Relation between a latent space and molecular biophysical properties

Author: Miguel Sánchez-Palomino, msanchpa135@alumnes.ub.edu
Facultat de Física, Universitat de Barcelona, Diagonal 645, 08028 Barcelona, Spain.

Advisor: Patrick Aloy Calaf, paloy@irbbarcelona.org
Advisor: Matteo Palassini, palassini@ub.edu

Abstract: Assessing the structure of latent spaces can be key when determining whether an AI model can be used for a given task. Through the analysis of nine physicochemical properties and two related magnitudes across six different latent spaces, this work attempts to provide a statistically robust methodology to determine whether those spaces can be used for tasks relying heavily in physicochemical properties. With it, one is able to rank the spaces based on their feasibility, emphasizing which model is more likely to outperform others in downstream tasks.
Keywords: Computational physics, linear algebra, numerical methods, statistical data analysis
SDGs: 9.5, 9.4, 12.2, 4.4

I. INTRODUCTION

Molecular embedding is a technique used to encode a molecule into a vector that preserves relevant chemical information. In this way, molecules are mapped to points in a well-defined mathematical space, known as latent space, where similar molecules are placed closer according to learned patterns. This representation allows easier analysis and manipulation using computational models, and enables downstream tasks such as predicting molecular properties, modelling chemical or biological behaviour or designing new compounds.

Often such embeddings are generated using artificial intelligence models, which tend to produce latent representations that are difficult to interpret. Even when the embeddings perform well in predictive tasks, it is often unclear what each of the individual dimensions of the vector represents in physical or chemical terms. This motivates the need for explainability, which is the ability to understand how the internal features of a model relate to meaningful real-world properties.

In this work, the explainability of molecular embeddings is studied by analysing how nine physicochemical properties are reflected across different latent spaces. Such properties are: molecular weight (MW, measured in unified atomic mass units u, also known as Daltons Da), partition coefficient (LOGP, a measure of the n-octanol/water partition coefficient, it indicates if a molecule is more likely to be absorbed in lipids or in water, with higher values indicating lipophilicity), molecular refractivity (MR, measure of the total polarizability of the molecule, calculated for a mole of substance), number of H-bond donors (HBD, strongly electronegative atoms such as N, O, F with at least one H atom directly attached to them), number of H-bond acceptors (HBA, electronegative atoms with a lone pair of electrons), polar surface area (PSA, measure of the total surface sum of all polar atoms including attached H, it measures the ability of a molecule to permeate cells), number of rotatable bonds (ROTB, rotatable bonds are non-ring bonds between non-terminal H atoms that can freely rotate around their axis, this means that the energy barrier of rotation can be overcome at room temperature), number of aromatic rings (AROM, an aromatic ring is a ring with a delocalized system of conjugated electrons, making it more stable) and number of structural alerts (ALERTS, a structural alert is a chemical substructure used as a predictor of adverse biological activity, such as toxicity).

Additionally, the study also covers two additional indicators of drug viability. The quantitative estimate of drug-likeness (QED [1], also known as chemical beauty, is a numerical score that predicts the performance of a molecule as a drug from eight physicochemical properties, higher is better). The number of Lipinski rule of 5 violations (RO5V, a set of four binary rules that determine the feasibility of a molecule to be orally consumed).

The analysis aims to provide a systematic way to determine whether a model structures its latent in a meaningful way and evaluating their suitability for tasks that rely on physicochemical understanding.

II. DATASETS

Data from the Chemical Checker (CC) [2] dataset has been used, it contains bioactivity information of over 1.2M chemical compounds among 25 spaces structured into five levels of increasing complexity: Chemistry, Targets, Networks, Cells and Clinics. Since the interest of this project revolves around small molecules, only molecules with a mass inferior to 700 Da have been considered. In addition, all molecular entities with more than one connected component (for example salts or ions) have been filtered out. After this distillation, a total of 1.13M molecules remained. Finally, all molecules were standardized using python package *rdkit* (v.2024.03.2). To conduct the studies ten randomly generated subsamples of 5k molecules have been considered.

III. MODELS

In this study, latent spaces from six different models have been used. Herein, a brief description of each model can be found. Unless stated otherwise the models use SMILES [3] representation of molecules as input.

Signaturizer. The signaturizers [4] are bioactivity descriptors for each of the spaces of the CC, providing a real valued 128-dimensional vector as the embedding of the molecule. Provided that the study focuses on physicochemical properties, the signature of the space A5 (Chemistry: Physicochemical properties) is considered. The encoding done by a signaturizer is done via a deep neural network, which is aware of the specific information of interest of the space. Additionally the signature of the space A1 (Chemistry: 2D fingerprints) has also been used. The resulting latent spaces are denoted by S and SA1 respectively.

Signaturizer3D. Signaturizer3D [5] is an enhancement of the original Signaturizer, where the model is aware of the stereochemistry of the molecule. Analogous to the Signaturizers those models provide a 128-dimensional real vector for each of the CC spaces. Those are trained on top of the Unimol [6] framework and later finetuned with the data from each of the spaces. Both A5 and A1 spaces have been considered, latent spaces are denoted by S3D and S3DA1 respectively.

SelfiesTED. SelfiesTED [7] is a transformer trained on SELFIES [8] for the vectorial representation of molecules. This transformer is based on BART (Bidirectional Auto-Regressive Transformer), allowing one to encode and decode molecules. In here, denoted by STED.

MolMetaLM. MolMetaLM [9] is a language model (LM) that uses a meta language (a language used to describe the language itself) to obtain a latent space for different molecules (subject) taking into account their physicochemical properties (predicate). For this case, the pretrained model from the authors has been used, denoted here by MMLM.

IV. METHODOLOGY

A. Ground truth spaces

Firstly, each of the subsamples is analysed to assess that they are representative of the CC universe. This assessment is done through the calculation of the mean, median, standard deviation, skewness and kurtosis for each individual subset and later comparing to the same values from the CC.

After this, an eleven-dimensional space was constructed such that each of its dimensions corresponds to a physicochemical property conveniently standardized to have a mean of 0 and a standard deviation of 1. By associating the euclidean metric to this space, a "ground truth" space (GT1) is constructed. This space sets one of the baselines to which the performance of the models is going to be compared.

An additional "ground truth" space (GT2) is constructed by considering the extended connectivity fingerprints (ECFP4) of the molecules and measuring their distance by means the Tanimoto distance [10]. This space is closer to biological truth since it considers structural elements of the molecule, allowing it to represent more than just its physicochemical properties.

B. Structure of the embeddings

For each latent space and each subgroup, a Principal Component Analysis (PCA) algorithm is performed to determine how many of many dimensions are relevant when attempting to explain the 95% of the variation within embeddings. This is equivalent to performing an eigendecomposition of the sample covariance matrix, and working with the square of the resulting eigenvalues. I.e. performing the eigendecomposition of the matrix $\frac{1}{N-1}X_c^T X_c$ where N denotes the number of samples and X_c the original matrix corrected by the per-feature mean.

