Heliyon 8 (2022) e10011

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Research article

Development of an innovative data-driven system to generate descriptive prediction equation of dielectric constant on small sample sets

Jiashun Mao [a,1], Amir Zeb [b,1], Min Sung Kim [c], Hyeon-Nae Jeon [c], Jianmin Wang [a], Shenghui Guan [d], Kyoung Tai NO [a,*]

[a] College of Integrative Biotechnology and Translational Medicine, Yonsei University, Incheon (21983), Republic of Korea
[b] Department of Natural and Basic Sciences, University of Turbat, Kech, Turbat, Balochistan (92600), Pakistan
[c] Department of Biotechnology, College of Life Science and Biotechnology, Yonsei University, Seoul (03722), Republic of Korea
[d] Department of Biology, School of Life Sciences, Southern University of Science and Technology, 1088 Xueyuan Avenue, Shenzhen (518055), Guangdong, China

ARTICLE INFO

Keywords:
Dielectric constant
Data-driven framework
Explainable
Small sample sets
Variable relationship network
QSPR

ABSTRACT

Dielectric constant (DC, ε) is a fundamental parameter in material sciences to measure polarizability of the system. In industrial processes, its value is an imperative indicator, which demonstrates the dielectric property of material and compiles information including separation information, chemical equilibrium, chemical reactivity analysis, and solubility modeling. Since, the available ε-prediction models are fairly primitive and frequently suffer from serious failures especially when deals with strong polar compounds. Therefore, we have developed a novel data-driven system to improve the efficiency and wide-range applicability of ε using in material sciences. This innovative scheme adopts the correlation distance and genetic algorithm to discriminate features' combination and avoid overfitting. Herein, the prediction output of the single ML model as a coding to estimate the target value by simulating the layer-by-layer extraction in deep learning, and enabling instant search for the optimal combination of features is recruited. Our model established an improved correlation value of 0.956 with target as compared to the previously available best traditional ML result of 0.877. Our framework established a profound improvement, especially for material systems possessing ε value >50. In terms of interpretability, we have derived a conceptual computational equation from a minimum generating tree. Our innovative data-driven system is preferentially superior over other methods due to its application for the prediction of dielectric constants as well as for the prediction of overall micro and macro-properties of any multi-components complex.

1. Introduction

Computer-aided material designing has been extensively reported as an emerging method with promising potential and widely applied in the discovery of new nanomaterials such as super graphene oxide, metal-organic framework (MOF), healthcare devises, automobile industries, environmental sciences, power generating plants and modern agriculture technologies [1, 2]. This stat-of-the-art approach has also contributed to global demands for energy storage and delivery, thereby accelerating the designing of efficient electrolyte solvents to improve battery storage and transmission efficacy [3, 4].

The dielectric constant (DC, ε), also known as the relative electrostatic permittivity, refers to the capacitance of a substance with respect to vacuum. Dielectric constant (hereafter ε) plays a fundamental role in reaction prediction, solvation free energy model, chemical reactivity analysis, and theoretical studies of solvents [5, 6, 7, 8, 9]. According to physical theory, the enhanced capacitance of dielectric materials is induced by a disparity in the orientation of metal's electrical charge density in response to the applied electrostatic field. Therefore, ε can also measure the polarization rate of the materials. In principle, charge-oriented polarization pattern is classified into two types. i) rotational polarization and ii) orientational polarization. Upon hot driving, the systematic pattern of permanent dipoles of a molecule rearrange into a randomly-scattered dipoles, which may produce distorted polarization. When a molecule generally experiences orientational polarization, the applied electrostatic field influences several molecular properties including bond length, bond angle, electron distribution, and many more. In order to measure the polarity of individual compound, ε can be used to

* Corresponding author.
E-mail address: ktno@yonsei.ac.kr (K.T. NO).
[1] The first co-authors contribute equally to this article.

https://doi.org/10.1016/j.heliyon.2022.e10011
Received 16 January 2022; Received in revised form 13 April 2022; Accepted 15 July 2022

Table 1. List of tools and/or relevant modules to generate the primary features.

Module Name	Amount	Software/lib	Language	Link
Mordred descriptor	1826	Mordred	python	https://mordred-descriptor.github.io/documentation/master/descriptors.html
Atomic level descriptor	18	RDKit	python	https://www.rdkit.org/docs/source/rdkit.Chem.rdchem.html
Molecular level descriptor	91	rdkit.Chem.Lipinski, rdkit.Chem.rdMolDescriptors	python	https://www.rdkit.org/docs/source/rdkit.Chem.Lipinski.html, https://www.rdkit.org/docs/source/rdkit.Chem.rdMolDescriptors.html
Rdkit ML descriptor	208	rdkit.ML.Descriptors	python	https://www.rdkit.org/docs/source/rdkit.Chem.Descriptors.html
CATS 2D descriptor	210	cats2d.rd_cats2d	python	https://github.com/iwatobipen/CATS2D
Mopac descriptor	50	MOPAC	Shell command	PM7 singlet bonds mullik STATIC polar
Coulomb matrix	100	dscribe.descriptors	python	https://singroup.github.io/dscribe/latest/
NET_ATOMIC_CHARGES	3	MOPAC	Shell command	PM7 singlet bonds mullik STATIC polar
Optimized Cartesian coordinates	5	MOPAC	Shell command	PM7 singlet bonds mullik STATIC polar
MACCS Fingerprint	167	Chem.MACCSkeys	python	https://www.rdkit.org/docs/source/rdkit.Chem.MACCSkeys.html
Morgan Fingerprint	2048	AllChem.GetMorganFingerprintAsBitVect	python	https://www.rdkit.org/docs/GettingStartedInPython.html
Avalon Fingerprint	512	rdkit.Avalon	python	https://www.rdkit.org/docs/source/rdkit.Avalon.pyAvalonTools.html
Topological fingerprints	64–600	Chem.Fingerprints	python	https://www.rdkit.org/docs/source/rdkit.Chem.Fingerprints.FingerprintMols.html
Pubchem property	19	Pubchempy	python	https://pubchempy.readthedocs.io/en/latest/

describe its properties at molecular level. To make in-depth understanding of solvent's polarity, the readers are advised to refer to these articles [10, 11, 12, 13, 14].

The available theoretical equations of ε have already leveraged good contribution to the field of physical theory of substances, but each one has its own discrepancies. For instance, the Clausius-Mosotti equation is based on Deby's dielectric theory and usually available for the rare gases and few liquids with finite polarities [15]. The extensions of Onsager and Kirkwood equations have considerably improved the polarity prediction for particular fluids, but the overall reliability is significantly poor [16]. Kirkwood–Frohlich's theory involves a correlation parameter "g", which measures local ordering only, and hence cannot be employed to calculate dipoles directly [17]. The poor prediction performance of these statistical mechanics methods indicate that the influence of charge orientation and polarization have not been adequately considered, in particular for fluids that may react with solutions bearing strong chemical influences. Such fluids are strongly hydrophilic fluids with hydrogen or water. Though, the Kirkwood's theory is a well-applied practice in the field of statistical mechanics, but its association factor cannot be calculated for parameter "g" may also be caused by multi-dimensional and distorted polarization effects, which makes it difficult to explicitly consider these properties.

The QSPR prediction model constructed by Liu and Rowley is based on four descriptors including dipole-dipole moment, solubility parameter, van der Waal's area and refractive index. This model is expected to have an average absolute percentage error <3% for hydrocarbons and non-polar compounds and <18% for polar compounds. By employing this model, the compounds with ε values ranging from 1.0 to 50.0, the accuracy of the predicted values are barely graded when the experimental values are not available [18]. Nevertheless, it is not reasonable for compounds with ε values greater than 50. In parallel, this is caused by the

Table 2. List of algorithms and/or relevant modules to process the diverse set of data.

Module Name	Dimension	Data Type	Algorithm	Language	Link
MACCS Fingerprint	167->18 167->4	One hot - > Ico	CS, SVD FFT, DCT, DST, SVD	MATLAB, python	https://github.com/aresmiki/CS-Recovery-Algorithms
Morgan Fingerprint	2048->28 2048->4	One hot- > Ico	CS, SVD FFT, DCT, DST, SVD	MATLAB, python	~
Avalon Fingerprint	512->24 512->4	One hot - > Ico	CS, SVD FFT, DCT, DST, SVD	MATLAB, python	~
Topological Fingerprints	64–600->40->8	One hot- > Ico	Transformer, SVD	python	https://pytorch.org/tutorials/beginner/basics/quickstart_tutorial.html
Atom sequence	L->40->8	Se- > Ico	Transformer, SVD	python	~
Coulomb matrix	100*100->1*64->50 L* L->1*76->40	Two- > Ico	Transformer, SVD	python	~
Atom sequence with logP_MR	2*L->1*4	Two- > Ico	FFT, DCT, DST, SVD	python	Scipy.fftpack, sparsesvd
Net atomic charges with coordinates	4*L ->1*40->1*3	Two- > Ico	FFT, DCT, DST, SVD	python	~
Mopac **descriptor: polarizability matrix**	3*3->1*1 3*3->1*3	Two- > Ico	FFT, DCT, DST, SVD	python	~

L: atom sequence length; Ico: independent continuous variables; Se: sequence variables; Two: two-dimensional variables; logP_MR: the octanol-water partition coefficient and the molar refractivity for each atom in a molecule; FFT: Fast Fourier Transform; DCT: Discrete Cosine Transform; DST: Discrete Sine Transform; CS: Compressed Sensing; SVD: Singular Value Decomposition; Transformer: Transformer. encoder; ~: the same as above.

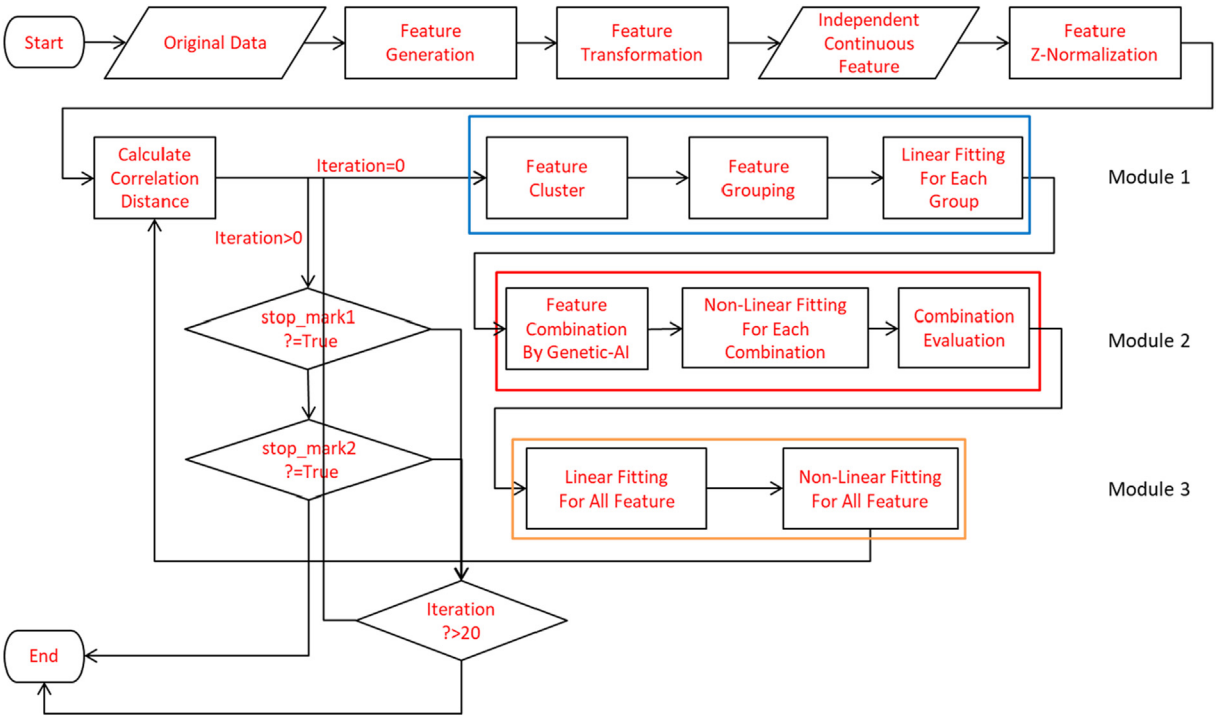


Figure 1. Flow chart of the model framework.

