al., arXiv:2209.01712

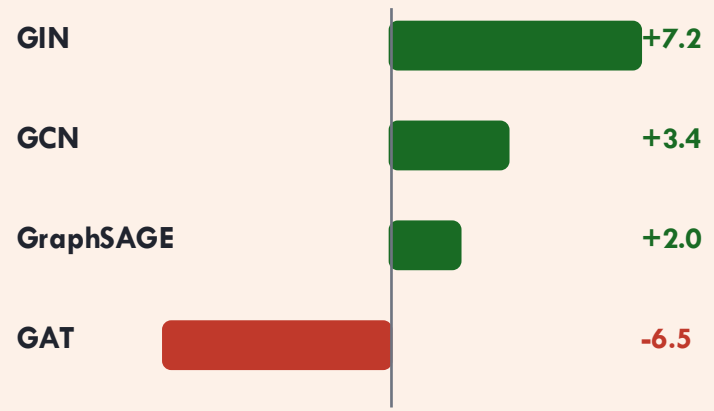
05

Transfer learning

- Pre-train once, fine-tune many times
- Frozen, partial, full
- When?

Pre-training GNNs: local and global, or negative transfer





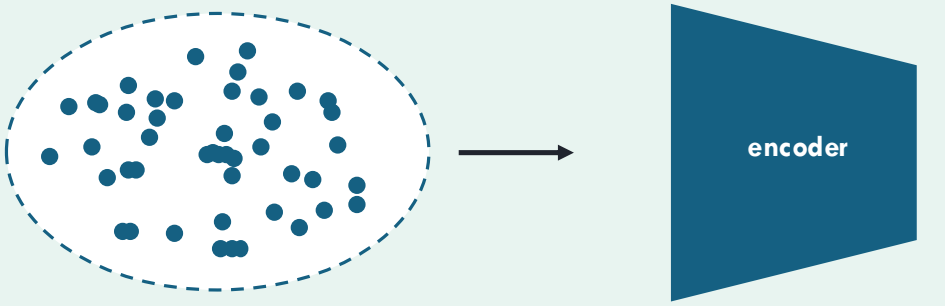
"We pre-trained it" guarantees nothing. **Pre-training a weaker architecture cost 6.5 points of ROC-AUC.**

Hu et al., ICLR 2020, arXiv:1905.12265 — chemistry average; the +9.4% in the abstract is the biology result

The recipe: pre-train once, fine-tune many times

STAGE 1 · PRE-TRAIN

once, by someone with a GPU cluster



10^6 – 10^9 molecules
no labels

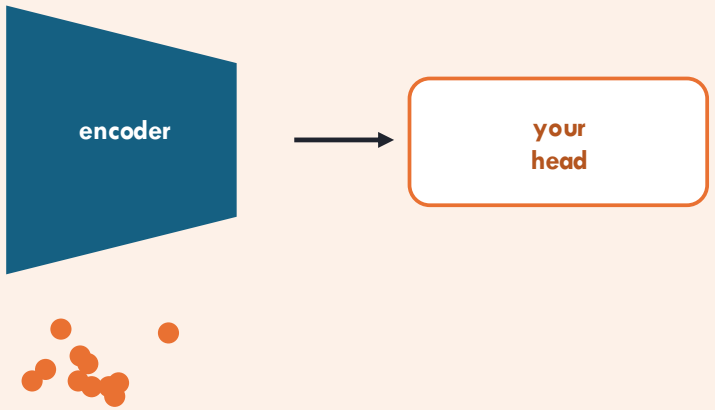
weights worth keeping

days of compute · done once · downloaded by everyone

reuse

STAGE 2 · FINE-TUNE

many times, by you, on a laptop



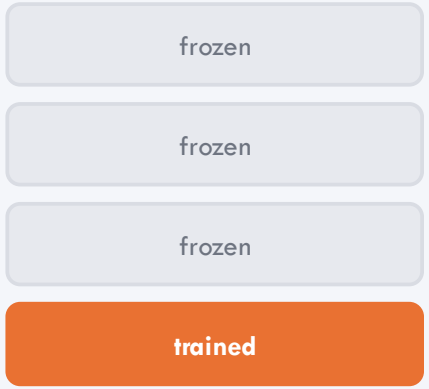
10^2 – 10^3 labelled
molecules

minutes of compute

The expensive half is **amortised across the whole community**. You are buying a billion molecules' worth of chemistry for a coffee break.