This allows one to find the directions of maximum variance (as eigenvalues) and the value of such variance (as eigenvectors).

In addition, the Spearman correlation coefficient is computed between each dimension and physicochemical property to determine whether some dimensions relate directly with the values of the property. Also, the confidence interval at 95% is computed by taking 250 random samples with replacement for each of the groups, calculating the coefficient and considering the 2.5th to 97.5th percentile interval. The same test is repeated considering the projection over the principal vectors of the embeddings.

C. Recapitulation of the spaces' global structure

Firstly, the projection of the latent spaces into two dimensions is performed by means of three commonly used methods, tSNE and UMAP algorithms as well as the projection over the two vectors of maximum variance. For the visualization of the spaces, all datapoints have been coloured according to the value of the physicochemical property being studied.

tSNE algorithm (t-Distributed Stochastic Neighbour Embedding) [11] consists on the projection of a space where the differences in the data are emphasized. I.e. closer points (in the higher dimensional space) appear closer (in the projected space), while distant points appear distant. UMAP (Uniform Manifold Approximation and Projection) [12] is a dimensional reduction technique that approximates the manifold that the datapoints form, thus being able to preserve clusters in its projection. The mathematical background of these is not covered here as it exceeds the scope of this work.

To determine the existence of correlation between a latent space and GT1, GT2, the Pearson (r_P) and Spearman (r_S) correlation coefficients between pairwise distance matrices is computed. The calculation of such coefficients is done by flattening all the off-diagonal superior elements of the pairwise matrices and computing the coefficients of the resulting vectors. r_P is calculated through

$$r_P = \frac{COV(X,Y)}{\sigma_X \sigma_Y} = \frac{\sum(x-\bar{x})(y-\bar{y})}{[\sum(x-\bar{x})^2 \sum(y-\bar{y})^2]^{\frac{1}{2}}} \quad (1)$$

where $COV(X,Y)$ denotes the covariance between the

spaces X and Y, σ_X (respectively σ_Y) denotes the standard deviation of X (resp. Y) and $\bar{X}$ (resp. $\bar{Y}$) denotes the mean of X (resp. Y).

r_S is computed in an analogous fashion but substituting each variable value by its ranked position, this way considering monotonic relations, extending further from the linear case.

To further quantify the preservation of a local neighbourhood in the latent spaces, overlap (O), local continuity meta criterion ($LCMC$) and trustworthiness (T) values are calculated for k-nearest neighbours (kNN), with $k \in \{3, 5, 10, 25, 50\}$.

Overlap is the average ratio of nearest neighbours (NN) that are the same in the considered latent space and in GT spaces. That is,

$$O = \frac{1}{kN} \sum_{i=1}^{N} |\mathrm{NN}_k(x_i) \cap \mathrm{NN}_k(g_i)| \tag{2}$$

Where k is the number of NN considered, N is the amount of points in the space and x_i (resp. g_i) denote the ith point of the X space (GT space). For all latent spaces, the euclidean distance has been used.

Local continuity meta criterion is a correction to overlap to remove the effects of randomness within the measure, that is

$$LCMC = O - \frac{k}{N-1} \tag{3}$$

Trustworthiness is a measure of how well the local structure of the space is preserved with relation to the GT space. Through T, any sample that is closer to a point in latent space X is penalised according to its real rank in the GT space. The resulting value is bounded between 0 and 1 and is defined as

$$T = 1 - \frac{2}{NK(2N-3k-1)} \sum_{i=1}^{N} \sum_{j \in \mathrm{NN}_k(x_i)} \max\left(0, (r(i,j) - k)\right) \tag{4}$$

Finally, in order to visualize the local preservation of neighbourhoods, a Receiver Operating Characteristic curve (ROC curve) can be generated. On it, NN in the GT spaces are marked as positives, while all other points are set to negatives. By increasing a distance threshold, and determining whether the points within it are positive or negative, one can compute the false positive rate (FPR) and the true positive rate (TPR). The ROC curve emerges when plotting TPR against FPR. The closer that the area under the curve (AUC) is to unity (with unity being attained for a perfect classifier), the better that the overall structure is preserved, while a classifier with AUC = 0.5 is random. The steepness of the curve close to the origin, indicates how well the closest of all neighbours are recovered. In this study, ROCs have been generated for all values of k and for both GT spaces.

D. Latent interpretability with classical methods

A final measure of the interpretability of the latent spaces has been done by attempting to predict the values of one subset of molecules by using the other nine to train simpler models. The models considered are the following

- Linear regression, denoted by LR. In particular a multiple multivariate linear regression, where coefficients W and w are fitted such that when applied to the embedding x, the result is a vector of physicochemical properties $\hat{y} = Wx + w$. The fitting of matrix W and vector w is done by minimizing RMSE.
- k nearest neighbours regression, denoted by $kNNR$. In this regression, the predicted output $\hat{y}$ is the average of the k nearest neighbours present on the training set.
- Kernel ridge regression, denoted by kRR. In this regression, the data is fitted in a higher dimensional space, allowing the fitting to be non-linear in the original space. In order to set the parameters of the regression (regularization strength and gamma parameter of the RBF kernel), a grid search is performed. Only the best performing model has been considered.
- Random forest, denoted by RF. It builds predictions by taking into account simpler decision trees, built with portions of the training data, the final prediction considers the average of those outputs.
- Single hidden-layer neural network, denoted by MLP. A single fully-connected feed-forward neural network (multilayer perceptron) with a single hidden-layer containing as many neurons as half the amount of dimensions of the embedding.

For each of those models the values of their coefficient of determination (R^2), their maximum absolute error (MAE) and their root mean square error ($RMSE$) are provided. Taking y the real values and $\hat{y}$ the predicted ones, the expressions for these metrics are

$$R^2 = 1 - \frac{\sum_{1=1}^{N} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{N} (y_i - \bar{y})^2} \tag{5}$$

$$MAE = \frac{1}{N} \sum_{i=1}^{N} |y_i - \hat{y}_i| \tag{6}$$

$$RMSE = \left[\frac{\sum_{i=1}^{N} (y_i - \hat{y}_i)^2}{N} \right]^{\frac{1}{2}} \tag{7}$$

Additionally, specific seeds were used to initialize all random processes so results are reproducible.

V. RESULTS

A. Ground truth spaces

After computing the above mentioned analyses it was observed that all statistics but kurtosis were preserved through all groups with a difference smaller than 1%. For kurtosis there was a larger variability, which is consistent given that this statistic measures the thickness of the distribution tail, which is highly sensitive to the choice of molecules.