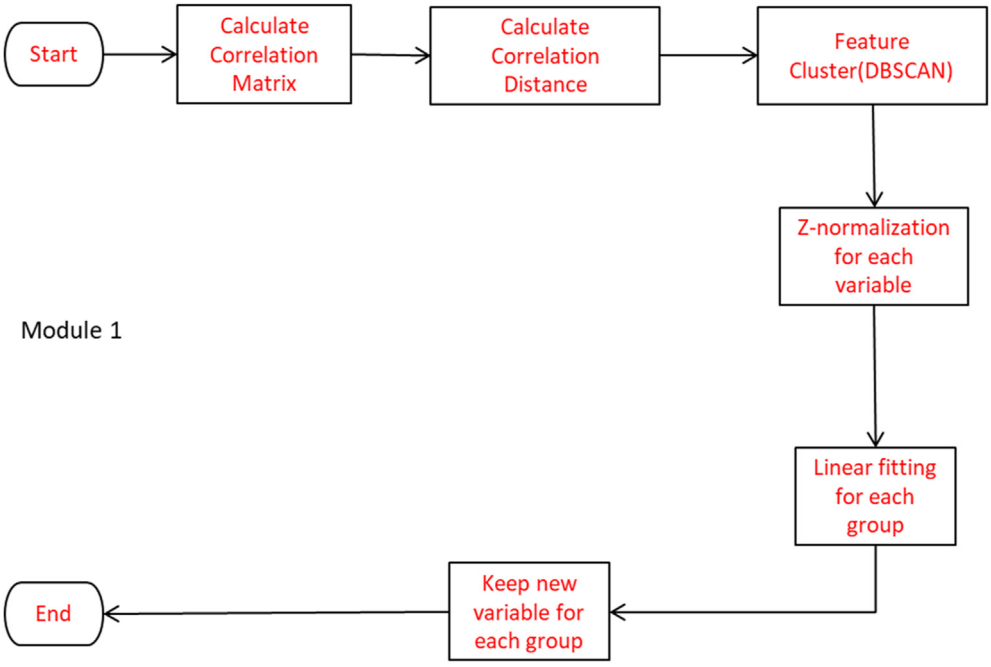


Figure 2. Linear fitting and cluster.

lack of experimental data, and the increase of the ε value, therefore, the polarity mechanism becomes much more skeptical phenomenon. It is not only associated with the polarity state of the individual molecule itself, but also related to the hydrogen bonds formed with other molecules as well as the electrostatic field to which they are exposed. It has been proved from the correlation between the dielectric constant, battery capacitance, solution polarity, and even chemical reaction rates. Hence ε property is a multi-layer concept between the microscopic level of its own atomic and electronic states and the macroscopic level of intermolecular interactions. Furthermore, Reichard has experimentally confirmed that it is not appropriate to interpret molecular-microscopic interactions exclusively by applying the concept of macroscopic dielectric constant [14].

In order to more precisely calculate the ε of compounds with larger range on small samples and understand the mechanism of macro-properties from micro level, the current study has been performed. Herein, we propose a new computational model, which is not only focusing on the micro-structural information of the molecule obtained through the quantum chemical calculations, including the partial atomic charge between each atom, the coulomb matrix between the atoms, the

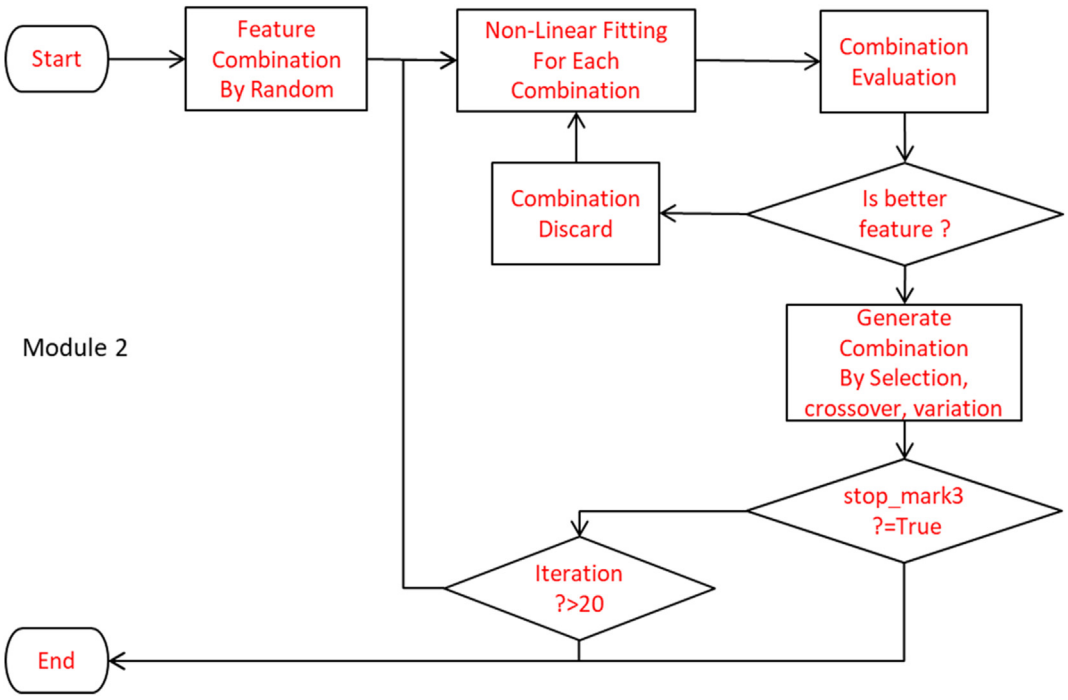


Figure 3. Non-linear fitting with genetic algorithm.

Table 3. Comparing the performance of single layer ML algorithms with different feature subsets.

Data set	Train	R2	Medae	Msle	Mse	Mae	Evs
1 xgboost in all data	**VALID**	0.93	0.21	0.01	34.35	1.72	0.93
2 gdbt in all data		0.78	0.33	0.02	109.12	2.48	0.78
3 xgboost drop coulomb matrix data(cm)		0.95	0.26	0.01	8.93	1.29	0.95
4 gdbt drop cm data		0.93	0.33	0.03	13.15	1.47	0.93
5 xgb drop cm and df_smile_seq_number.csv		0.82	0.72	0.09	33.60	2.65	0.82
6 gdbt drop cm and df_smile_seq_number.csv		0.71	0.81	0.08	53.15	2.66	0.72
7 xgb only df_smile_seq_number.csv		0.95	0.21	0.02	9.65	1.32	0.95
8 gdbt only df_smile_seq_number.csv		0.92	0.24	0.02	15.28	1.41	0.92
1 xgb all data		0.87	1.53	0.13	33.70	3.08	0.87
2 gb all data		0.83	1.13	0.11	43.61	3.13	0.83
3 xgb drop cm		0.80	1.04	0.10	50.07	2.96	0.80
4 gb drop cm		0.87	1.30	0.11	32.36	2.87	0.87
5 xgb drop cm and df_smile_seq_number.csv		0.31	2.77	0.28	176.43	5.97	0.34
6 gb drop cm and df_smile_seq_number.csv		0.41	1.74	0.19	149.74	4.89	0.45
7 xgb only df_smile_seq_number.csv		0.75	1.26	0.13	63.54	3.27	0.75
8 gb only df_smile_seq_number.csv		0.87	1.31	0.12	32.73	2.87	0.87

R2: coefficient of determination; Medae: median absolute error; Msle: mean squared log error.
Mse: mean squared error; Evs: explained variance score.

dipole moment at different orientations, and the molecular orbital energies, but also introducing macroscopic properties on the local chemical groups, hydrogen bond acceptors and donors, aromatic rings, the accessible surface area and overall volume of the molecule.

1.1. In-depth analysis

Since ε has a strong correlation with electrostatic field, therefore, researchers have developed the correlation associating ε with other measurable attributes. The relationship of ε with refractive index of non-polar molecules and dipole moment (μ) are well-known and ascertained from the previously established theories [18]. Furthermore, other empirical correlations have also been tested. For example, Papazian and Holmes established strong correlation between the ε and surface tension in a comprehensive simple correlation [19, 20]. Paruta and coworkers investigated strong correlation between the ε and solubility parameters [21]. Arnoldus HF unveiled the relationship between the surface and dipole moment in interface vicinity [22]. Paruta correlations were found to be profoundly useful for hydrogen-bonded chemicals [21]. Overall, these correlations provide useful but approximate estimates of ε, but do not establish accurate predictive equations.

