Abstract

Pointwise mutual information (PMI) is a measure of association used in information theory. In this paper, PMI is used to characterize several publicly available databases (DrugBank, ChEMBL, PubChem and ZINC) in terms of association strength between compound structural features resulting in database PMI interrelation profiles. As structural features, substructure fragments obtained by coding individual compounds as MACCS, PubChemKey and ECFP fingerprints are used. The analysis of publicly available databases reveals, in accord with other studies, unusual properties of DrugBank compounds which further confirms the validity of PMI profiling approach. Z-standardized relative feature tightness (ZRFT), a PMI-derived measure that quantifies how well the given compound’s feature combinations fit these in a particular compound set, is applied for the analysis of compound synthetic accessibility (SA), as well as for the classification of compounds as easy (ES) and hard (HS) to synthesize. ZRFT value distributions are compared with these of SYBA and SAScore. The analysis of ZRFT values of structurally complex compounds in the SAVI database reveals oligopeptide structures that are mispredicted by SAScore as HS, while correctly predicted by ZRFT and SYBA as ES. Compared to SAScore, SYBA and random forest, ZRFT predictions are less accurate, though by a narrow margin (Acc\_ZRFT = 94.5%, Acc\_SYBA = 98.8%, Acc\_SAScore = 99.0%, Acc\_RF = 97.3%). However, ZRFT ability to distinguish between ES and HS compounds is surprisingly high considering that while SYBA, SAScore and random forest are dedicated SA models, ZRFT is a generic measurement that merely quantifies the strength of interrelations between structural feature pairs. The results presented in the current work indicate that structural feature co-occurrence, quantified by PMI or ZRFT, contains a significant amount of information relevant to physico-chemical properties of organic compounds.

Keywords: Hashed fingerprint, Structural key, Information theory, Pointwise mutual information, Synthetic accessibility

Introduction

Information theory is a mathematical approach for the quantification, storage and communication of information. Information theory concepts, such as Shannon entropy [1] or mutual information (MI) [2], are used across a wide variety of scientific areas. Due to the generic nature of information theory, sometimes even very distant scientific fields independently develop methodologies that are built upon the same underlying information theory framework. In one such framework, MI is used to profile and compare objects based on the interrelations between their features. MI is commonly used in linguistics to identify unusual word combinations [3] with the aim to estimate text complexity [4]. In bioinformatics, gene coinheritance among different organisms, expressed by MI, was profiled to elucidate functional
linkages among proteins [5]. In medicinal sciences, MI was applied to profile relations between stressors, health conditions, genes and other factors in order to build comorbidity charts useful for disease study and preventive medicine [6–8].

In cheminformatics, the use of information theory concepts is widespread [9, 10]. Shannon entropy was applied, for example, to design and evaluate molecular descriptors [11, 12] and fingerprints [13], to determine the information content of chemical structures based on their topology and symmetry [14], to create the aggregate fingerprints of whole chemical databases [15] or to evaluate the significance of individual fingerprint bits in order to improve similarity search methodologies [16]. MI was applied to improve feature selection in similarity search [17] and QSAR [18, 19] and to improve performance of topological molecular descriptors in the modeling of the physico-chemical properties of 2-furylethylene derivatives [20]. However, a more straightforward MI application, the comparison of compound sets based on interrelations between their structural features, was not reported so far. In this paper, we demonstrate the use of pointwise mutual information (PMI) for the profiling of structural feature interrelations within several publicly available chemical databases (DrugBank [21], ChEMBL [22, 23], PubChem [24] and ZINC15 [25]) using PubChem [26] and MDL MACCS [27] structure keys, as well as extended connectivity fingerprints (ECFP) [28]. Z-standardized relative feature tightness (ZRFT), a PMI-based measure that quantifies how the given compound fits into the particular compound set, is postulated and its utility is demonstrated in the analysis of compound synthetic accessibility (SA), as well as in the classification of compounds as easy (ES) and hard (HS) to synthesize.

Methods

Methodology of feature interrelation profiling

In linguistics, PMI is used to expresses the extent to which the observed frequency of the co-occurrence of two different words differs from what would be expected if they were independent [29]. PMI is the measure of the strength of the association between words x and y and, for a given corpus, it is calculated using the number of times the word pair (x, y) is observed in one sentence versus the number of times words x and y are observed separately. The concept of PMI can be easily adopted for the analysis of the interrelations between structural features (i.e., words) within individual molecules (i.e., sentences) from a compound set (i.e., a corpus). In this work, two types of structural features are employed: dictionary-based and hashed structural fragments [30–32]. Dictionary-based fragments are used to convert a compound into a binary fingerprint called “a structure key”. Though fragment dictionaries are constructed from fragments perceived as most relevant to the intended purpose, some important fragments may be omitted. To circumvent this aspect of explicit fragment selection, hashed fingerprints were developed. They are formed by fitting all fragments present in the molecule up to a defined size into the bit-string of the defined length. In the present work, PubChem [26] and MDL MACCS [27] structure keys and ECFP4 and ECFP6 [28] hashed fingerprints are used to decompose molecules into structural features. Structural features/fragments will be, in the following text, referred to simply as features.