B. Structure of the embeddings

As shown in table I, S3D, S, S3DA1, SA1 show that their embedding size is close to optimal, provided that

TABLE I. Average number of required eigenvectors to explain 95% variance for different models over 10 groups

Model	S3D	S	S3DA1
# vecs/ total	94.4/128	84/128	96/128
Model	SA1	STED	MMLM
# vecs/ total	87/128	178.8/1024	3/768

most of the eigenvectors are required to explain the 95% of the variance. For the case of STED, less than 20% of vectors are required, indicating, a priori, that the embedding size is much larger than necessary. For MMLM, one can infer that the model is incapable of discerning between the molecules in the dataset, as only three eigenvectors are required.

When computing the Spearman correlation coefficients between each dimension and physicochemical property, one can observe that in general, some dimensions present a significantly non-zero correlation coefficient, while most are close to zero. This indicates that the embeddings likely have some preferred dimensions when inferring the physicochemical properties (fig. 1a for S3D, other in the appendix). When taking principal components (PC) as dimensions, such entries cannot be as well determined, as the correlation values are lower indicating that dimensions are more intertwined, and virtually all are required to explain each property (figures for PCA and remaining models are present in the appendix).

In addition, it is also observed that some of those high-correlation dimensions are shared among different physicochemical properties, hence indicating that both properties are also strongly correlated; for example, the correlation between molecular weight and molecular refractivity can be clearly seen, which is physically backed up by the fact that larger molecules have larger electron clouds (in volume), giving place to a higher polarizability in the molecule, which is proportional to molecular refractivity.

C. Recapitulation of the spaces' global structure

The representation of the S3DA1 through UMAP algorithm can be found in fig. 1b (other models in the appendix), in general no clear structure can be observed, but for some of the physicochemical properties (such as aromaticity or alerts) there are mild gradients, this is observed for S3D, S, S3DA1, SA1. For STED. For MMLM, many points are projected into the same coordinates, which is consistent with embeddings not being discriminative enough. Similar results are obtained for tSNE and PCA projections, figures can be found in the appendix.

Local recapitulation of neighbourhood is extremely good for S3D and S models, with trustworthiness values over 0.9, followed by STED and S3DA1, SA1, with values of around 0.85 and 0.75 respectively and finally by MMLM with values of around 0.5. Performance decreases as the number of NN increases.

The value of the AUC for ROC curves (considering $GT1$ spaces) is over 0.9 for S3D and S indicating a defined local structure, and over 0.8 when considering $GT2$ space. This manifests that the models still exhibit a good comprehension of the molecule beyond its physicochemical properties. For A1 spaces, it can be observed that $GT2$ scores are superior to $GT1$ (over 0.9 and over 0.75 respectively), indicating that those models are not so well aligned with physicochemical properties but present a better explanation of the molecular 2D structure. STED AUCs present better recapitulation for $GT1$ than for $GT2$ (0.74 and 0.71 respectively), while MMLM presents results are close to random (0.568 and 0.6) and with flat slopes around the origin, showing a bad recapitulation of neighbours.

D. Latent interpretability with classical methods

The results obtained when attempting to model the spaces via some classical methods can be found in table II.

For signaturizers, the performance of the linear regression is significantly lower than the one achieved with non-linear methods, thus manifesting that the latent space structure is non-linear. For STED, we observe that even if non-linear methods are superior, linear regression approximates well the results, meaning that the space could be almost-lineal. For MMLM, there is no good method that approximates the space, this could once more be indicating that the space is unable to discern between different molecules.

VI. CONCLUSIONS

Throughout this work we have observed that the model S3D and S preserve the local neighbourhood of GT1 much better than the other models, hinting a potential superiority of performance in tasks involving physicochemical properties. However, when comparing the neighbourhood preservation of GT2, the models that perform best are S3DA1 and SA1, while other also have high recapitulation metrics. This shows how to some extent the remaining models are also useful for other tasks that require a comprehension of the molecule.

For the case of MMLM, a failure in the construction of embeddings was observed, as those were unable to discern between different molecules. This could possibly have been overcome by finetuning the model with a representative set of molecules from the dataset.

It is also worth noting the difference in dimensions for different models, as the better-performing signaturizer family ones are 128 dimensions long, while STED contains 1024 dimensions, thus indicating that the dimensions of signaturizer ones are more aligned with the purposes of this project. Superior performance of Signaturizers was expected since these are tailored to be aware of the magnitudes here measured.

As it is the case for SA1 and S3DA1, STED could be focusing on different features of the molecules beyond physicochemical properties, yielding a lower performance on those tasks, and not manifesting their complete potential in this study. That is also the case for MMLM, where the lack of good embeddings forbid the model to show its usefulness. More general tests than the ones

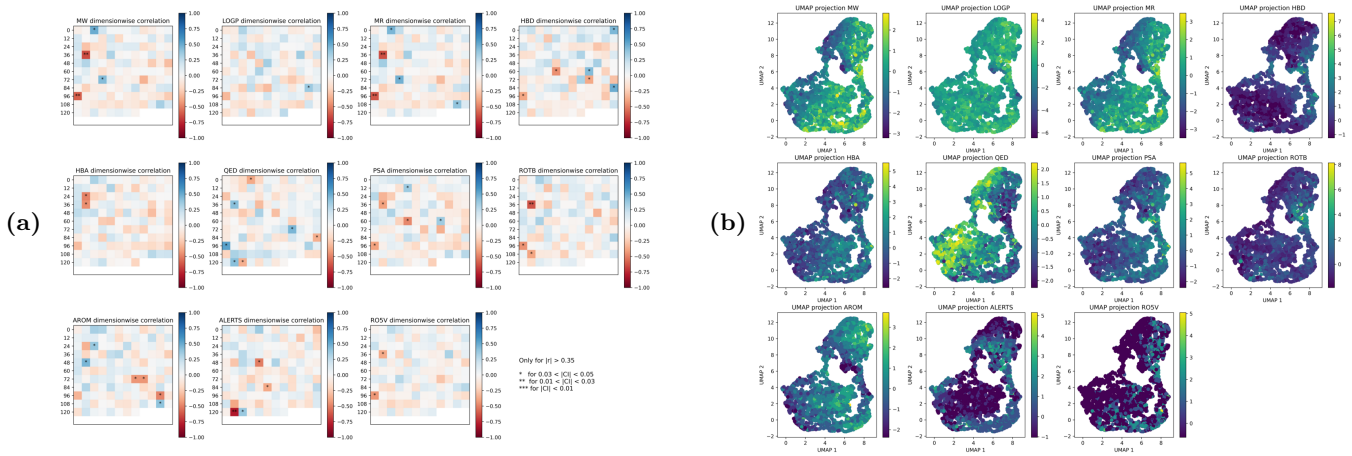

FIG. 1. **(a)** Spearman correlation for each dimension (ordered from left to right, top to bottom) of Signaturizer3D and each physicochemical property indicated in the title. Colour indicates the value of the correlation coefficient. Through asterisks it is indicated the width of the confidence interval of the coefficient for all correlations larger than 0.35. **(b)** UMAP projection of Signaturizer3D model for each physicochemical property. Colour indicates the value of the property standardized.

