What Can Neural Network Embeddings Do That Fingerprints Can't?

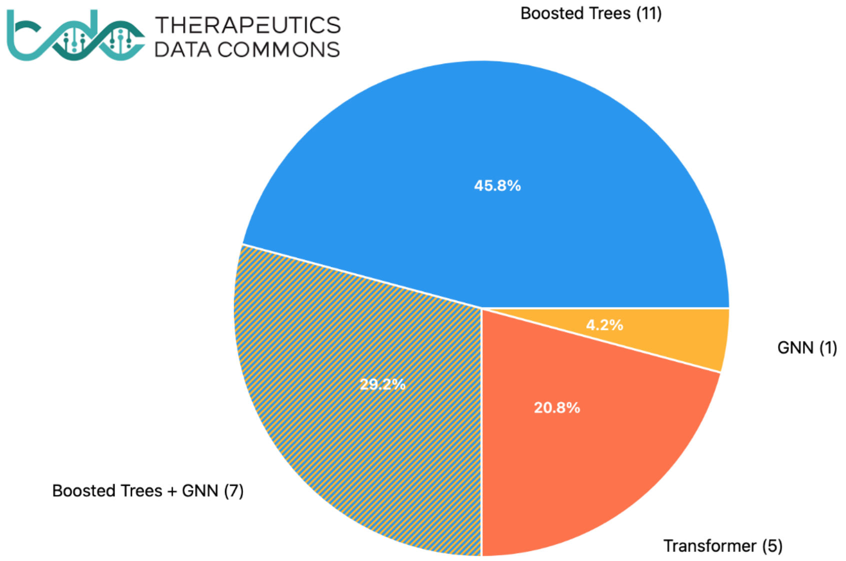
Nov 13 2024

Molecular fingerprints, like Extended-Connectivity Fingerprints (ECFP), are widely used because they are simple, interpretable, and efficient, encoding molecules into fixed-length bit vectors based on predefined structural features. In ADMET property prediction tasks fingerprints are still the state-of-the-art when combined with traditional machine learning methods like XGBoost or Random Forest. In contrast, neural network embeddings are dense, high-dimensional vectors learned directly from data using models like GraphConv, Chemprop, MolBERT, ChemBERTa, MolGPT, Graphformer and CHEESE. These models, trained on millions of drug-like molecules represented as SMILES, graphs, or 3D point clouds, capture continuous and context-dependent molecular features, enabling tasks such as property prediction, molecular similarity, and generative design. The rise of neural network-based representations has raised an important question: Do AI embeddings offer advantages over fingerprints?

The Performance Paradox

Here's the catch: in many standard predictive tasks, neural network embeddings do not necessarily outperform fingerprints. Benchmarks from the Therapeutic Data Commons (TDC), which include datasets for properties like bioavailability, lipophilicity, hERG toxicity, and half-life, reveal a surprising trend. As illustrated in the pie chart below, the majority of state-of-the-art (SOTA) results are achieved using "old-school" gradient-boosted trees (e.g., Random Forest or XGBoost) with molecular fingerprints. Only one in four datasets sees SOTA performance from more advanced architectures like Graph Neural Networks (GNNs) or Transformers.

This paradox raises a question: > If Neural Network or Transformer Models are so much larger and computationally powerful while not outperforming fingerprints in simple prediction tasks, what are their advantages?



A pie chart showing SOTA results on the TDC ADMET benchmark, indicating which model types achieved the best performance. All the boosted trees models used fingerprints, boosted trees + GNN means it is an ensemble model.

Neural Networks are Superior on Unstructured Data

Traditional algorithms like boosted trees perform exceptionally well on structured data, where relationships and patterns are well-defined and features are engineered. However, they falter on unstructured data, where modalities are diverse and relationships more complex. Neural networks excel in these cases, learning directly from unstructured or qualitative data such as images, audio, or natural language.