The evidence of wide-range correlations between the aforementioned physicochemical properties implies that there must exist some commonalities, which come from a microscopic level relative to the molecule as a whole. Since, correlation is the only major parameter and can be obtained from observational data, thus, we determine the features of molecules at atomic and quantum chemical levels from these microscopic concepts. Such features include Coulomb matrix between the atoms, the

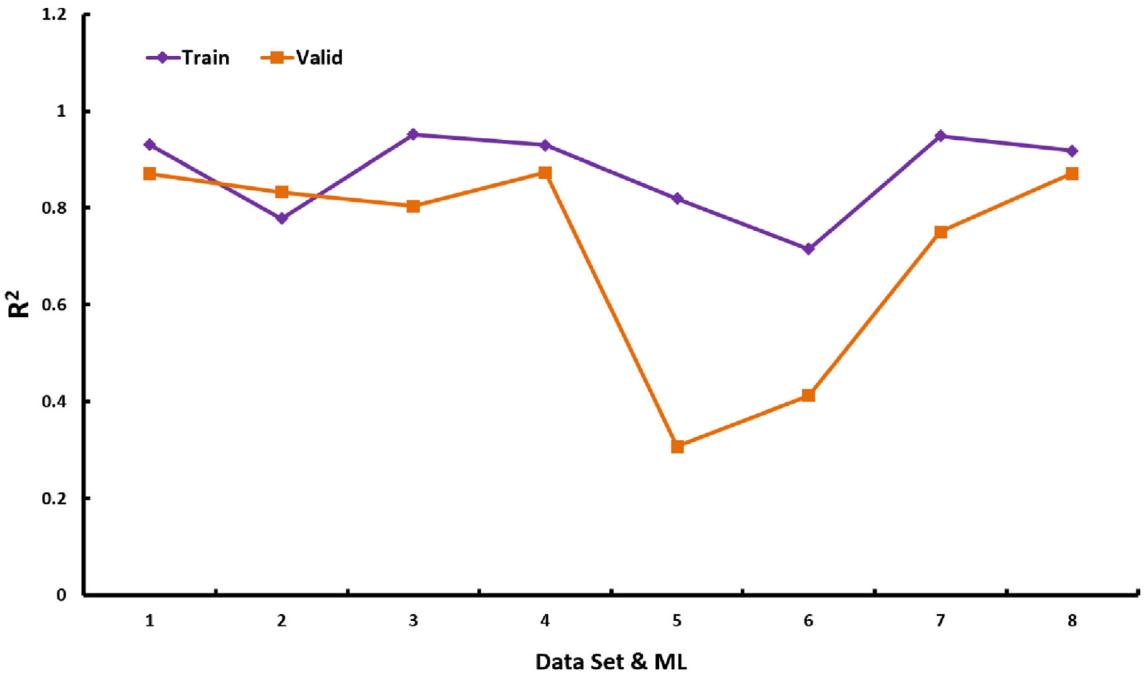


Figure 4. Different performance of R^2 on the training and validation sets.

dipole moment matrix of molecule oriented in XYZ dimensions, the HOMO and LUMO gap energy from molecular orbital theory, the distribution of atomic charges, and the molecular fingerprints based on the functional group(s) of fragments. Such information covers the most essential characteristics of each molecule and has multi-scale description from microscopic electronic and atomic states to fragment-based chemical functional group(s), the entire molecular orbital energy, and dipole moment distribution. We explore strong correlation between these features in different scales, and answer a plethora of questions including whether the properties on the same scale with strong correlation have common features at the bottom scales, and what is the true relationship? The skepticism behind this philosophy is very inspiring for us to construct the computational equation which could answer the aforementioned queries.

Although the prediction target in this study is the dielectric constant, the predictions of solvent-related properties (such as solubility, dipole moment, pH value, acidity and alkalinity, ADMET properties), drug-like properties (such as pKa (negative log of the dissociation constant), IC_{50}, logP (the octanol-water partition coefficient)) and chemical reaction-related properties (such as reaction rate, catalytic capacity, Kd (the equilibrium dissociation constant)) are also lies underneath this skepticism.

We strongly encourage the wide-applicability of our methodology, if the new target is highly correlated with our target by employing the same features, because they must have some common characteristics. Furthermore, it is possible to find exactly on which underlying features or weight, the two targets differ slightly, thus giving a new interpretation of these properties.

1.2. Main idea

It is noted that our model is quite different from traditional QSP (A) R/ML models. The core purpose of such models is always based on the training of ideal matrices and assures the feature's suitability and importance for the target. However, the mutual relationship between features and description of the highly abstracted features from micro features on a layer-by-layer scale is the most challenging issue. To address this issue, our method combines linear fitting on local features and global combination of the features, and searches most optimal descriptive relationship among the features through the layer-by-layer directed iteration. Moreover, we calculate correlation coefficients between the generated features and target to decide whether the iteration ends or not? Since our core idea is to combine differentially originated features to generate new and more abstracted features, thus, the supervised machine learning algorithm is regarded as a feature encoder, a kind of generator. Therefore, the entire framework is completely based on generative models, only considering the concept of linear and non-linear transformations. In other words, we consider the learning process of how the macroscopic features are developed from microscopic features, which is fundamentally different from the traditional QSP (A) R thinking, as well as the generative models in deep learning. In fact, constructing stronger features (have the higher correlation with the target than each composed features) from the underlying features through layer-by-layer iteration is a kind of greedy strategy. In order to prevent the generated new features from falling into the local optimum, we have also introduced the genetic algorithm for global search, and continuously change the features combination and evaluation strategies. When the newly generated features have the specified correlation coefficient with target, or when iterations reach the specified number, the program terminates.

2. Method

In this section, we have described the original features extracted from a small molecule, and ultimately explained the significance of choosing these features to be obtained. The formation of original features is diverse from sequence to multi-dimensional matrix and from fixed vector to indefinite vector in terms of length. However, we need the unified form of input data. Herein, we also provide the adjective method for transformation. Finally, the entire input variables are either independent continuous variables or Boolean variables.

After preprocessing the features, all input data are incorporated into the framework for formal processing, including the calculation of correlation matrix between features, distances between the variables, and grouping features by density clustering, linear fitting for reducing redundancy of variables, non-linear fitting to generate and evaluate new combination variables, and iterative computation by combining local greedy and global directed optimization until the steady state or threshold value is obtained.

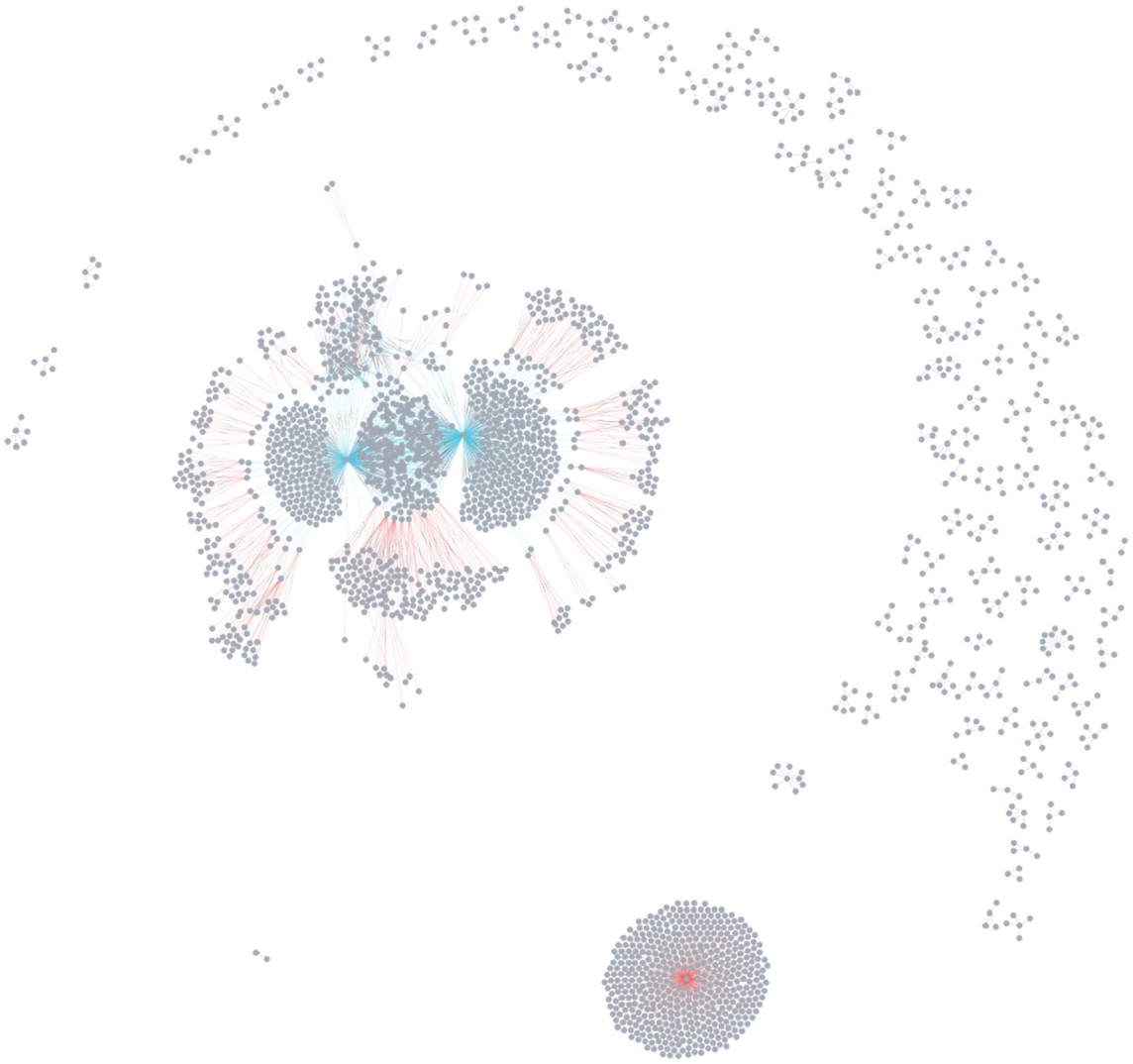


Figure 5. Relationship network composed of partial nodes and edges.

2.1. Features preparation

The ε values established through experiments are available for the frequently used chemicals, yet there are several industrially important chemicals, for which the ε measurements are not available in the literature. Fortunately, we retrieved our original data from this article (http://www.rsc.org/suppdata/c9/cp/c9cp01704f/c9cp01704f3.xlsx) [15]. Thereafter, according to the CAS registry number and compound name, we generated molecules in their SMILE and PDB formats. Subsequently, the electrification of all molecules was detected and manually processed to make them electrically neutral. Additionally, the MOPAC program was used for quantum chemistry optimization and property calculation, while the RDKit lib and Mordred tool were used for molecular and atomic level features calculation, and molecular fingerprint calculations. In order to reduce the dimensionality of the features, we performed Fourier and cosine transformers on molecular fingerprints, Coulomb matrices and polar distribution features. In parallel, we introduced sparse transformations for further compression, and performed embedding coding which was based on Transformers model [23, 24, 25, 26].

Initial dimensions for original features were between 5321 and 5857 due to the non-fixed vector. After preprocessing, the total number of features was reduced to 3648. (The physical and chemical meaning and name of all the features can be found in supplement Part 1). To calculate these features, refer to Table 1. Herein, the features were categorized as Mordred descriptors, atomic level descriptors, molecular level descriptors, Rdkit ML descriptors, CATS 2D descriptors, MOPAC descriptors, Coulomb matrix, NET_ATOMIC_CHARGES, Optimized Cartesian coordinates, MACCS Fingerprint, Morgan Fingerprint, Avalon Fingerprint, Topological fingerprints, and PubChem properties. Most of these features are independent continuous variables. For example, Mordred descriptors, carbon atom number, molecular weight, predicted values of LogP (XLogP [27], ALogP [28]), properties calculated from two-dimensional (2D) structures (Eccentric connectivity index [29] and three-dimensional (3D) structures (charged partial surface area (CPSA) [30]), and quantum mechanics-based properties [(highest occupied molecular orbitals (HOMO), lowest unoccupied molecular orbitals (LUMO), orbital energies)] etc.