TABLE II. $R^2/MAE/RMSE$ scores for linear regression, k nearest neighbours regression (k=3), kernel ridge regression (best performing hyperparameters), random forest and single-hidden layer neural networks for each model. Values obtained for the prediction of one group when considering the other nine to train.

Model	LR	kNNR	KRR	RF	MLP
S3D	0.619/0.384/0.264	0.786/0.278/0.180	0.832/0.296/0.166	0.845/0.279/0.153	0.830/0.303/0.170
S	0.584/0.383/0.270	0.744/0.288/0.207	0.812/0.307/0.186	0.822/0.289/0.176	0.806/0.315/0.191
S3DA1	-0.089/0.538/0.488	0.439/0.447/0.408	0.599/0.479/0.396	0.624/0.461/0.371	0.570/0.498/0.424
SA1	-0.402/0.565/0.539	0.325/0.486/0.478	0.539/0.514/0.455	0.576/0.489/0.419	0.518/0.530/0.476
STED	0.806/0.268/0.146	0.638/0.349/0.263	0.877/0.240/0.122	0.685/0.416/0.312	0.905/0.209/0.094
MMLM	-96.79/0.783/0.966	-3.279/0.896/1.277	0.162/0.714/0.827	0.198/0.699/0.792	0.019/0.782/0.792

analysed in this paper could also help illustrate the value of these embeddings.

ACKNOWLEDGMENTS

I would like to thank my supervisors both at IRB Barcelona and at UB for their attent supervision and valuable feedback. I am also grateful to these institutions by providing me both the knowledge and the opportunities required for the completion of this work. Finally, I want to acknowledge my family and friends by providing support and making easier the confection of this project.

[1] Bickerton, G. et al. "Quantifying the chemical beauty of drugs" Nature Chem, **4**, 90-98 (2012)
[2] Duran-Frigola, M. et al. "Extending the small-molecule similarity principle to all levels of biology with the Chemical Checker" Nature Biotechnology **38**: 1087-1096 (2020)
[3] Weininger, D. "SMILES, a Chemical Language and Information System. 1. Introduction to methodology and encoding rules" Journal of Chemical Information and Computer Sciences. **28**: 31-36 (1988)
[4] Bertoni, M. et al. "Bioactivity descriptors for uncharacterized chemical compounds" Nature Communications **12**, Article number 3932 (2021)
[5] Comajuncosa-Creus, A. et al. "Stereochemically aware bioactivity descriptors for uncharacterized chemical compounds" Journal of Cheminformatics. **16**, Article number 70 (2024)
[6] Zhou, G et al. "Uni-Mol: A Universal 3D Molecular Representation Learning Framework" The 11th Intl. Conference on Learning Representations, (2023)
[7] IBM Research "Selfies-TED : A Robust Transformer Model For Molecular Representation Using SELFIES" International Conference on Learning Representations, 8268 (2025)
[8] Krenn, M. et al. "Self-Referencing Embedded Strings (SELFIES): A 100% robust molecular string representation" Machine Learning: Science and Technology. **1**, Article number 4 (2020)
[9] Wu, Y. et al. "MolMetaLM: a Physicochemical Knowledge-Guided Molecular Meta Language Model" ArXiV, 2411.15500 (2024)
[10] Tanimoto, T.T. "An elementary mathematical theory of classification and prediction" IBM IR, (1957)
[11] Van Der Maaten, L. and Hinton, G. "Visualizing data using t-SNE" Journal of Machine Learning Research. **9**, 2579-2605 (2008)
[12] McInnes, L. et al. "UMAP: Uniform Manifold Approximation and Projection" Journal of Open Source Software, **3(29)**, 861 (2018)

Relació entre un espai latent i propietats biofísiques moleculars

Author: Miguel Sánchez-Palomino, msanchpa135@alumnes.ub.edu
Facultat de Física, Universitat de Barcelona, Diagonal 645, 08028 Barcelona, Spain.

Advisor: Patrick Aloy Calaf, paloy@irbbarcelona.org
Advisor: Matteo Palassini, palassini@ub.edu

Resum: Determinar l'estructura dels espais latents pot ser clau per a determinar si un model d'IA pot ser emprat per una tasca donada. A partir de l'anàlisi de nou propietats fisicoquímiques i dues magnituds relacionades en sis espais latents diferents, aquest treball intenta proporcionar una metodologia estadística robusta per a determinar quins d'aquests espais poden ser usats per a tasques basades en propietats fisicoquímiques. Amb aquesta, som capaços d'ordenar els espais basats en la seva viabilitat, deixant clar quin model és més propens a superar els altres en tasques relacionades.
Paraules clau: Física computacional, anàlisi de dades estadístic, àlgebra lineal, mètodes numèrics
ODSs: 9.5, 9.4, 12.2, 4.4

OBJECTIUS DE DESENVOLUPAMENT SOSTENIBLE (ODSS O SDGS)

1. Fi de la es desigualtats		10. Reducció de les desigualtats	
2. Fam zero		11. Ciutats i comunitats sostenibles	
3. Salut i benestar		12. Consum i producció responsables	X
4. Educació de qualitat	X	13. Acció climàtica	
5. Igualtat de gènere		14. Vida submarina	
6. Aigua neta i sanejament		15. Vida terrestre	
7. Energia neta i sostenible		16. Pau, justícia i institucions sòlides	
8. Treball digne i creixement econòmic		17. Aliança pels objectius	
9. Indústria, innovació, infraestructures	X		

Podem relacionar el treball amb els objectius de desenvolupament sostenible 9.5 donat que es proposa una metodologia estadísticament robusta per evaluar la viabilitat d'espais latents per a tasques donades, contribuint a una millora de les eines emprades en recerca científica, d'aquesta manera el projecte reforça la capacitat computacional i tecnològica de les infraestructures de recerca. La tria d'un model més eficient provoca una menor despesa de recursos computacionals en aquestes, fent que el treball assoleixi la fita 9.4. En una línia similar, aquest estalvi té un impacte directe sobre els recursos naturals, alineant-se amb 12.2. Finalment el treball es troba alineat amb la fita 4.4 ja que contribueix al desenvolupament d'habilitats relacionades amb la innovació.