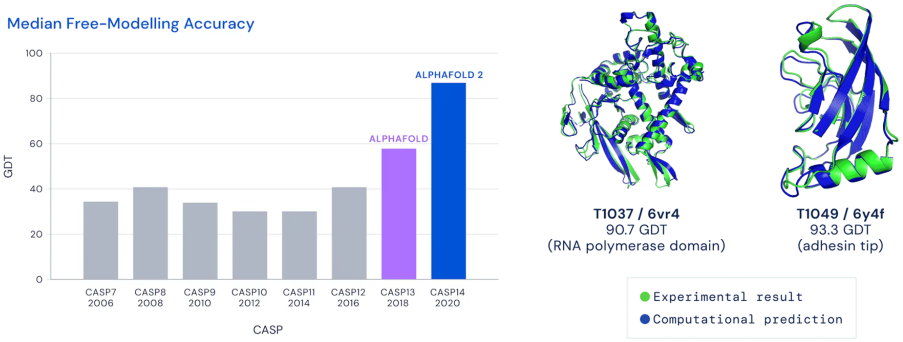
Yet in chemistry, some modalities are fairly structured. Representations like 2D molecular graphs, SMILES strings or fingerprints (which encode subgraph patterns) are discrete and inherently systematic, which is why traditional algorithms often perform as well—if not better—than neural networks, especially on small datasets.

> On structured modalities, particularly with small datasets, neural networks offer no inherent advantage over traditional boosted trees paired with binary fingerprints. This trend is clearly reflected in the TDC Benchmark results, where simple approaches often outperform fancy AI methods.

However, when it comes to continuous and unstructured data—especially in the context of large datasets—neural networks reveal their true potential. These are domains where crafting exact equations or simulations is infeasible due to the complexity of the data. Neural models can uncover patterns and relationships that are otherwise hidden, pushing the boundaries of discovery. **Examples include learning from 3D molecular shapes, electrostatics, or solving problems such as protein folding, conformer prediction, partial charge prediction and docking.**

Some of these applications, like protein folding, have already led to Nobel Prize breakthroughs, as seen with AlphaFold. Others, like docking and conformer generation, hold significant promise but are currently constrained by dataset size and the need for better benchmarks to avoid overfitting.

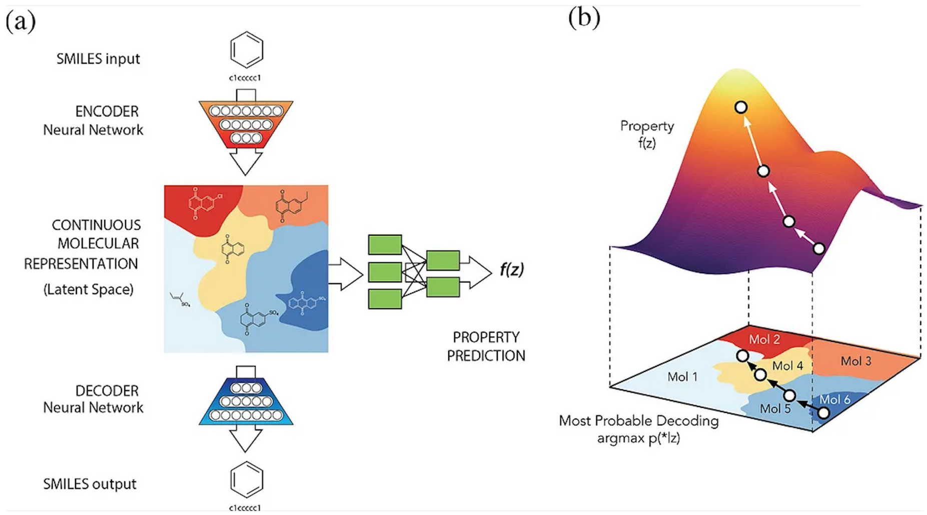
> Neural networks shine in unstructured or continuous modalities, such as 3D molecular shapes, electrostatics and large-scale datasets. Think of protein folding, conformer generation, or docking. In these domains, their ability to learn intricate, non-linear relationships directly from data—and even generate entirely new molecules—enables breakthroughs that traditional approaches cannot match.



A figure with Alphafold success in CASP visible on a bar graph (left) and example of protein tertiary structure predictions by it closely matching experimental results.

Smooth Latent Spaces: A Key Advantage

One of the most compelling strengths of neural embeddings lies in their ability to create **smooth latent spaces**, where similar inputs can be interpolated naturally. This feature underpins many modern generative models, such as VAEs, GANs, and diffusion models, and facilitates the **continuous optimization of molecular properties.** Moreover, it gives interpretability to their otherwise black-box nature. Discrete molecular representations, like SMILES strings or molecular graphs, can be transformed into continuous latent spaces, enabling seamless navigation and manipulation through mathematical operations. Selected points in this latent space can then be decoded back into molecular structures, making this approach invaluable for tasks like molecular design and property prediction.