Profiling feature interrelations requires to retain information on how many times each feature pair appears in the compound set S. This information is stored in the co-occurrence relation matrix (CORM). If each molecule in the compound set S is encoded by the feature vector k, CORM is calculated as the sum of the outer products of all feature vectors k:

\[ CORM(S) = \sum_{o=1}^{\vert S \vert} k_o \otimes k_o = \sum_{o=1}^{\vert S \vert} k_o k_o^T \]  

where \( \vert S \vert \) is the number of molecules in the compound set S. COPRM is a symmetrical square matrix of nonnegative integers with dimensions equaling to the number of features, i.e. to the length of the feature vector k.

The division of co-occurrence counts in CORM by compound set size \( \vert S \vert \) leads to the co-occurrence probability relation matrix (COPRM):

\[ COPRM(S) = \frac{CORM(S)}{\vert S \vert} \]  

On its diagonal, COPRM contains probabilities with which individual features are observed in the compound set S. Its off-diagonal elements contain probabilities of the occurrence of feature pairs in the compound set S.

The strength of the interrelation between two features x and y can be inferred using pointwise mutual information (PMI):

\[ PMI = \log_2 \frac{p(x,y)}{p(x)p(y)} \]

PMI quantifies the divergence between feature pair co-occurrence probability \( p(x, y) \) and individual occurrence probabilities \( p(x) \) and \( p(y) \). Positive PMI indicates the enrichment of feature co-occurrences compared to their separate occurrences, e.g., PMI of 1 means that both features appear together (i.e., in one compound) twice as often as they appear separately (i.e., in two different compounds). PMI equalling to 0 means that two features appear together about as often as they appear...
separately. Negative PMI indicates negative inter-relation between a pair of features, e.g., a feature pair with PMI of -1 appears only half as often as could be expected from their individual occurrence probabilities.

From COPRM, a pointwise mutual information relation matrix (PMIRM) containing PMI values for all possible feature pairs can be constructed. Its individual elements $PMIRM(S)_{ij}$ are given as:

$$PMIRM(S)_{ij} = \log_2 \frac{COPRM(S)_{ij}}{COPRM(S)_{i,}} \frac{COPRM(S)_{j,}}{COPRM(S)_{j,}}$$  (4)

PMIRM diagonal contains zeros and feature pairs involving features that are never observed in the compound set $S$ have undefined PMI. PMIRM constitutes the interrelation profile of the compound set $S$. PMIRM interrelation profile is intrinsically affected by the choice of features. For example, overlapping structural features can interact in a complementary manner which leads to the shift of PMI distribution towards positive values. These shifts can be, if desired, corrected by normalizing PMI values into Z-scores (ZPMI) leading to the Z-standardized pointwise mutual information relation matrix (ZPMIRM):

$$ZPMIRM(S)_{ij} = \frac{PMIRM(S)_{ij} - \mu(PMIRM(S))}{\sigma(PMIRM(S))}$$  (5)

where $\mu$ is the mean and $\sigma$ is the standard deviation of all values in PMIRM. The construction of relation matrices (RMs) CORM, COPRM, PMIRM and ZPMIRM is summarized in Fig. 1.

Apart from the analysis of interrelations within the compound set $S$, PMI methodology also enables to measure how tightly the query compound set $S'$ matches the reference compound set $S'$ meaning how similar are, on average, the query and reference compound sets in terms of feature pair co-occurrence probabilities. This is quantified by the relative feature tightness (RFT):

$$RFT = \mu \left( COPRM(S) \times PMIRM(S') \right) = \mu \left( \sum_{o=1}^{\left| S \right|} k_o k_o^T \frac{PMIRM(S')}{|S|} \right) \times PMIRM(S')$$  (6)
where COPRM(\(S\)) is the co-occurrence probability relation matrix (Eq. 2) of the query compound set \(S\), PMIRM(\(S'\)) is the pointwise mutual information relation matrix (Eq. 4) of the reference compound \(S'\) and \(\mu\) is the mean of all values in the \(COPRM(S) \times PMIRM(S')\) matrix. Based on the choice of \(S\) and \(S'\), three different cases can occur:

1. The query compound set \(S\) consists of only one compound, the reference compound set \(S'\) consists of several compounds. In this case, RFT measures how well the feature combinations of a compound \(S\) fit these within the reference compound set \(S'\).
2. Both \(S\) and \(S'\) compound sets consist of several compounds. In this case, RFT measures how close are feature interrelations within compounds from the query compound set \(S\) to feature interrelations within the reference compound set \(S'\).
3. The reference compound set \(S'\) is the same as the query compound set \(S\), i.e., \(S = S'\). In this case, RFT measures the “inner tightness” of the compound set \(S\), i.e., how strong are the feature interrelations within the compound set \(S\).