Appendix A: ADDITIONAL FIGURES

1. Signaturizer3D

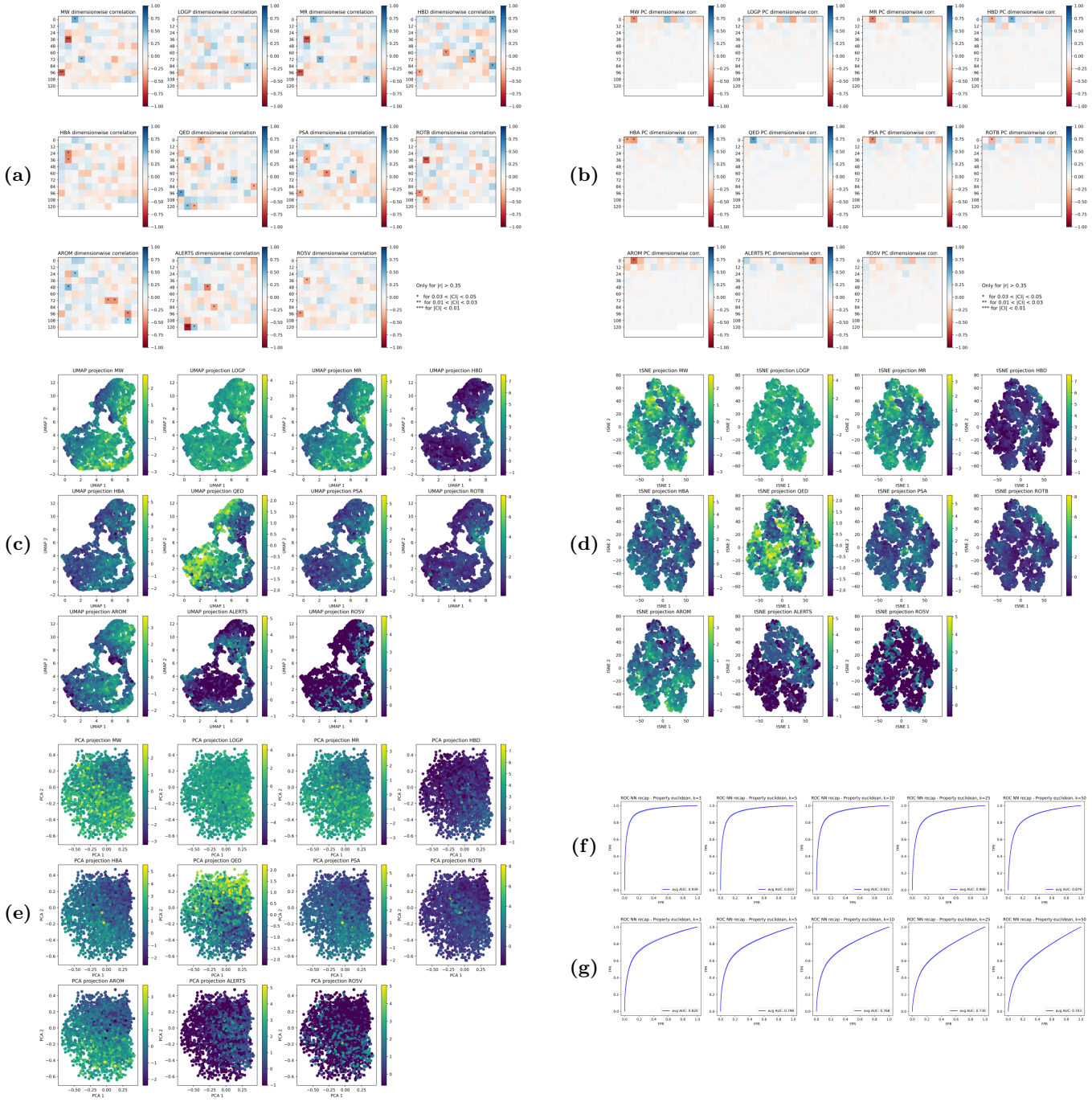

FIG. 2. **(a)** Spearman correlation for each dimension (ordered from left to right, top to bottom) of Signaturizer3D and each physicochemical property indicated in the title. Colour indicates the value of the correlation coefficient. Through asterisks it is indicated the width of the confidence interval of the coefficient for all correlations larger than 0.35. **(b)** Same as (a) but taking principal components of variance instead of dimensions. **(c)** UMAP projection of Signaturizer3D model for each physicochemical property. Colour indicates the value of the property standardized. **(d)** Same as (c) but using tSNE projection. **(e)** Same as (c), (d) but over the two principal components from the sample covariance matrix. **(f)** ROC curves for neighbourhood recapitulation with respect to $GT1$ for $k \in \{3, 5, 10, 25, 50\}$ neighbours. **(g)** Same as (f) but considering ground truth space $GT2$.