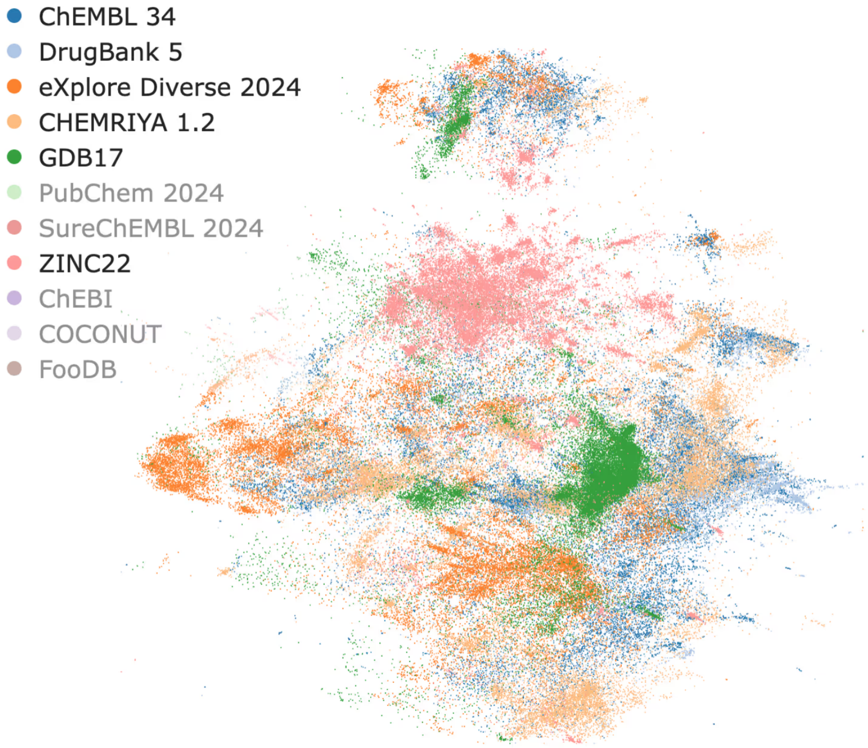


ChemVAE is a generative variational autoencoder from Goméz-Bombarelli et. al. which is able to continuously optimize properties of molecules in a latent space.

In these latent spaces, molecular similarity can be computed efficiently using simple metrics like **Euclidean Distance or Cosine Similarity** between vectors. These operations are highly optimized for GPU acceleration, allowing tools like CHEESE [citation] to perform clustering and searches on billion-scale chemical spaces with remarkable speed and efficiency. For example, CHEESE first identifies the most similar cluster by comparing centroid vectors and then refines the search within the cluster, achieving significant performance gains over traditional methods.

In contrast, traditional approaches face severe limitations at scale. For instance, while ROCS without GPU acceleration would take 50 years of CPU time to do shape-similarity screening on Enamine-REAL (5.5B database), CHEESE takes only 6 seconds for the same task by efficiently by leveraging the inherent advantages of its neural embedding latent space. Similarly, while clustering methods relying on Tanimoto similarity (e.g., Taylor-Butina) are computationally prohibitive on large datasets and would require a supercomputer to operate on such a large scale, CHEESE accomplishes billion-scale molecular clustering with commodity hardware.

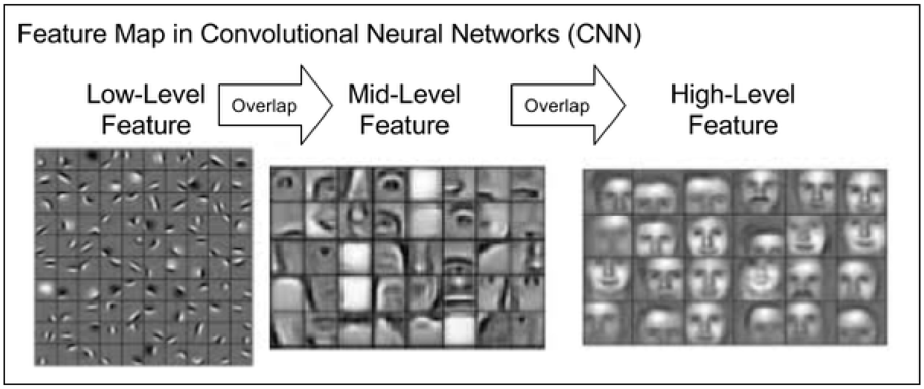
> A common criticism of neural networks is their perceived “black box” nature. However, neural latent spaces can be visualized, interpreted, and mapped back to the original molecular representations. Tools like CHEESE Explorer make this process intuitive, ensuring that neural embeddings are both practical and interpretable.