Three knobs — pick by your dataset size

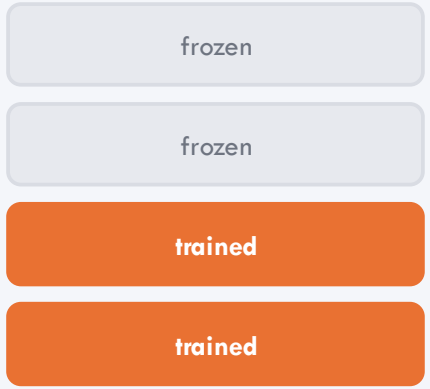
FROZEN (linear probe)



10 – 500 labels

fast · impossible to overfit the encoder

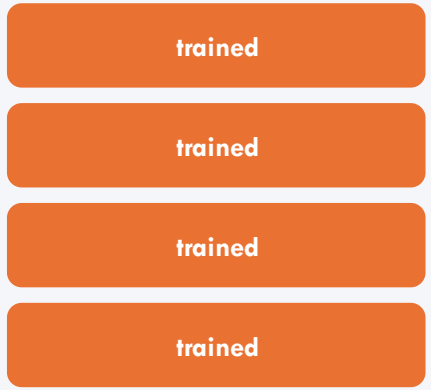
PARTIAL



500 – 5,000 labels

unfreeze the top layers only

FULL

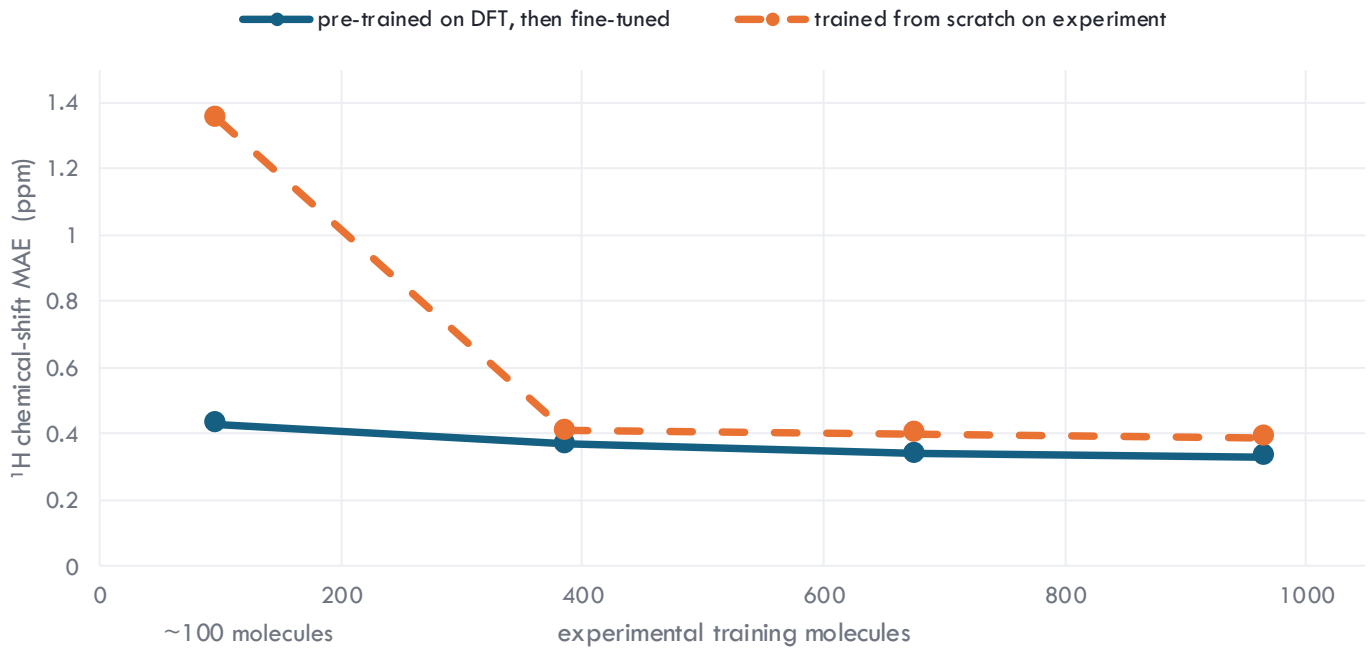


> 5,000 labels

best ceiling · needs a small learning rate

Always run the frozen probe first. Two minutes, no risk, and it tells you whether the representation contains anything useful for your task.

Where pre-training genuinely pays



the whole lecture in one plot

- Pre-training buys you data you do not have.
- Its value is largest when you have almost nothing.
- It decays to zero — and for ^{13}C it went negative — once you have ~1,000 labels.