2. Signaturizer

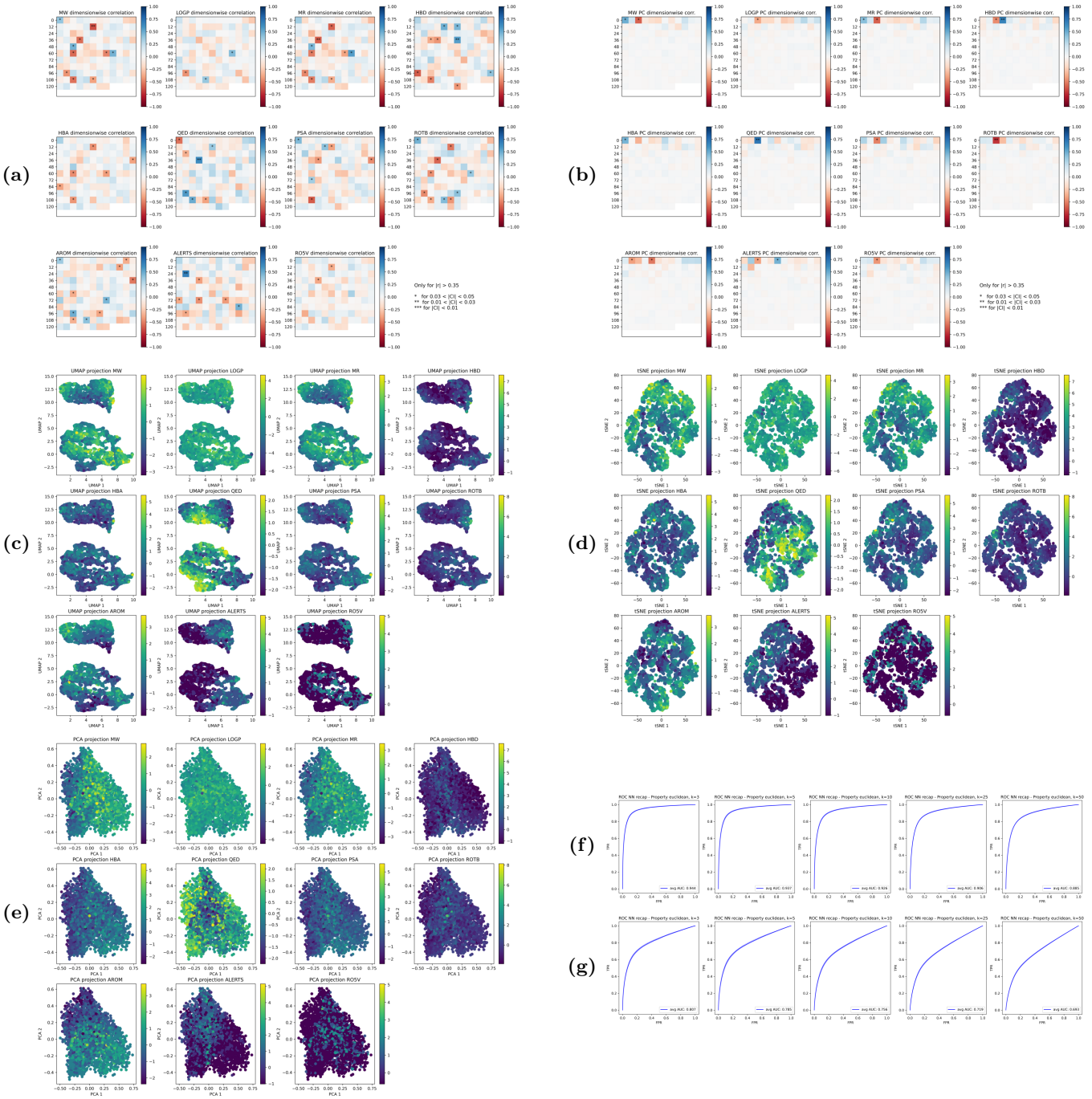

FIG. 3. **(a)** Spearman correlation for each dimension (ordered from left to right, top to bottom) of Signaturizer and each physicochemical property indicated in the title. Colour indicates the value of the correlation coefficient. Through asterisks it is indicated the width of the confidence interval of the coefficient for all correlations larger than 0.35. **(b)** Same as (a) but taking principal components of variance instead of dimensions. **(c)** UMAP projection of Signaturizer model for each physicochemical property. Colour indicates the value of the property standardized. **(d)** Same as (c) but using tSNE projection. **(e)** Same as (c), (d) but over the two principal components from the sample covariance matrix. **(f)** ROC curves for neighbourhood recapitulation with respect to $GT1$ for $k \in \{3, 5, 10, 25, 50\}$ neighbours. **(g)** Same as (f) but considering ground truth space $GT2$.

3. Signaturizer3D_A1

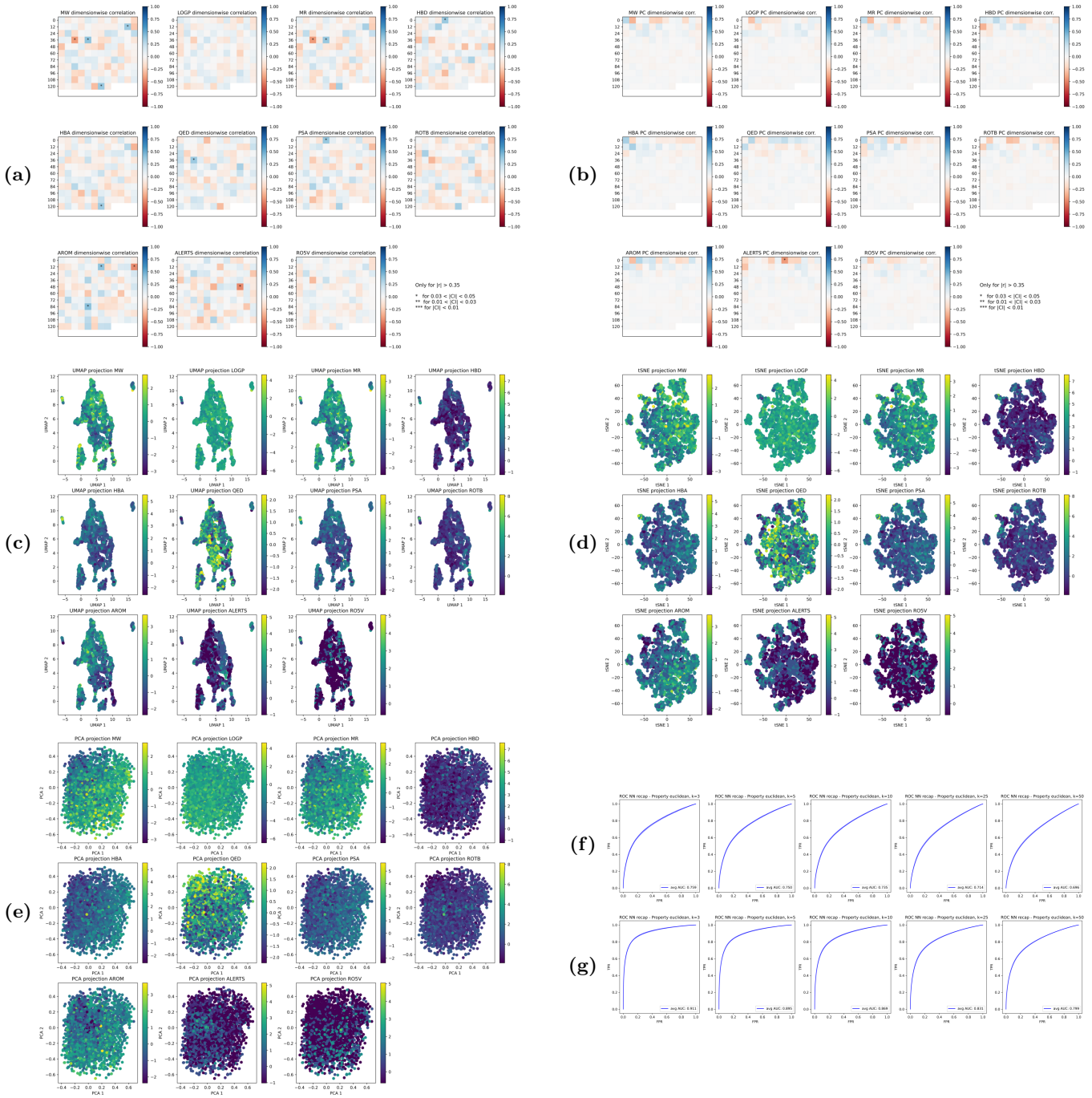

FIG. 4. **(a)** Spearman correlation for each dimension (ordered from left to right, top to bottom) of Signaturizer3d_A1 and each physicochemical property indicated in the title. Colour indicates the value of the correlation coefficient. Through asterisks it is indicated the width of the confidence interval of the coefficient for all correlations larger than 0.35. **(b)** Same as (a) but taking principal components of variance instead of dimensions. **(c)** UMAP projection of Signaturizer3d_A1 model for each physicochemical property. Colour indicates the value of the property standardized. **(d)** Same as (c) but using tSNE projection. **(e)** Same as (c), (d) but over the two principal components from the sample covariance matrix. **(f)** ROC curves for neighbourhood recapitulation with respect to $GT1$ for $k \in \{3, 5, 10, 25, 50\}$ neighbours. **(g)** Same as (f) but considering ground truth space $GT2$.