CHEESE Explorer: a Visualisation app for latent spaces of chemical databases

Embeddings as “Inceptions”

To understand what neural embeddings bring to the table, it’s helpful to consider why neural representations often surpass traditional approaches in other fields. A compelling example comes from computer vision, where convolutional neural networks (CNNs) have revolutionized image recognition. Initially, handcrafted features like shape matching or Fourier coefficients dominated, but they were eventually outpaced by neural approaches.



A depiction of feature maps in a convolutional neural network, progressing from low-level features (e.g., edges) to high-level abstractions (e.g., shapes or objects). CNN learns these features by itself on a face-recognition task.

This transformation occurred because capturing all possible image modalities manually is nearly impossible. Neural networks, like humans, can learn these modalities automatically from data. Consider the problem of recognizing a cat in an arbitrary image: defining an exhaustive set of equations or handcrafted features to account for every possible pose, background, or lighting condition is impractical. Neural networks overcome this challenge by learning directly from data, producing representations that generalize across scenarios.

Inception Distance

Neural embeddings can be likened to “Inceptions” in that they provide perceptual representation of data, allowing comparisons at a higher semantic level. Building on this analogy, the “distance” or distribution shift between embeddings can be measured using methods inspired by computer vision, such as the Fréchet Inception Distance (FID). FID, which derives its name from a Convolutional Neural Network (CNN) called Inception, is widely used to evaluate the quality of generated images. By analyzing embeddings rather than raw pixels, FID captures semantic differences, distinguishing visually similar but meaningfully different images (e.g., those with slight shifts, rotations, or added noise). This makes FID a robust measure for assessing the “realness” of images generated by models.

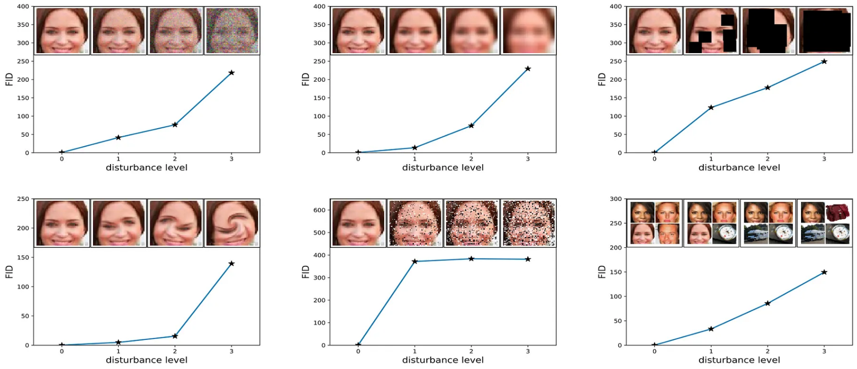
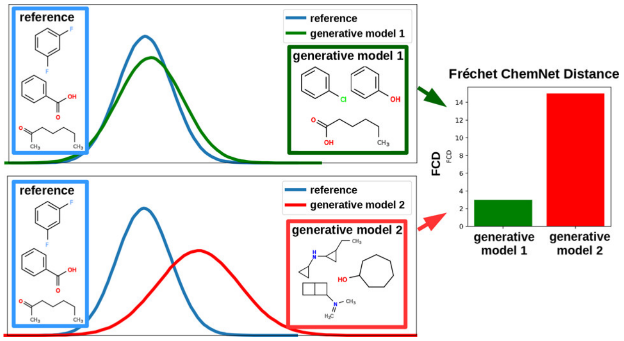


Figure 3: FID is evaluated for **upper left:** Gaussian noise, **upper middle:** Gaussian blur, **upper right:** implanted black rectangles, **lower left:** swirled images, **lower middle:** salt and pepper noise, and **lower right:** CelebA dataset contaminated by ImageNet images. The disturbance level rises from zero and increases to the highest level. The FID captures the disturbance level very well by monotonically increasing.

Figure showing various deformations of an image (noise, blur, mask, vortex, salt and pepper or collage) and corresponding increase in Fréchet Inception Distance.