Although, both open-source and proprietary software have been developed for calculating molecular descriptors, such as PaDEL-Descriptor [31], ChemoPy [32], PyDPI [33], Rcpi [34], Cinfony [35], but each of them has advantages and disadvantages. In order to easily integrate with our python framework, as well as the advantages of themselves, we chose Mordred as our core software.

2.2. Further processing

Special transformations need for discontinuous features including the features based on sequences (atomic symbol sequences, topological fingerprinting, Coulomb matrix, optimized 3D coordinates, atomic sequence of partial charge distribution, atomic sequence of partial logP

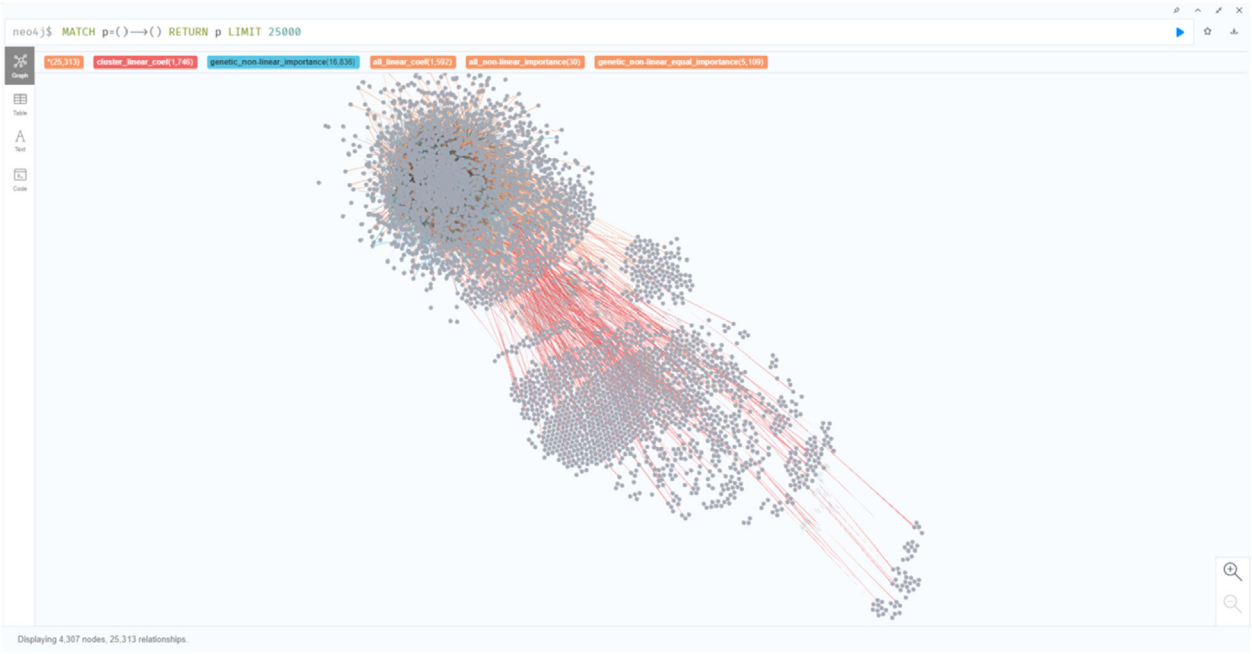


Figure 6. Relationship network for randomly selected 25,000 nodes.

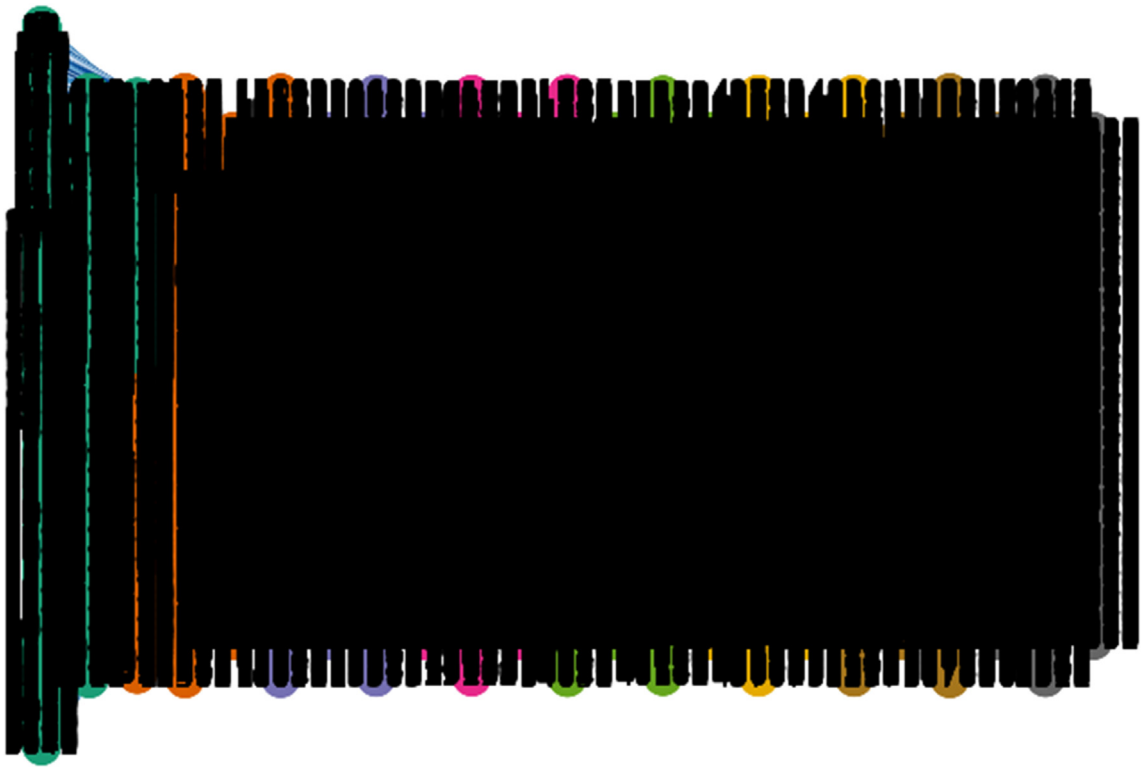


Figure 7. Variable relationship network for 21 layers with more than 250,000 nodes.

and MR distribution) and the features based on quantum chemistry (contribution of dipole moment for heat of formation in 3-D coordinate system and contribution of dipole moment for polarizability).

Interestingly, we have a number of mathematical theories for aforementioned special transformations, including Digital signal processing, Compressed Sensing, Matrix Decomposition, Natural Language Processing (NLP), and Digital Image Processing method to process these data. We have applied algorithms such as Fast Fourier transform (FFT), Discrete Cosine transform (DCT), Discrete Sine transform (DST), Compressed Sensing algorithm (CS), Singular Value Decomposition (SVD), Transformer (a deep learning model that adopts the mechanism of attention, differentially weighing the significance of each part of the input data, It is used primarily in the field of NLP), and Convolutional Neural Network (CNN) for the special transformations.

First of all, the discontinuous data including sequence data, two-dimensional matrices, and one-hot vectors, are handled by the

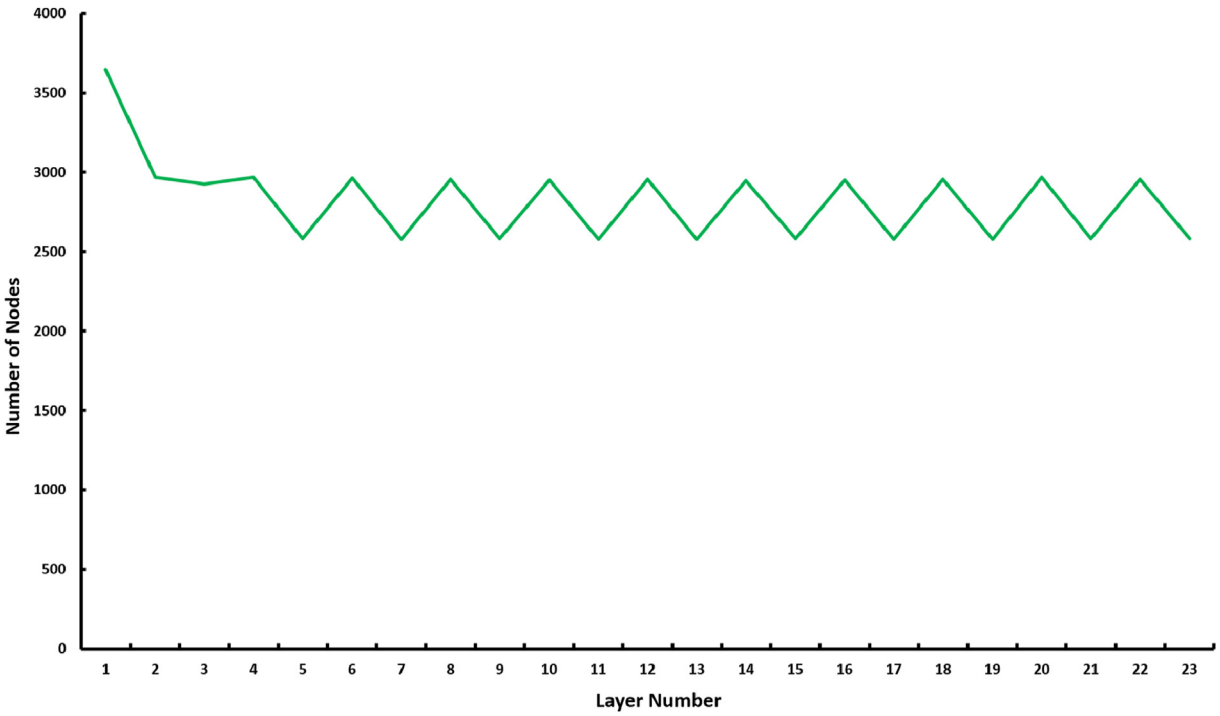


Figure 8. The change in the number of variables indicates that a large number of similar variables were generated in the previous step 2.