Pre-training is a prior: worth 3× when you have ~100 labels, worth nothing by ~1,000, and occasionally worth less than nothing.

Han & Choi, J. Phys. Chem. Lett. 2021, 12, 3662 — ^{1}H MAE vs experimental training-set size (SI Table S2)

The same shape, in three different fields

REACTION YIELD

98

reactions are enough for a pre-trained reaction encoder to match DFT descriptors + RF

Buchwald–Hartwig HTE · Yield-BERT

INTERATOMIC POTENTIALS

3 %

of the data needed to specialise a foundation MLIP to liquid water, vs training from scratch

MatterSim · 17 M structures pre-trained

SUBLIMATION ENTHALPIES

tens

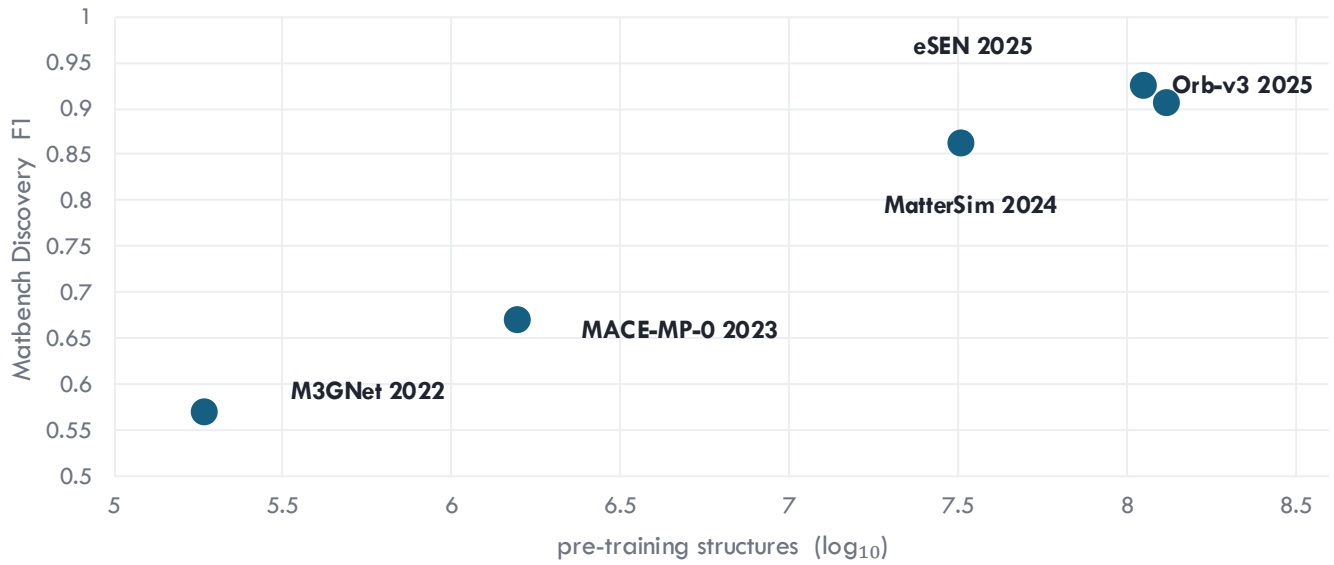
of training structures give sub-kJ/mol accuracy after fine-tuning MACE-MP-0