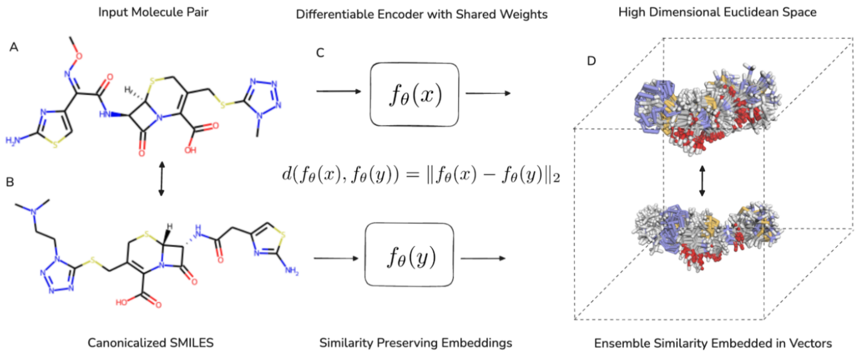
In the molecular domain, the Fréchet ChemNet Distance (FCD) adapts this concept to compare distributions of molecular data. FCD evaluates how closely the distribution of molecules generated by a model matches that of a reference dataset, such as ChEMBL, PubChem, or ZINC, by leveraging the activations of a pre-trained neural network. Instead of relying on predefined rules (e.g., Lipinski's Rule of Five), FCD uses the learned patterns in high-dimensional latent space to assess "drug-likeness" or other chemical properties. This approach excels at detecting subtle yet meaningful differences between distributions.



Graphical Abstract showing distributions of reference database molecules vs molecules by generative models. Valid, yet unplausible or undruglike molecules will have higher FCD from the reference database.

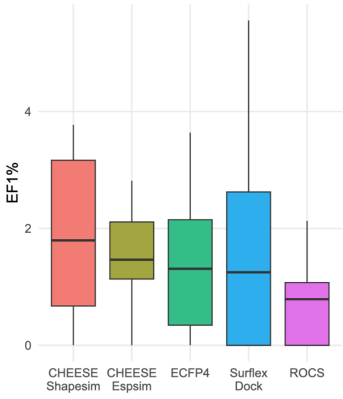
CHEESE Chemical Embeddings

CHEESE (Chemical Embeddings Search Engine) leverages neural embeddings to perform advanced similarity searches, prioritizing molecules that align closely with the query not only in 2D structure but also in **3D shape and electrostatic properties.** While any trained neural network learns an “inception distance” sui generis, this does not inherently guarantee chemical relevance in applications like virtual screening. CHEESE goes a step further by optimizing its loss function to emphasize physicochemically important metrics such as 3D shape and electrostatic similarity, ensuring that retrieved molecules are both intuitively relevant and chemically meaningful.


Input Molecule Pair
Differentiable Encoder with Shared Weights
High Dimensional Euclidean Space
A
B
C
D
$f_\theta(x)$
$d(f_\theta(x), f_\theta(y)) = \|f_\theta(x) - f_\theta(y)\|_2$
$f_\theta(y)$
Canonicalized SMILES
Similarity Preserving Embeddings
Ensemble Similarity Embedded in Vectors

CHEESE accepts pairs of molecules and an encoder with shared weights processes maps them into a continuous vector space. The encoder maintains differentiability throughout the encoding process and computes the Euclidean distance between the embeddings of the two molecules. The similarity preserving loss function ensures that the embeddings preserve molecular similarities in the induced euclidean vector space, penalizing deviations from isometry.

This targeted optimization delivers a significant improvement in virtual screening performance, as demonstrated on the LIT-PCBA benchmark, where CHEESE consistently outperformed traditional approaches in recovering active hits.



Target	**CHEESE Espsim**	**CHEESE Shapesim**	**ECFP4 Fingerprint**	**ROCS Combo**	**Surflex Dock**
ADRB2	**9.37**	3.12	0.00	0.00	0.00
ALDH1	1.20	1.18	1.58	1.08	1.25
ESR1_ago	2.02	**8.10**	0.00	0.00	0.00
ESR1_ant	1.21	0.49	2.67	1.07	1.60
FEN1	2.20	2.20	1.09	0.00	3.26
GBA	0.00	0.41	1.63	0.81	4.47
IDH1	1.59	1.59	1.59	0.79	0.79
KAT2A	1.39	3.47	0.69	0.69	4.17
MAPK1	1.46	0.86	0.95	1.39	1.99
MTORC1	0.00	3.77	0.00	0.00	1.52
OPRK1	0.00	0.00	**16.67**	0.00	0.00
PKM2	1.63	1.80	1.31	2.13	0.90
PPARG	1.07	3.21	**5.56**	**5.56**	**5.56**
TP53	2.82	0.00	0.00	0.88	0.00
VDR	**6.66**	2.42	3.64	0.00	0.00
Median	1.46	1.80	1.31	0.79	1.25

(a) Enrichment Ratio (Mean EF1%, $N = 15$)

(b) Table with Results (Mean EF1%). Values $> 5.$ in bold.