4. Signaturizer_A1

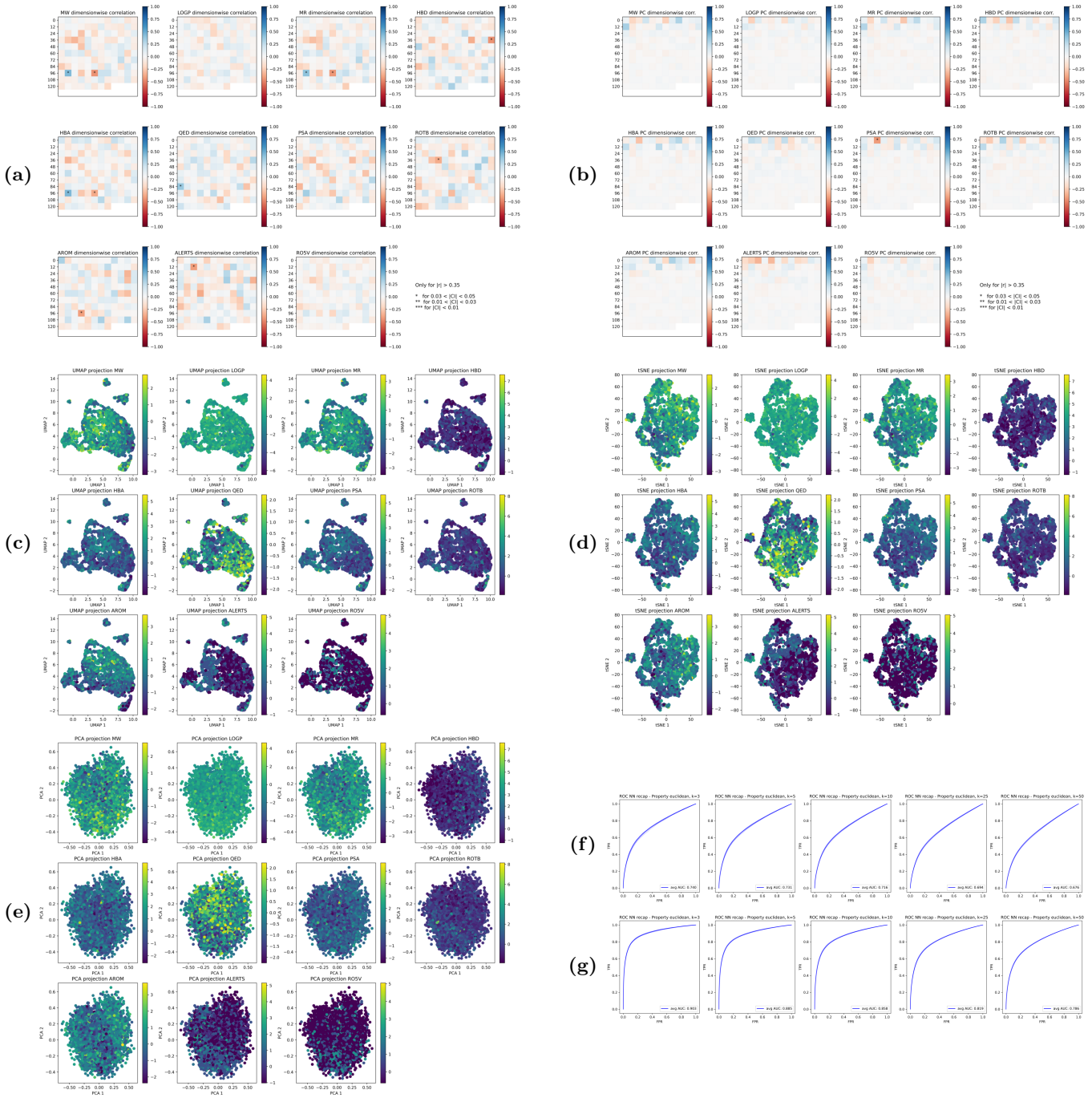

FIG. 5. **(a)** Spearman correlation for each dimension (ordered from left to right, top to bottom) of Signaturizer_A1 and each physicochemical property indicated in the title. Colour indicates the value of the correlation coefficient. Through asterisks it is indicated the width of the confidence interval of the coefficient for all correlations larger than 0.35. **(b)** Same as (a) but taking principal components of variance instead of dimensions. **(c)** UMAP projection of Signaturizer_A1 model for each physicochemical property. Colour indicates the value of the property standardized. **(d)** Same as (c) but using tSNE projection. **(e)** Same as (c), (d) but over the two principal components from the sample covariance matrix. **(f)** ROC curves for neighbourhood recapitulation with respect to $GT1$ for $k \in \{3, 5, 10, 25, 50\}$ neighbours. **(g)** Same as (f) but considering ground truth space $GT2$.