Kaur et al., Faraday Discuss. 2025

Different fields, different architectures, **the same shape** — the value of pre-training is concentrated where labels are most expensive.

Foundation models for atoms — the same play, at scale

~600× more training structures in three years



why chemists care

You will almost never train one of these.

You download it, then fine-tune on the 50–500 structures you can afford to compute for your own system.

Exactly the recipe from the last slide — applied to potentials instead of properties.

You will never train one of these — you will fine-tune one. Same recipe, applied to potentials instead of properties.

Matbench Discovery (matbench-discovery.materialsproject.org), values as of 11 Aug 2026 · Riebesell et al., Nat. Mach. Intell. 2025, 7, 836

06

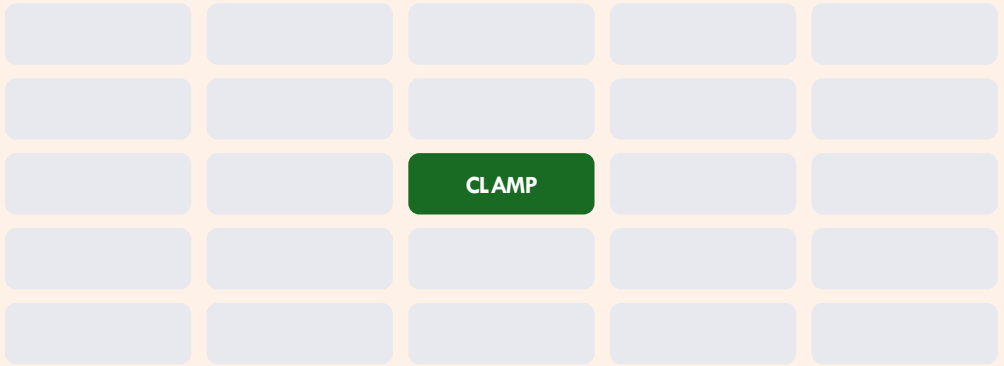
When it fails

- Negative transfer and the replication problem
- Activity cliffs
- The baseline you must always run

The uncomfortable benchmark

25 pre-trained models × 25 datasets

re-evaluated with hierarchical Bayesian testing
instead of leaderboard means



one model beats the ECFP baseline significantly —
and it is itself fingerprint-based

> *nearly all neural models show negligible or no improvement over the baseline ECFP molecular fingerprint.*

Praski, Adamczyk & Czech, arXiv:2508.06199 (2025)

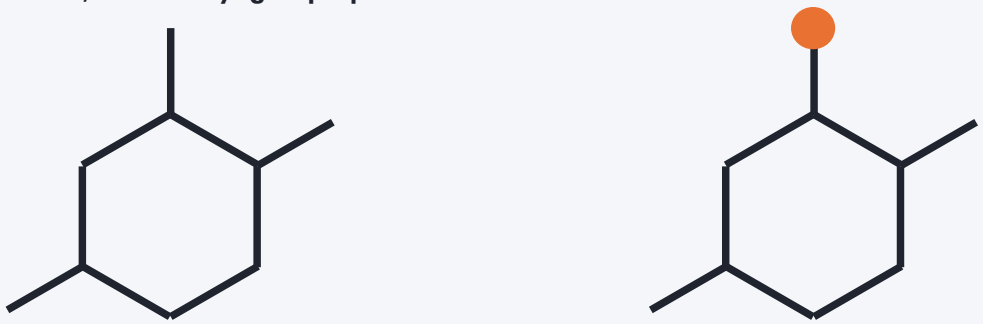
Not a reason to give up on learned representations.
A reason to always run the ECFP + random-forest baseline first.