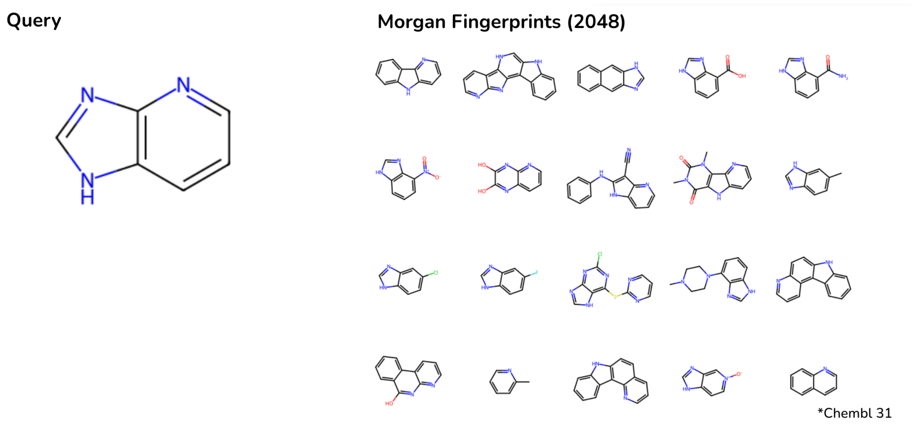
Figure 6: LIT-PCBA Unbiased (Single-Ligand) Evaluation

Evaluation of enrichment factor on a LIT-PCBA Benchmark (a benchmark based on PubChem Bioassays measuring a success rate of retrieving experimentally measured active molecules).

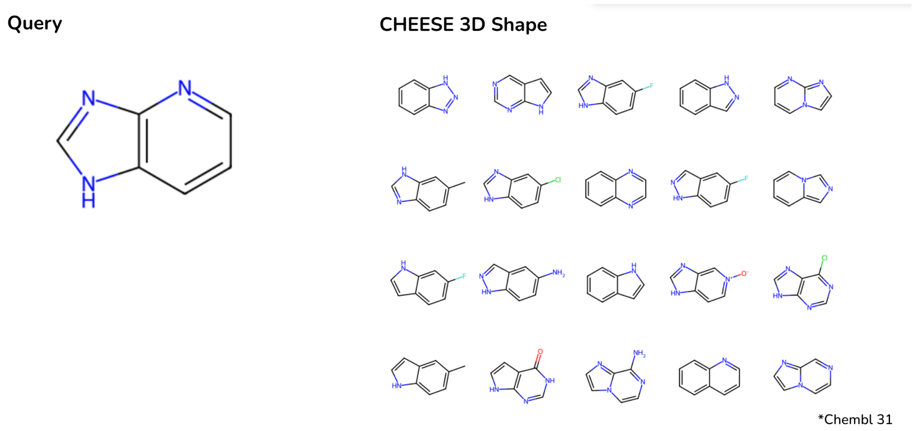
3D Shape Similarity

Traditional fingerprints, such as Morgan fingerprints, often fall short when molecular similarity relies heavily on 3D conformation rather than substructural patterns. Neural networks, particularly those designed for 3D molecular data, excel in capturing these spatial relationships, making them invaluable for tasks requiring precise shape and electrostatic comparisons.

For instance, in a ChEMBL database search for a small heterocyclic molecule, Morgan fingerprints retrieved molecules with significantly different scaffolds, ring counts, or molecular weights. In contrast, CHEESE embeddings accurately prioritized results based on 3D shape similarity, offering chemically and biologically relevant matches.