5. Selfies-TED

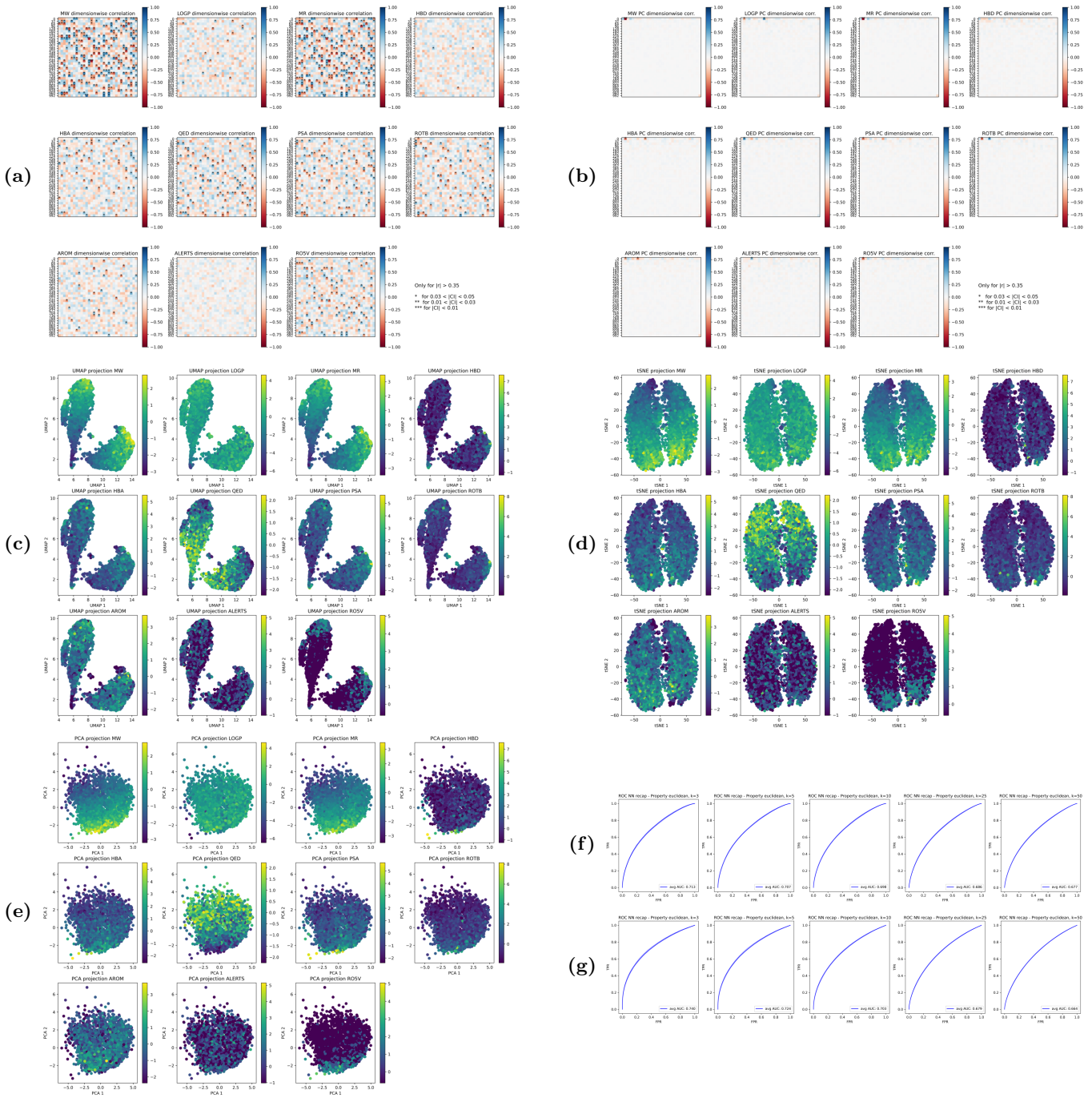

FIG. 6. **(a)** Spearman correlation for each dimension (ordered from left to right, top to bottom) of Selfies-TED and each physicochemical property indicated in the title. Colour indicates the value of the correlation coefficient. Through asterisks it is indicated the width of the confidence interval of the coefficient for all correlations larger than 0.35. **(b)** Same as (a) but taking principal components of variance instead of dimensions. **(c)** UMAP projection of Selfies-TED model for each physicochemical property. Colour indicates the value of the property standardized. **(d)** Same as (c) but using tSNE projection. **(e)** Same as (c), (d) but over the two principal components from the sample covariance matrix. **(f)** ROC curves for neighbourhood recapitulation with respect to $GT1$ for $k \in \{3, 5, 10, 25, 50\}$ neighbours. **(g)** Same as (f) but considering ground truth space $GT2$.

6. MolMetaLM

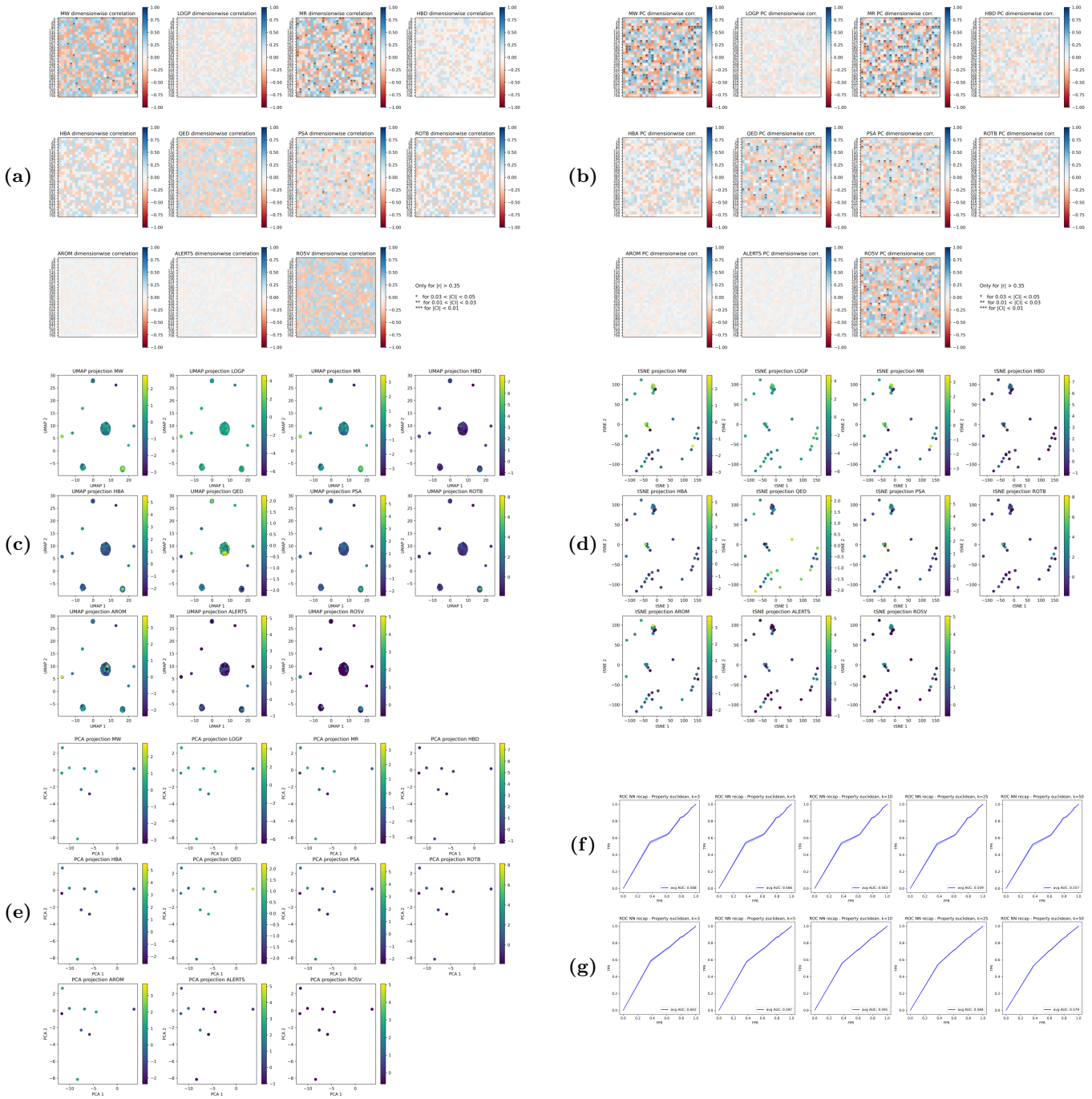

FIG. 7. **(a)** Spearman correlation for each dimension (ordered from left to right, top to bottom) of MolMetaLM and each physicochemical property indicated in the title. Colour indicates the value of the correlation coefficient. Through asterisks it is indicated the width of the confidence interval of the coefficient for all correlations larger than 0.35. **(b)** Same as (a) but taking principal components of variance instead of dimensions. **(c)** UMAP projection of MolMetaLM model for each physicochemical property. Colour indicates the value of the property standardized. **(d)** Same as (c) but using tSNE projection. **(e)** Same as (c), (d) but over the two principal components from the sample covariance matrix. **(f)** ROC curves for neighbourhood recapitulation with respect to $GT1$ for $k \in \{3, 5, 10, 25, 50\}$ neighbours. **(g)** Same as (f) but considering ground truth space $GT2$.