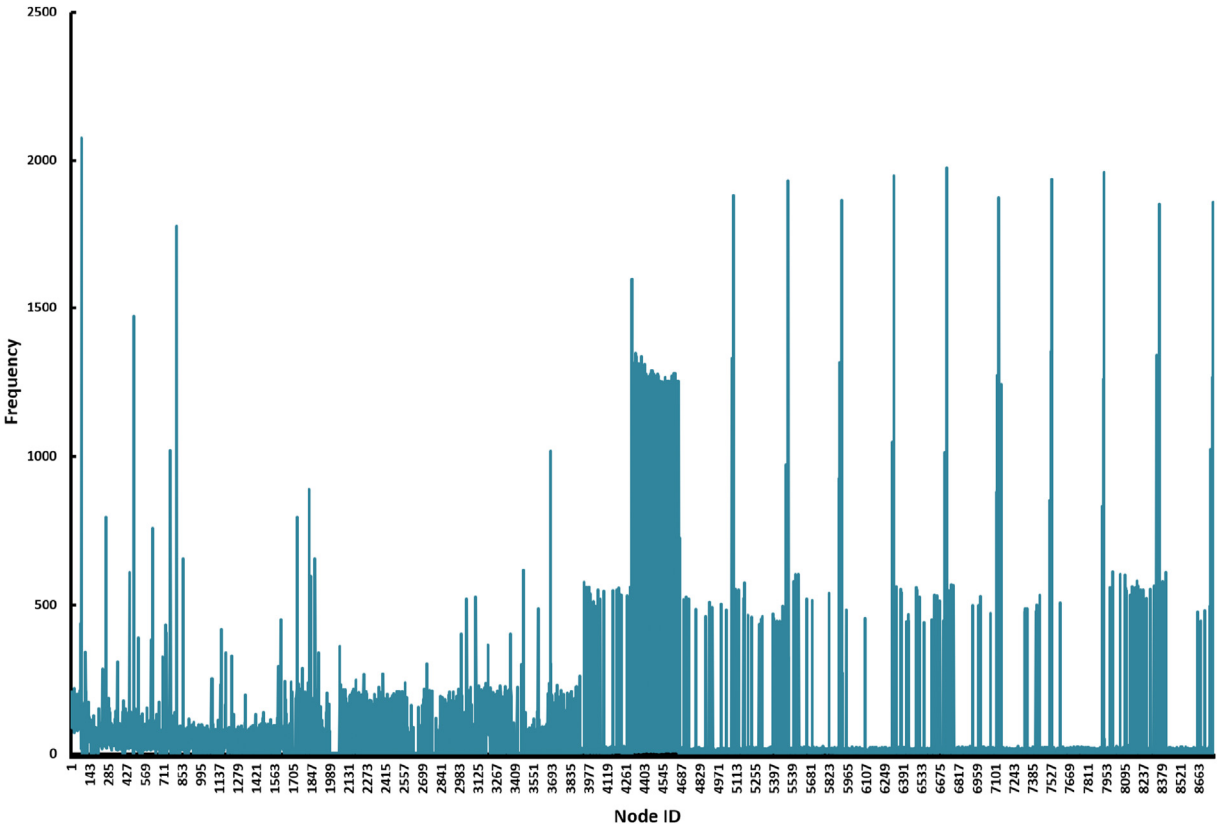


Figure 9. The frequency of occurrences of each variable in all layers.

aforementioned algorithms to capture specific perspective information, and finally converted uniformly to a representation in dense continuous space.

For one-dimensional MACCS, Morgan, and Avalon fingerprints, since they mainly express the existence of a particular type of chemical group in a molecule and do not involve atomic contextual sequence information. Therefore, it is more reasonable to treat them as one-dimensional sparse signals. Furthermore, according to the theory of compressed perception, they can be compressed into a denser expression without changing their original meaning, which is a one-to-one mapping.

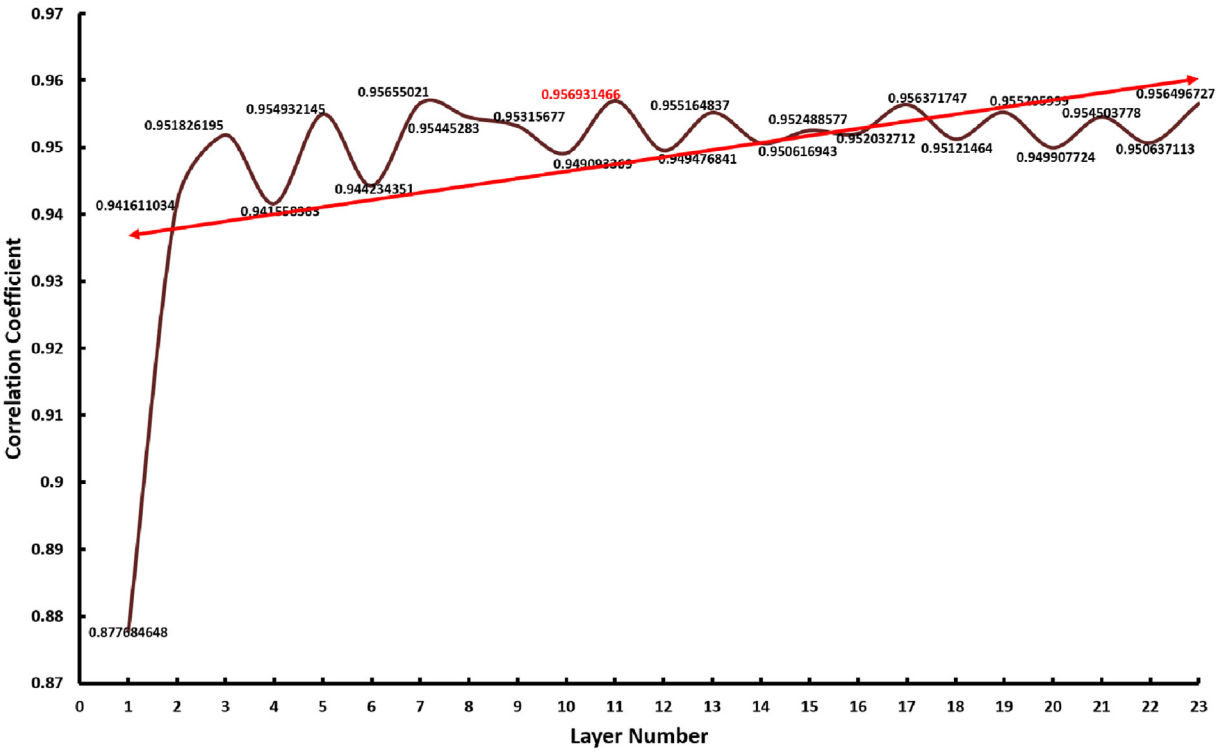


Figure 10. The highest correlation value for each layer in the test set forms a spiral curve.

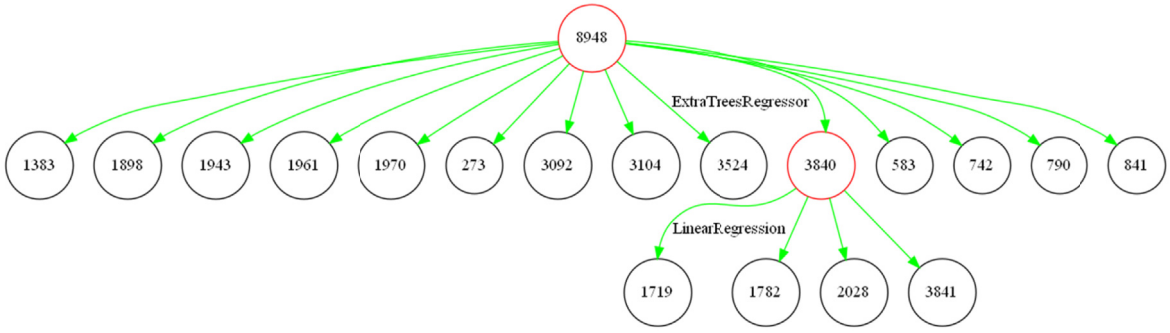
1. 8947: 21_iter_0_tree_model_ExtraTreesRegressor_2566_11
2. 1692: fingerprints_morgan_2047 fingerprints_morgan
3. 1702: numheteroatoms mol_level_descriptor1_rdMolDescriptors_Properties
4. 1898: peoe_vsa12 rdkit_descriptors_2D
5. 1943: vsa_estate3 rdkit_descriptors_2D
6. 1970 : fr_nh0 rdkit_descriptors_2D
7. 3092 : ic0 mordred_descriptors
8. 3104 : sic0 mordred_descriptors
9. 427: fingerprints_avalon_224 fingerprints_avalon
10. 499: fingerprints_avalon_300 fingerprints_avalon
11. 614: fingerprints_avalon_418 fingerprints_avalon
12. 732: fingerprints_maccskeys_58 fingerprints_maccskeys
13. 742: fingerprints_maccskeys_68 fingerprints_maccskeys
14. 746: fingerprints_maccskeys_74 fingerprints_maccskeys
15. 793 : fingerprints_maccskeys_121 fingerprints_maccskeys
16. 841 : fingerprints_morgan_6 fingerprints_morgan

Figure 11. 21_iter_0_tree_model_ExtraTreesRegressor_2566_11.

Certainly, relevant information could also be extracted using signal processing related methods.

For topological fingerprints, the topological distribution of atoms in each molecule is different, which results in length variation. Since, it represents the contextual information of atoms in the sequence, therefore, it is more appropriate to deal with it as a sequence.

Regarding the Coulomb matrix, which is composed of Coulomb forces between any two atoms in a molecule, hence it could be treated as a



1. 8948: 21_iter_0_tree_model_ExtraTreesRegressor_2566_12
2. 1383: fingerprints_morgan_1283 fingerprints_morgan
3. 1898: peoe_vsa12 rdkit_descriptors_2D
4. 1943: vsa_estate3 rdkit_descriptors_2D
5. 1961: fr_al_oh rdkit_descriptors_2D
6. 1970: fr_nh0 rdkit_descriptors_2D
7. 273: fingerprints_avalon_58 fingerprints_avalon
8. 3092: ic0 mordred_descriptors 0-ordered neighborhood information content
9. 3104: sic0 mordred_descriptors 0-ordered structural information content
10. 3524: df4_logp_all_2 fft_relavant
11. 3840: 0_96_mol_level_calcnumheteroatoms_lipinski_numheteroatoms_nhetero_mid_h_lineRg_LinearRegression
12. 583: fingerprints_avalon_386 fingerprints_avalon
13. 742: fingerprints_maccskeys_68 fingerprints_maccskeys
14. 790: fingerprints_maccskeys_118 fingerprints_maccskeys
15. 841: fingerprints_morgan_6 fingerprints_morgan
16. 1719: mol_level_calcnumheteroatoms mol_level_descriptor1
17. 1782: lipinski_numheteroatoms mol_level_descriptor_3_Lipinski
18. 2028: nhetero mordred_descriptors
19. 3841: 0_96 (intercept distance) intercept distance

Figure 12. 21_iter_0_tree_model_ExtraTreesRegressor_2566_12.

whole for SVD transformation to simplify the expression. It can also be regarded as certain type of embedding expressed sequence, because it contains implicit information about the sequence of atoms.

The two-dimensional matrices such as the charged properties and logP, MR, dipole moment, and polarity matrix of each atom, reflect a signal of the molecule as a whole. Therefore, they are considered as two-dimensional signals and hence extracting information by signal processing related methods.