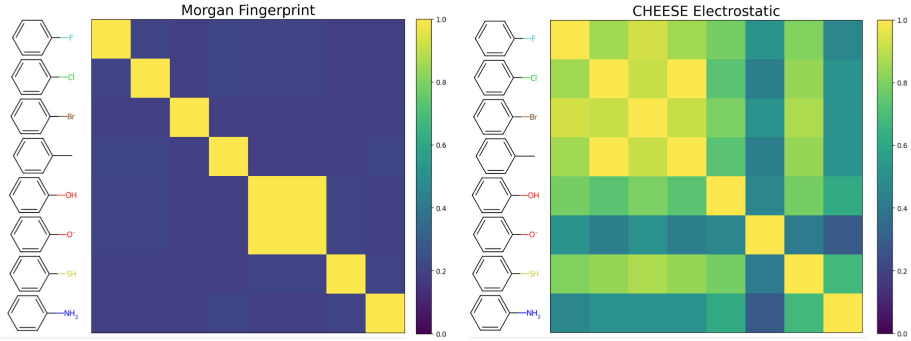
Search results on Chembl database when searching a small heterocyclic molecule with morgan fingerprints. Depicted are most tanimoto-similar molecules to fingerprint of the query. Results lack consistency in scaffold shape and molecular weight.



Search results on Chembl database when searching a small heterocyclic molecule with CHEESE Shapesim Embeddings. Depicted are most cosine-similar molecules to the embedding of the query. Results show closer matches based on 3D shape and molecular relevance.

Electrostatic Similarity

Beyond 3D shape, CHEESE embeddings shine in their ability to compare molecules based on **electrostatic properties**, a key determinant in molecular interactions. Traditional fingerprints reduce molecular information to fixed-length binary vectors, which are **inherently limited in representing continuous features like electrostatics.** In contrast, CHEESE uses continuous embedding vectors, enabling more nuanced and contextually relevant comparisons.

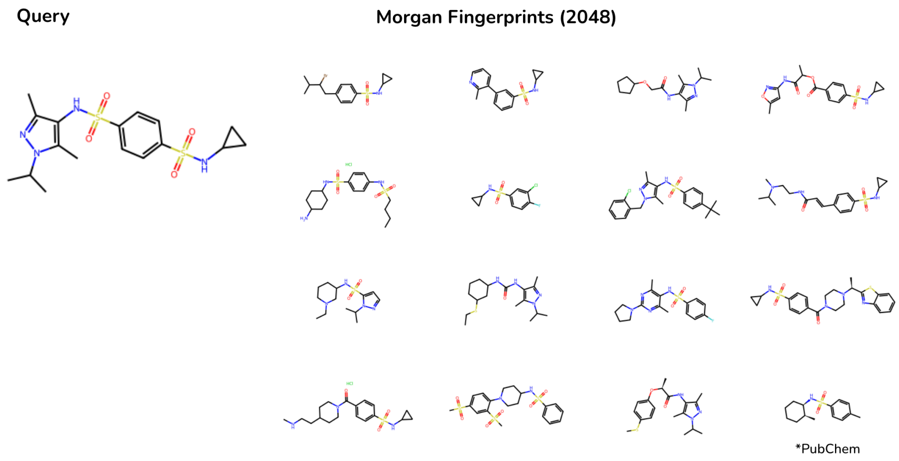


Comparison of similarity matrices: Tanimoto similarity for fingerprints (left) versus cosine similarity for CHEESE embeddings (right).

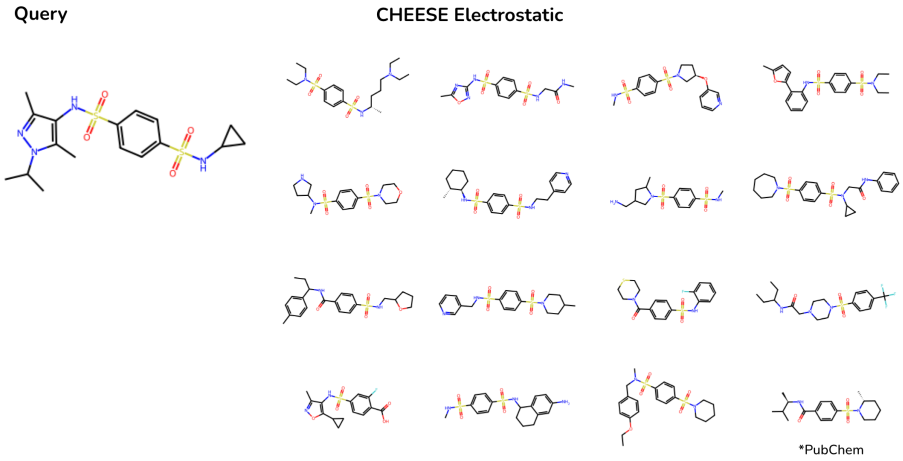
Consider the task of identifying molecules with similar electrostatic distributions in a large dataset, such as PubChem. Fingerprint-based methods often retrieve results that vary widely in electrostatic profiles, as they cannot directly encode or prioritize these properties. CHEESE, however, preserves electrostatic information in its embeddings. This enables it to consistently prioritize molecules with similar **charge distributions and interaction potentials**, even when structural similarities are weak.