Always run ECFP + random forest first. If your model cannot beat it, you have learned something valuable for the price of two minutes.

Praski, Adamczyk & Czech, arXiv:2508.06199 (2025) · Jiang et al., J. Cheminform. 2021, 13, 12 · Deng et al., Nat. Commun. 2023, 14, 6395

Activity cliffs: where smoothness is a lie

two molecules, one methyl group apart



IC_{50} = 8 nM

IC_{50} = 4,200 nM

Tanimoto similarity 0.95

500× potency difference

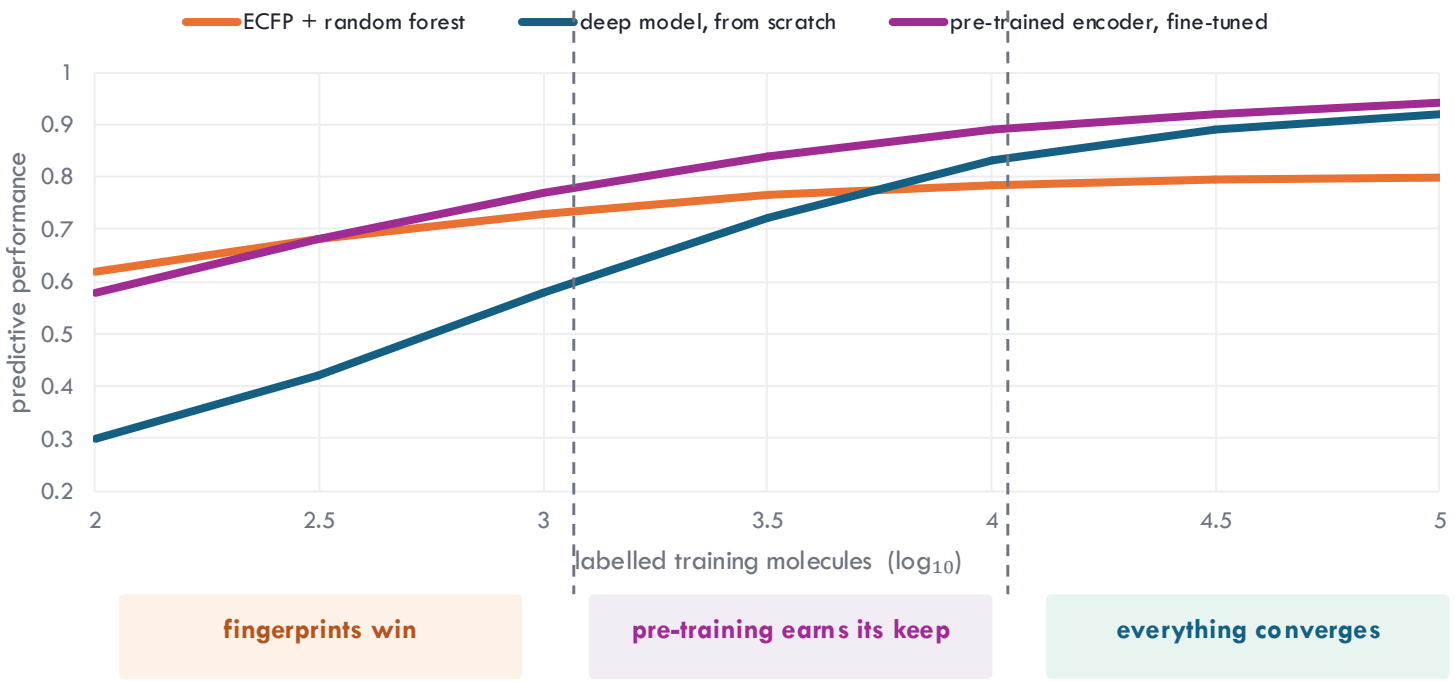
what breaks

- Smooth embeddings put A and B in the same place.
- Errors on cliff compounds reach 1.4 log units — a 25× error in potency.
- Across 30 targets, SVM on ECFP was the best method overall.
- Transformers pre-trained on 10 M SMILES did not beat LSTMs pre-trained on 36 k.

Smoothness is an assumption, and cliffs violate it. **Report cliff performance separately** — overall RMSE hides exactly the compounds that matter.

van Tilborg, Alenicheva & Grisoni, J. Chem. Inf. Model. 2022, 62, 5938 (MoleculeACE)

So when is a learned representation worth it?



schematic

The shapes are real;
the exact crossover is not.

Published estimates put it anywhere between 10^3 and 10^4 labelled molecules, and it moves with label noise, split type and how narrow your chemistry is.

Measure it on your own data — that is what the tutorial does.

The shapes are robust; **the crossover is yours to measure**, and it moves with label noise, split type and how narrow your chemistry is.

When to pretrain?

How many labelled examples do you have?

< 100

Do not train a deep model.
ECFP + RF, or a Gaussian process.
Consider frozen pre-trained embeddings as extra features.

100 – 5,000

This is where pre-training pays.
Frozen embeddings first, then partial fine-tuning. Always report the ECFP baseline.

> 10,000

Train from scratch on your own data.
Pre-training becomes a convenience, not a necessity.