In addition, different information is extracted and simplified. Variables of different lengths become fixed-length variables and multi-dimensional variables become one-dimensional variables. In principle, one-dimensional sparse signal matrices can be well-compressed with the CS algorithm. If the matrix is not one-dimensional sparse, it is firstly converted to a one-dimensional sparse matrix by FFT (including Z-transform, DCT (symmetric)). For details, refer to Table 1. It should be noted that Principal Component Analysis (PCA) method for redundancy and noise reduction will actually change the information contained in the original (directly change a group of bases for representation, the data need to be normalized minus the mean divided by the variance, i.e. z-score normalization) dataset, while SVD is only used to find a more simplified approach to express the original signal, without normalization. So it will not alter the information contained in the original dataset. In contrast, the method such as FFT represents the same information from different perspectives, such as from time domain to the frequency domain, or vice versa. This can make particular features more distinct in different dimensions. Conclusively, for dimensionality reduction, SVD is better than others, while for de-redundancy, PCA method is superior, and for sequence embedding representation, Transformer encoder is the best approach.

In the final step of data pre-processing, normalization, alignment of variable dimensions, filling in zeros for insufficient dimensions, and merging all variables were performed. Eventually, all variables incorporated into the model were continuous variables, and conduct Z-score normalization so that the linear and non-linear transformations are not influenced by the scale between different variables:

The standard Z-score of a variable X is calculated as Eq. (1).

$$Z = (X - \mu)/\sigma \tag{1}$$

Here μ is the mean of X, and σ is the standard deviation of$X$.

As shown in Table 2, the integration of different modality information is implemented by uniformly transforming different forms of data into continuous variables and reducing the dimensionality.

Indeed, it is easy to introduce artificial preferences and parameters when handling discontinuous data, for instance, the selection of compressed dimensions, the accuracy between compression and reconstruction, and the selection of compression parameters, etc. But these processes are determined by personal experience and calculation accuracy. We argue that the processing of these data is fair as long as all features are processed with the same parameters, or the same accuracy, and of course we need the same parameters processing when predicting.

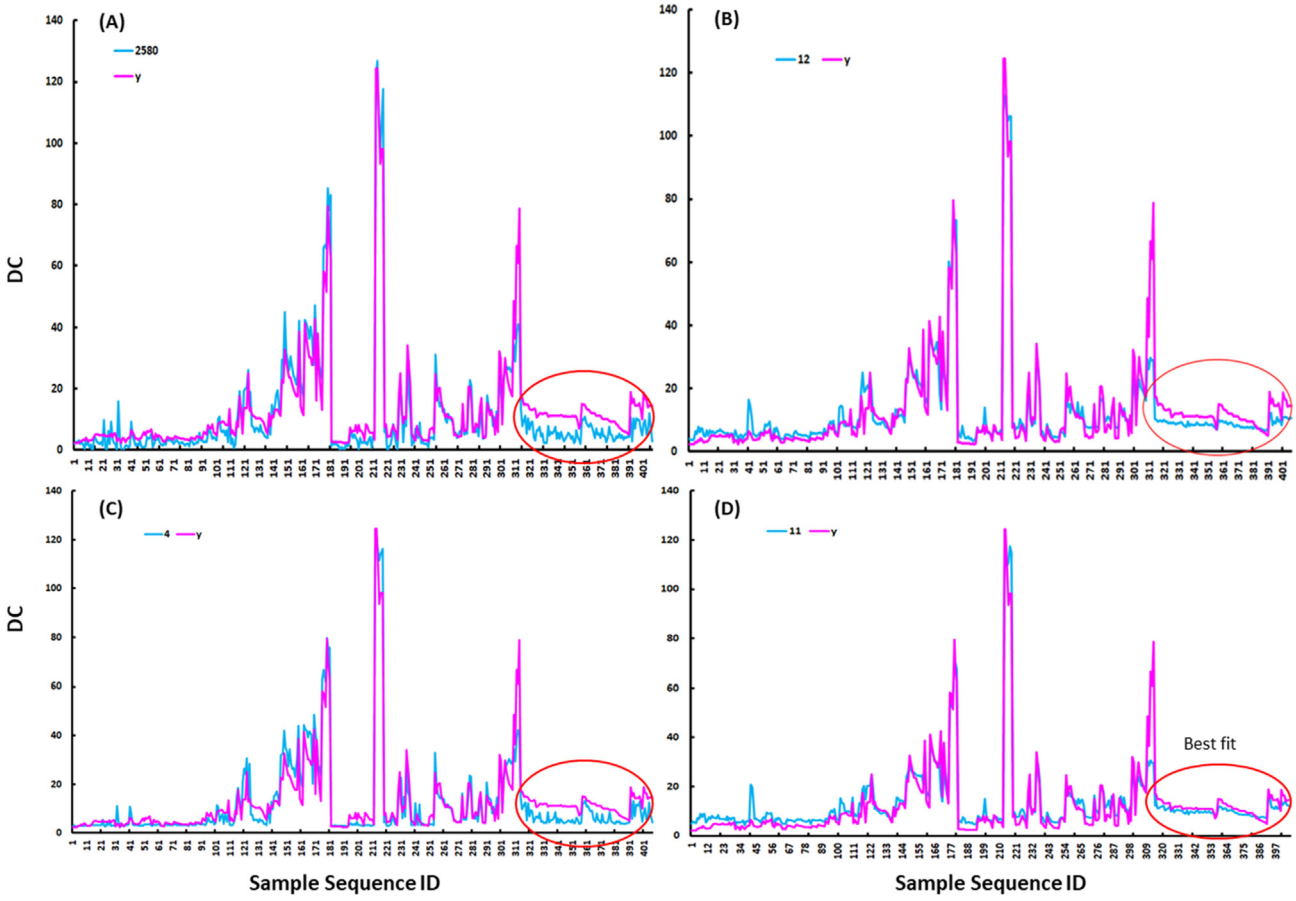


Figure 13. Comparison of the distribution of the four new variables with the target variable.

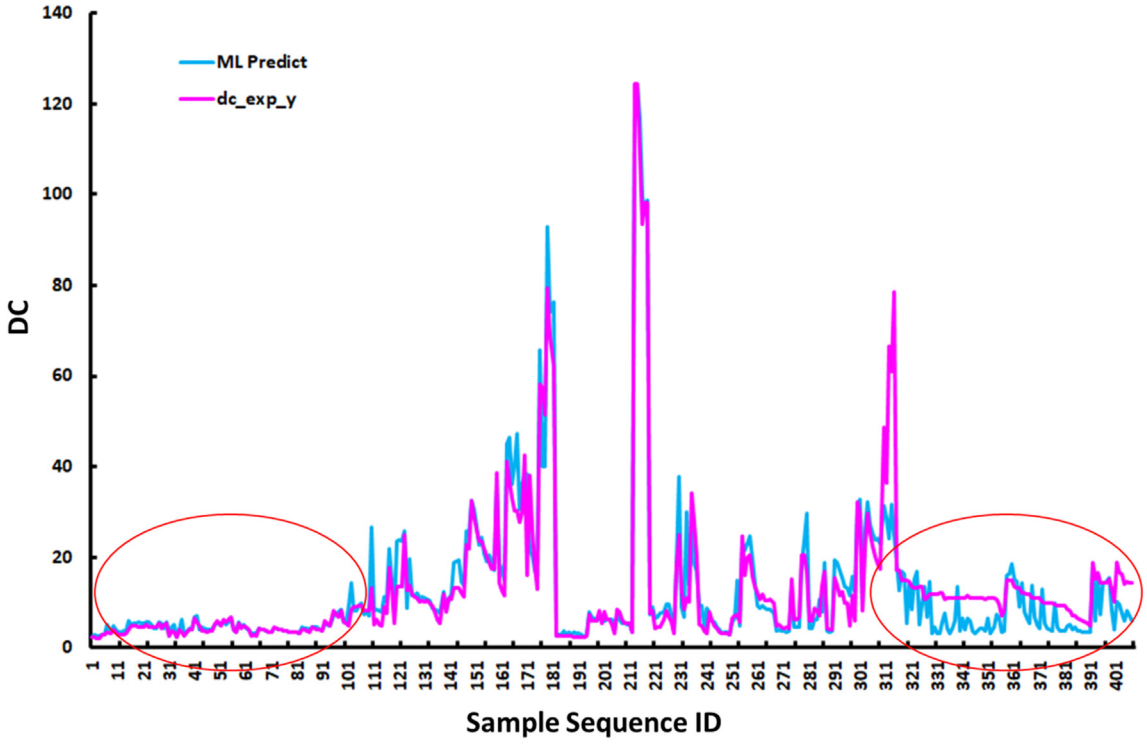


Figure 14. Best prediction based on ML (traditional QSA/PR) and real DC, the horizontal coordinate is the Sample sequence id, and the vertical coordinate is the DC value.

Initially, we have no idea which modal information is relevant to the target, and we can only provide as much information as possible in different modalities and in different dimensions, since it directly determines the volume of feature space, and the upper limit of the subsequent model.

2.3. Model framework

Through pre-processing of features and a further transformation of information, all input data are successfully turned into uniform independent continuous variables, which is a prerequisite to enable the next step of linear and non-linear combination (Figure 1). Module 1 is correlation distance-based clustering, which results in the reduction of linear redundant variables. Module 2 is the combination of feature subspaces, which is optimized based on the genetic algorithm, resulting in the generation of new variables by applying non-linear operators. Reserved new variables should be more relevant to the target. Otherwise, the non-linear encoding of this combination of features' subspaces would be dropped by the genetic algorithm. Final module encodes all variables with linearity and non-linearity and these new variables are not discarded whether or not they are closer to the target variables. The intention behind this approach is to preserve the globality as much as possible, yet these variables are usually closer to the target in terms of correlation.

The hyperparameters for our model are set as follows: in variance filtering, threshold = 0, filling Nan to 0, applying StandardScaler before trainning, in clustering algorithm DBSCAN, eps = 0.01, MinPts = 3. All linear (LinearRegression, SGDRegressor, RidgeCV, LassoCV, ElasticNetCV, BayesianRidge, LinearSVR) and nonlinear (KernelRidge, BaggingRegressor, ExtraTreesRegressor, RandomForestRegressor, GradientBoostingRegressor, HistGradientBoostingRegressor, XGBRegressor, LGBMRegressor) model parameters use default values, and the evaluation indicators apply r2 score, in the genetic module, DNA_SIZE = the dimension of features, POP_SIZE = 20, CROSSOVER_RATE = 0.8, MUTATION_RATE = 0.1, N_GENERATIONS = 20, final_target = 0.999. The Spearman correlation coefficient is used when calculating the correlation distance between the variables, while the Pearson correlation coefficient is used when calculating the correlation between all the features and the target variables.

Considering these three modules as a layer, and repeating in the next iteration, there are two criteria to judge whether to stop the iteration or not. First, whether the number of iterations meets the specified number, and second, whether the correlation coefficient between at least one variable and the target reaches the specified threshold in the test set, which is set to 0.999. The whole idea of the modular framework is derived from the conclusions in our previous review article [36]. Detail of module 1 and 2 (Figure 1) are shown in Figures 2 and 3, respectively. It is observed that the number of variables will be reduced after passing module 1, which is due to the merging of variables with high correlation. According to the linear correlation theorem, it does not lose the information of the merged variables and enables to reconstruct their component variables by Independent Component Analysis (ICA) or PCA.