> Fingerprints may retrieve molecules with electrostatically similar fragments or motifs, however they cannot capture the overal “big picture”. Tanimoto similarity is not always intuitive: parts of the found molecules may contain the same chemical groups, but as a whole it may be chemically entirely different molecule.

For instance, when searching for molecules with a sulfonamide functional group, Morgan fingerprints produced inconsistent results: some retrieved molecules had one sulfonamide group, others had three, and some lacked it entirely. CHEESE embeddings, on the other hand, correctly identified the electrostatically significant role of the sulfonamide group, prioritizing molecules with similar charge distributions and interaction profiles.



Search results on Chembl database when searching a molecule with sulfonamides with morgan fingerprints. Depicted are most tanimoto-similar molecules to fingerprint of the query. Results vary in electrostatic similarity.



Search results on Chembl database when searching a molecule with sulfonamides with CHEESE Espsim Embeddings. Depicted are most cosine-similar molecules to the embedding of the query. Results show improved alignment in electrostatic profiles.

Conclusion: Choosing the Right Tool

Neural network embeddings and traditional fingerprints each bring unique strengths to the table, and their utility depends on the specific task at hand. Fingerprints excel in scenarios where simplicity, interpretability, and computational efficiency are paramount—particularly for structured data and small-scale predictive tasks. On the other hand, neural embeddings shine when tackling unstructured or continuous molecular data, enabling breakthroughs in areas such as 3D molecular shape comparison, electrostatic similarity, and generative modeling.

The smooth latent spaces offered by embeddings open up new possibilities for molecular discovery, from generating novel compounds to finding nuanced relationships that fingerprints might miss. Tools like CHEESE exemplify how embeddings can redefine workflows in drug design and materials science, providing chemically meaningful insights that align with experimental observations.

Ultimately, rather than viewing these representations as competitors, it's more productive to see them as complementary tools. As datasets grow in size and complexity, and as AI models improve, the lines between structured and unstructured tasks will blur further, making the choice of representation less about technical limitations and more about aligning with the goals of the research. For chemists and data scientists alike, the future lies in leveraging the best of both worlds to accelerate discovery.

References

- Lžičař, M., & Gamouh, H. (2024). CHEESE: 3D Shape and Electrostatic Virtual Screening in a Vector Space. ChemRxiv. doi:10.26434/chemrxiv-2024-cswth.
- Rogers, D., & Hahn, M. (2010). Extended-connectivity fingerprints. Journal of Chemical Information and Modeling, 50(5), 742–754.
- Gilmer, J., Schoenholz, S. S., Riley, P. F., Vinyals, O., & Dahl, G. E. (2017). Neural message passing for quantum chemistry. arXiv preprint arXiv:1704.01212.
- Veličković, P., Cucurull, G., Casanova, A., Romero, A., Lio, P., & Bengio, Y. (2018). Graph attention networks. ICLR 2018.
- Yang, K., Swanson, K., Jin, W., Coley, C., Eiden, P., Gao, H., … & Barzilay, R. (2019). Analyzing learned molecular representations for property prediction. Journal of Chemical Information and Modeling, 59(8), 3370–3388.
- Grant, J. A., Gallardo, M., & Pickup, B. T. (1996). A fast method of molecular shape comparison: a simple application of a Gaussian description of molecular shape. Journal of Computational Chemistry, 17(14), 1653–1666.
- Gómez-Bombarelli, R., Wei, J. N., Duvenaud, D., Hernández-Lobato, J. M., Sánchez-Lengeling, B., Sheberla, D., … & Aspuru-Guzik, A. (2018). Automatic chemical design using a data-driven continuous representation of molecules. ACS Central Science, 4(2), 268–276.
- Preuer, K., Renz, P., Unterthiner, T., Hochreiter, S., & Klambauer, G. (2018). Fréchet ChemNet distance: a metric for generative models for molecules in drug discovery. Journal of Chemical Information and Modeling, 58(9), 1736–1741.
- Tran-Nguyen, V.-K., Jacquemard, C., & Rognan, D. (2020). LIT-PCBA:

An unbiased data set for machine learning and virtual screening. Journal of Chemical Information and Modeling, 60(9), 4263–4273.