In all three columns: run the baseline first, and split by scaffold, not at random.

Take-home messages

- **Labels are the bottleneck, not molecules.** Everything today is a way of spending unlabelled data to make up for scarce labels.
- **Unsupervised methods find real structure** — reaction classes, the periodic table — but a t-SNE plot tells you about your fingerprint, not about chemistry.
- **The VAE's decoder disappointed; its encoder changed the field.** Pretext task → keep the encoder → reuse it. That is the whole pattern.
- **Validity was the easy metric.** 100% valid is achieved by construction and says nothing about quality, novelty or synthesizability.
- **Pre-training is a prior.** Worth ~3× at a hundred labels, worth nothing at ten thousand, occasionally worth less than nothing.
- **Always run ECFP + random forest, and always split by scaffold.** In the largest independent replication, almost nothing beat that baseline.

Further reading

- **Chemical VAE** — Gómez-Bombarelli et al., ACS Cent. Sci. **2018**, 4, 268
- **SELFIES** — Krenn et al., Mach. Learn.: Sci. Technol. **2020**, 1, 045024 · and the critique: Skinnider, Nat. Mach. Intell. **2024**, 6, 437

- **Pre-training GNNs** — Hu et al., ICLR **2020**, arXiv:1905.12265
- **MoLFormer** — Ross et al., Nat. Mach. Intell. **2022**, 4, 1256 · **MolCLR** — Wang et al., Nat. Mach. Intell. **2022**, 4, 279

- **Reaction space** — Schwaller et al., Nat. Mach. Intell. **2021**, 3, 144
- **Transfer for NMR** — Han & Choi, J. Phys. Chem. Lett. **2021**, 12, 3662
- **Do they actually help?** — Deng et al., Nat. Commun. **2023**, 14, 6395 · Praski et al., arXiv:2508.06199 (**2025**) · van Tilborg et al., JCIM **2022**, 62, 5938
- **Evaluation guidelines** — Bender et al., Nat. Rev. Chem. **2022**, 6, 428

Learned Representations

Tutorial · 15:15 · RSC 275

Chemical space maps · frozen vs learned embeddings · the learning-curve crossover

jul@caltech.edu