After passing through the non-linear oriented combinations in Module 2, only retain the combinations of stronger correlations with target variables. This also achieves a layer-by-layer improvement of the intermediate variables at the abstraction level and thus can simulate the function of the integration operator from the non-linear perspective. It is observed from the experimental results that the newly generated non-linear variables, in terms of correlation, are indeed closer to target variables and show an oscillating upward trend in their maximum correlation coefficients in all layers (Figure 10).

In order to obtain better robustness of the experimental results, we setup six different sets of parameters, which are as follows.

1: fit = 1,0,1,0,1,0,1,0,1,0 and eps = 0.05 with all data
2: fit = 1,0,0,0,0,0,0,0,0,0 and eps = 0.05 with all data
3: fit = 0,0,0,0,0,0,0,0,0,0 and eps = 0.05 with all data
4: fit = 0,1,0,1,0,1,0,1,0,1 and eps = 0.05 with all data
5: fit = 1,0,1,0,1,0,1,0,1,0 and eps = 0.10 with all data
6: fit = 1,0,1,0,1,0,1,0,1,0 and eps = 0.05 with all data except for coulomb features

Fit = 1 means that the best matrix in the training set is chosen among the numerous tree models in Module 2, while 0 means that the best matrix in the test set is chosen. The "eps" denotes the correlation distance setting in the clustering algorithm DBSCAN, where default is 0.05, and this distance ranges from 0 to 1. With all of the above six conditions completed for training, a comprehensive analysis is performed.

First, four variables with the greatest relevance to the target are taken as root nodes of the tree. From the entire hierarchical network of variables, a minimum generating tree is retrieved and each leaf node is the original input variable. Then, according to the meaning and category information of the leaf nodes and correlation matrix, the intermediate generating variables can be explained. Finally, a new estimation model for DC is given by the theoretical experience and computational complexity of the minimum generating tree. By analyzing the predicted values and experimental values, our method shows promising performance on predictions of $\varepsilon > 50$ as compared to the single layer ML approach.

3. Results and analyses

3.1. Single layer ML

To compare the performance of our proposed new framework, the results obtained from the single layer ML method were analyzed. The specific experimental data are shown in Table 3. The hypothesis of experimental conditions is given below.

(a) By changing features space to observe the matrices change with the same algorithm.
(b) By changing algorithms with the same features to observe the matrices variation.

Table 3 shows that maximum R^2 on validation set is about 0.87, while on the training set, it can reach a maximum of about 0.95. Observing the four groups (1–2, 3–4, 5–6, and 7–8) in the training set (for the same data), the performance of xgb is better than gbdt. However, in the validation set, gbdt outperforms xgb in the other three groups except for group 1–2, which indicates that xgb has best fitting ability in part, while generalization ability is uncertain.

Again, observing the R^2 of 1–3 and 5–7 of the training set, it is evident that different features' subspace leads to differences of prediction performance under the same algorithm. This indicates that there exists a special binding relationship between feature subspace and the prediction target. It also supported assumption that introducing other irrelevant features will degrade the performance efficiency. For instance, the comparison between 1 and 3, after removing the Coulomb matrix data (cm) features, R^2 rises from 0.92 to 0.95, however, in the test set, it decreases from 0.87 to 0.80 in contrast.

By analyzing experimental results in Table 3, we assume that the correlation value with target may be the upper limit of accuracy of the model determined by the input data. Therefore, we argue that correlation value is an essential matrix, which is ignored by the traditional QSP (A) R ML approaches.

Next, we performed an evaluation matrix for regression model on the training and test sets and obtained R^2 cure (Figure 4). It is demonstrated that differences in algorithms, training and test sets, and features subspaces, contributed to the final R^2 matrix. Therefore, we strongly recommend that developing a QSP (A) R model, all attempts should be made according to practical situation. Because there is no universal algorithm that is perfect for all the problems, no subset of features that is

appropriate for all predictive targets, and no algorithm that could work well on both training and test sets.

Based on the above analytical results of single-layer machine learning experiments, it is cleared that an intrinsic special mechanism for adaptive matching between the algorithm and feature subspace is required.

In our framework, both Module 1 and Module 2, have applied a number of algorithms for selection, and tested on training and validation sets, respectively. This accomplishes self-adaptive matching for specific feature subspace and specific prediction target.

For the linear fitting Module 1, the candidate models include Linear Regression, SGD Regressor, Ridge CV, Lasso CV, ElasticNet CV, Bayesian Ridge, and Linear SVR. These models can be found in python library called sklearn, which are commonly linear fitting models. For the non-linear fit, taking into account the running time, we chose KernelRidge, BaggingRegressor, ExtraTreesRegressor, RandomForestRegressor, GradientBoostingRegressor, HistGradientBoosting Regressor, XGB Regressor, and LGBM Regressor. These candidate models are trained in parallel on the same dataset and the model with best performance matrix is finally considered as the encoding model. Fit = 1 means that a 0.2 percentage of the dataset on the training set is chosen as the criteria for selection, and fit = 0 means that the validation set is taken as the criteria for selection.

During the training period, R^2 score is our evaluation method. After training, we calculated the Pearson correlation coefficient between all the features and target variable until the maximum Pearson correlation coefficient is over 0.999 or the number of iterations is over 20. We assume that this variable is closest to target, and the target can be replaced by this variable, we just generated this variable means that we have completed this prediction task. In order to avoid the overfitting issues, we do not change parameters of all the repressors, and just operated them as mapping function.

3.2. Relationship network based on combination of the variables

The whole network generates over six million edges, yet over 99% edges have weights below 0.01. After removing the edge with weight <0.01, it filters 253,288 edges and 8952 nodes.

Figure 5 shows the partial relationship network around 2500 nodes, where the blue line is the non-linear relationship generated by Module 2. The red line represents linear fitting relationship by Module 1. The fitted relationship of the local variable after clustering has been illustrated (Figure 5, right side). The larger circle at the bottom of Figure 5 is generated by the global linear fitting. The whole relationship network has both complex non-linear relationship, simple local linear relationship, and global linear relationship, but they have no connection with each other, indicating that many variables are not associated with the target, while some are involved in both local linear and non-linear fitting as shown in Figure 6.

Since, the entire 21 layers network with more than 250,000 edges is very hard to draw, thus, we only obtained the approximate and very fuzzy network graph (Figure 7). If we plot the number of changes of nodes in each layer, it appears jagged as shown in Figure 8. Since, a large number of new variables are generated in Module 2 and large amounts of redundant variables are merged together in Module 1 in the next iteration. Therefore, there will be such a back and forth oscillation in the number of changes.

To consider the frequency of occurrence of each variable in all layers, we realized that some variables appeared very frequently while other appeared rarely (Figure 9). It should be noted that node with id < 3648 are the original feature variables. The frequencies revealed that some nodes are repeatedly integrated into new variables, indicating that these variables are very important. For the distribution relationship between these variables with the highest and lowest frequencies and the target variables, the reader can further refer to Part 2 in supplementary information.

If we consider the highest correlation with target variable in each layer together, the trend of maximum correlation between the fitting variable and ε increases in layer-by-layer pattern. In Figure 10, the global peak among all network layers appeared in layer 11, at 0.956931466, although it couldn't reach the set criterion of 0.999. This indicates the limitation of correlation between the target variables and generated novel variables from the original 3648 variables. However, this value is already significantly improved from the highest value of 0.877 in the original input variables. The spiral pattern indicates that the generated variables are temporarily suppressed due to the removal of redundancy and reappear in the next iteration. From the above analyses, we can summarize as:

- The stronger the correlation between the two variables, the more similar they are.
- The correlation between the original variables and the target variables is not strong, but after non-linear fitting, the variables are more strongly correlated with the target variables.
- By combining non-linear fit, these weak variables become stronger variables and are continuously improved through layer-by-layer iterations.
- The frequency of occurrence of a variable does not affect correlation of the variable with other variables.
- A big number of genetic combinations at the same level of iteration do not result in a significant increase in correlation.
- Low correlation variables are gradually integrated into high correlation variables with target, which indicates that the low correlation variables also include some useful information, thus making the newly generated variables stronger.

In addition, the objective of this study is not only obtaining a novel variable whose correlation is 0.956931466, but also to explore interpretation and relationship between the variables.

For interpretation, we mainly search for new variables that could most approximately be closed to the target in terms of correlation and obtained a minimum generating tree with that variable as the root node and the original variable as the leaf node, including its fitted linear and non-linear paths and weights. For the exploration of the relationship between the variables, middle variables are mainly explained and defined by their components (the original variables) or by other variables with high correlation with them from the correlation heat map.

3.3. Minimum generating tree

The variables in last layer with correlations above 0.95 were selected as a candidate:

- ['21_all_f_linear_rg_RidgeCV_2580', 0.9564967270184328],
- ['21_iter_0_tree_model_ExtraTreesRegressor_2566_4',0.9516725385817788]
  ['21_iter_0_tree_model_ExtraTreesRegressor_2566_11',0.950804537293214]
- ['21_iter_0_tree_model_ExtraTreesRegressor_2566_12',0.95061201226503]

The minimum generating tree with variable as the root node can be obtained by retracing the generation path from these variables respectively, and each leaf node must be the original variable. Herein, for variables 21_all_f_linear_rg_RidgeCV_2580 and 21_iter_0_tree_model_ExtraTreesRegressor_2566_4, they have tons of nodes, which is hard to display. From the names of the aforementioned variables, there are 10/16 molecular fingerprint features. The specific values of molecular fingerprints correspond to a chemical group, for the single layer calculation model, features based on fragment or functional group are more favorable for calculating the properties of molecule as a whole.

Compared to the single-layer calculation model in Figure 11, the tree in Figure 12 includes two layers, where the df4_logp_all_2 is incorporated as new feature from the atomic level. It indicates that these features from essential nature of the bottom layer are required for generating efficient predictions. Interestingly, such representation can be interpreted as an extension that deep learning can learn directly from the underlying 3D

coordinate's space, where traditional ML relies heavily on expert features.

To collect a more detailed analyses of performance of the predicted values on different ε regions, we compared the distributions of the above four variables with those of the target variables (Figure 13). In parallel, the distribution of prediction results of the machine learning model trained from the original input variables compared to target variables were also demonstrated (Figure 14). From the comparison in Figure 13 and Figure 14, it is cleared that the variables generated after several iterations have a better similarity of distribution to target variables in general. Yet in some local areas, traditional ML methods have also advantages, for example, the bottom-left panel of Figure 14, the fitting in red circle is clearly more accurate than our new method. Our framework is significantly better than traditional ML methods for predictions with $\varepsilon > 50$, indicating that non-linear multi-level iterations have indeed introduced new perspectives for the prediction of high values. In order to strengthen the interpretability as well as reduce the computational complexity, these minimum generating trees were rewritten to an equation which will enable our understanding and comparison with known physical experiences.

3.4. Equation generation

Based on the type of leaf node and its own physicochemical meaning, we try to interpret these intermediate nodes in minimum generating tree, combining with the empirical formula (Kirkwood–Frohlich equation), the dielectric constant (ε) of a polar associated liquid is given by the Kirkwood–Frohlich equation. For more detail, refer to references [15, 16, 18].

Though in Figure 13, we have considered that 21_iter_0_tree_model_ExtraTreesRegressor_2566_11 is the best choice to derive an equation. The equation is re-written as below:

f(x1,x2,x3…,x15) = F(Z(numheteroatoms)*0.0180910444769773,
Z(sic0)* 0.0169836141314585,
Z(ic0)* 0.0264475279022733,
Z(fr_nh0)* 0.0117640775273062,
Z(vsa_estate3)* 0.0576742145982364,
Z(peoe_vsa12)* 0.0393558545204146,
Z(fingerprints_morgan_2047)* 0.0103208569437381,
Z(fingerprints_morgan_6)* 0.1398486568677072,
Z(fingerprints_maccskeys_121)* 0.0149855944177587,
Z(fingerprints_maccskeys_74)* 0.0162316055895788,
Z(fingerprints_maccskeys_68)* 0.0163947108147842,
Z(fingerprints_maccskeys_58)* 0.0209220970182179,
Z(fingerprints_avalon_418)* 0.014709145198778,
Z(fingerprints_avalon_300)* 0.0239342453469886,
Z(fingerprints_avalon_224)* 0.012844074838388).

$Z = (X - \mu)/\sigma$, here μ is the mean of X, and σ is the standard deviation of X.

F is non-linear mapping function, an instance of the ExtraTreesRegressor model.

Finally, the overall fit is certainly better with our framework, and only slightly weak on few points. In contrast, the single layer ML has a significant overall error rate and almost impervious to boost on the basis of same input descriptors or fingerprints.

Although, this study has focused on the prediction of properties of pure compounds only, but it could also be applied to the system involving features from the micro to macro level via the layer-by-layer extraction approach, for instance, the prediction of the overall properties of solvent complexes and ionic solutions.

4. Conclusion

A primary challenge in developing accurate and reliable prediction models to predict properties in the absence of sufficient experimental data has emerged as the foremost difficulty in modern computational sciences.

How to develop interpretable and generalized models and fully exploit the features' space on small sample sets is the core target of this study. In our framework, various non-linear coding have been retained according to the actual prediction performance, resulting in adaptive matching between different ML algorithms and several features' subspaces. An abstract computational equation is generated from the minimum generating tree which is derived from the variables network.

Main conclusions and advantages of our system are:

(a) The abstract intermediate variables can bridge gap between the microscopic and macroscopic features within the molecule based on layer-by-layer iterations. Such architecture generates a hierarchical relationship network where the features are associated with each other and eventually understands the complex relationships between these features.
(b) Our approach fundamentally addresses the inconsistency between the local optimum (based on greedy combination for each layer) and the combination explosion (caused by searching in global space). Our system is smart enough to accomplish an adaptive balance between the real optimal features' subspace and matching algorithm with target.
(c) Both linear and non-linear transformations were performed for each layer, and dimensionality improvement was realized as the integral. Compare to straightforward downward partial derivative of objective function in DL, our system employs upward integration which is based on locally efficient combinations.
(d) From breadth of the features' space, our framework applies clustering based on correlation distance to filter out redundant features, and ultimately reduces the dimensionality of features' space. From the depth of the features' space, our framework reduces the search space for accelerated finding of effective combinations by applying the layer-by-layer extraction and optimizing the feature combinations via genetic algorithm.

Overall, the data-driven system by applying multi-layer iteration for depth combination based on a massive amount of available descriptors, adopting automated feature extraction methods, such as the implicit space-based method and kernel transformation-based method, perhaps represents the future for developing QSP(A)R models.

Declarations

Author contribution statement

Jiashun Mao, Amir Zeb: Conceived and designed the experiments; Performed the experiments; Analyzed and interpreted the data; Wrote the paper.

Min Sung Kim, Hyeon-Nae Jeon, Jianmin Wang, Shenghui Guan, Kyoung Tai NO: Contributed reagents, materials, analysis tools or data.

Funding statement

This work was supported by Yonsei University graduate school "Integrative Biotechnology & Translational Medicine".

Data availability statement

Data included in article/supplementary material/referenced in article.

Declaration of interests statement

The authors declare no conflict of interest.

Additional information

Supplementary content related to this article has been published online at https://doi.org/10.1016/j.heliyon.2022.e10011.

References

[1] S. Pyo, et al., Super-expansion of assembled reduced graphene oxide interlayers by segregation of Al nanoparticle pillars for high-capacity Na-ion battery anodes, ACS Appl. Mater. Interfaces 12 (21) (2020) 23781–23788.

[2] A.E. Baumann, et al., Metal-organic framework functionalization and design strategies for advanced electrochemical energy storage devices, Communications Chemistry 2 (2019).

[3] K. Xu, Nonaqueous liquid electrolytes for lithium-based rechargeable batteries, Chem. Rev. 104 (10) (2004) 4303–4417.

[4] T. Husch, M. Korth, Charting the known chemical space for non-aqueous lithium-air battery electrolyte solvents, Phys. Chem. Chem. Phys. 17 (35) (2015) 22596–22603.

[5] J.L. Gao, Hybrid quantum and molecular mechanical simulations: an alternative avenue to solvent effects in organic chemistry, Acc. Chem. Res. 29 (6) (1996) 298–305.

[6] C. Reichardt, Empirical parameters of solvent polarity as linear free-energy relationships, Angew Chem. Int. Ed. Engl. 18 (2) (1979) 98–110.

[7] J.R. Pliego, J.M. Riveros, The cluster-continuum model for the calculation of the solvation free energy of ionic species, J. Phys. Chem. A 105 (30) (2001) 7241–7247.

[8] E.L.M. Miguel, et al., How accurate is the SMD model for predicting free energy barriers for nucleophilic substitution reactions in polar protic and dipolar aprotic solvents? J. Braz. Chem. Soc. 27 (11) (2016) 2055–2061.

[9] M. Maroncelli, J. Macinnis, G.R. Fleming, Polar-solvent dynamics and electron-transfer reactions, Science 243 (4899) (1989) 1674–1681.

[10] R.C. Schweitzer, J.B. Morris, The development of a quantitative structure property relationship (QSPR) for the prediction of dielectric constants using neural networks, Anal. Chim. Acta 384 (3) (1999) 285–303.

[11] P.M. Wang, A. Anderko, Computation of dielectric constants of solvent mixtures and electrolyte solutions, Fluid Phase Equil. 186 (1-2) (2001) 103–122.

[12] M. Cocchi, et al., Development of quantitative structure-property relationships using calculated descriptors for the prediction of the physicochemical properties (n(D), p, bp, epsilon, eta) of a series of organic solvents, J. Chem. Inf. Comput. Sci. 39 (6) (1999) 1190–1203.

[13] S. Sild, M. Karelson, A general QSPR treatment for dielectric constants of organic compounds, J. Chem. Inf. Comput. Sci. 42 (2) (2002) 360–367.

[14] A.R. Katritzky, et al., Quantitative measures of solvent polarity, Chem. Rev. 104 (1) (2004) 175–198.

[15] R. Bouteloup, D. Mathieu, Predicting dielectric constants of pure liquids: fragment-based Kirkwood-Frohlich model applicable over a wide range of polarity, Phys. Chem. Chem. Phys. 21 (21) (2019) 11043–11057.

[16] N. Deb, A.S. Tiwary, A.K. Mukherjee, Calculation of the Kirkwood-Frohlich correlation factor and dielectric constant of methanol using a statistical model and density functional theory, Mol. Phys. 108 (14) (2010) 1907–1917.

[17] H. Fröhlich, General theory of the static dielectric constant, Trans. Faraday Soc. 44 (1948) 238–243.

[18] J.P. Liu, et al., A quantitative structure property relation correlation of the dielectric constant for organic chemicals, J. Chem. Eng. Data 55 (1) (2010) 41–45.

[19] Papazian, A. Harold, Correlation of surface tension between various liquids, J. Am. Chem. Soc. 93 (22) (1971) 5634–5636.

[20] Holmes, F. Curtis, Relation between surface tension and dielectric constant, J. Am. Chem. Soc. 95 (4) (1973) 1014–1016.

[21] A.N. Paruta, B.J. Sciarrone, Lording, Correlation between solubility parameters and dielectric constants, J. Pharmaceut. Sci. 51 (7) (1962) 704–705.

[22] H.F. Arnoldus, Surface contribution to the electric dipole moment near an interface, and its effect on power emission, Journal of the Optical Society of America a-Optics Image Science and Vision 38 (5) (2021) 606–615.

[23] J.J.P. Stewart, Special issue - mopac - a semiempirical molecular-orbital program, J. Comput. Aided Mol. Des. 4 (1) (1990) 1–45.

[24] G. Landrum, RDKit: Open-Source Cheminformatics from Machine Learning to Chemical Registration, Abstracts of Papers of the American Chemical Society, 2019, p. 258.

[25] H. Moriwaki, et al., Mordred: a molecular descriptor calculator, J. Cheminf. 10 (2018).

[26] M. Reutlinger, et al., Chemically advanced template search (CATS) for scaffold-hopping and prospective target prediction for "orphan' molecules, Molecular Informatics 32 (2) (2013) 133–138.

[27] R.X. Wang, Y. Fu, L.H. Lai, A new atom-additive method for calculating partition coefficients, J. Chem. Inf. Comput. Sci. 37 (3) (1997) 615–621.

[28] R. Ahmedi, T. Lanez, Calculation of octanol/water partition coefficients of ferrocene derivatives, Asian J. Chem. 22 (1) (2010) 299–306.

[29] V. Sharma, R. Goswami, A.K. Madan, Eccentric connectivity index: a novel highly discriminating topological descriptor for structure-property and structure-activity studies, J. Chem. Inf. Comput. Sci. 37 (2) (1997) 273–282.

[30] D.T. Stanton, et al., Charged partial surface area (CPSA) descriptors QSAR applications, SAR QSAR Environ. Res. 13 (2) (2002) 341–351.

[31] C.W. Yap, PaDEL-Descriptor: an open source software to calculate molecular descriptors and fingerprints, J. Comput. Chem. 32 (7) (2011) 1466–1474.

[32] D.S. Cao, et al., ChemoPy: freely available python package for computational biology and chemoinformatics, Bioinformatics 29 (8) (2013) 1092–1094.

[33] D.S. Cao, et al., PyDPI: freely available Python package for chemoinformatics, bioinformatics, and chemogenomics studies, J. Chem. Inf. Model. 53 (11) (2013) 3086–3096.

[34] D.S. Cao, et al., Rcpi: R/Bioconductor package to generate various descriptors of proteins, compounds and their interactions, Bioinformatics 31 (2) (2015) 279–281.

[35] N.M. O'Boyle, G.R. Hutchison, Cinfony – combining Open Source cheminformatics toolkits behind a common interface, Chem. Cent. J. 2 (1) (2008) 24.

[36] J.S. Mao, et al., Comprehensive strategies of machine-learning-based quantitative structure-activity relationship models, iScience 24 (9) (2021).