Generally, the higher RFT is, the more similar are the compound sets \(S\) and \(S'\) in terms of feature co-occurrences. If ZPMIRM is used instead of PMIRM in Eq. 6, a Z-standardized relative feature tightness (ZRFT) is obtained:

\[
ZRFT = \mu \left( COPRM(S) \times ZPMIRM(S') \right) = \mu \left( \frac{\sum_{i,j} |S| k_0 k_1^T}{|S|} \times ZPMIRM(S') \right)
\]  

ZRFT is interpreted much like RFT with the added convenience of standardization: chemical structures containing predominantly feature pairs that are rated above average within the reference interrelation profile will receive positive ZRFT values and vice versa. However, it must be stressed that neither RFT nor ZRFT can be considered as metrics because they are not symmetric: \(RFT/ZRFT(A, B)\) is unlikely to be the same as \(RFT/ZRFT(B, A)\).

**Applications of feature interrelation profiling**

The utility of feature interrelation profiling is demonstrated for chemical database and synthetic accessibility analysis.

**Chemical database analysis**

In this application, the DrugBank 5.0.3 [21], ChEMBL22 [22, 23], PubChem (downloaded in 12/2016) [24] and ZINC15 [25] databases (Fig. 2) are analyzed using their PMI profiles. The merged_dbs compound set is created by merging all four databases with duplicates removed. Feature interrelations are profiled using the RDKit [33] cheminformatics toolkit and the ChemFP Python library [34, 35]. Compound stereochemistry is removed, compounds are standardized by the IMI eTox standardizer [36] and duplicates are identified using InChIKeys. For each compound, four fingerprints are generated: the PubChemKey (881 bits long) [26], MACCS key (166 bits long) [27] and ECFP4 and ECFP6 fingerprints, both 1024 bits long [28]. To estimate the influence of compound set size on PMI profile, a series of five overlapping ZINC subsets containing 8000, 32,000, 128,000, 512,000 and 2,048,000 randomly selected compounds is prepared.

**Synthetic accessibility analysis**

In this application, ZRFT profiles of several compound sets (Table 1, Fig. 2) with easy (ES) and hard (HS) to synthesize molecules are investigated under the premise that compounds containing feature pairs common in existing molecules are likely to be synthetically accessible.

HS compound set (Additional file 1) is generated by the Nonpher methodology [45]. Nonpher is based on the molecular morphing algorithm [46] in which new structures are constructed by the iterative application of simple structural changes, such as the addition or removal of an atom or a bond. In Nonpher, molecular morphing is stopped when the proposed structure exceeds the thresh-

old [45] of at least one of four monitored complexity metrics (Bertz [41], Whitlock [42], BC [43] and SMCM [44] indices). This procedure was previously optimized [45] to ensure that though generated molecules can be deemed as HS, they are not excessively complex. Nonpher algorithm and compound set construction are described in a detail in the Nonpher and SYBA publications [37, 45].

Three ES compound data sets (Additional file 1) are obtained from the following sources: the Synthetically Accessible Virtual Inventory (SAVI) Database [38, 39], Screenable Chemical Universe Based on Intuitive Data OrganizatiOn (SCUBIDOO) database [40] and ZINC15 database [25]. While the SAVI and SCUBIDOO databases were computationally generated by the application of selected chemical reactions (11 reactions for SAVI and 58 reactions for SCUBIDOO generation) to the given set of chemical building blocks (~ 230,000 building blocks for SAVI and ~ 8000 building blocks for SCUBIDOO generation), the ZINC15 database contains already synthesized
commercially available organic compounds. Therefore, compounds in SAVI, SCUBIDOO and ZINC15 databases can be considered as ES. The examples of the nonpher, savi, scubidoo and random_zinc compounds are shown in Additional file 2.

Though savi and scubidoo compound sets are expected to contain only ES compounds, some of these are
Table 1 Compound sets used in synthetic accessibility assessment

| Compound set       | Type | Number of compounds |
|--------------------|------|---------------------|
| nonpher            | HS   | 693,353             |
| savi               | ES   | 610,245             |
| scubidoo           | ES   | 999,794             |
| zinc_random        | ES   | 693,353             |
| nonpher_complex    | HS   | 161                 |
| savi_complex       | HS   | 2930                |
| scubidoo_complex   | HS   | 104                 |
| gdb_complex        | HS   | 3581                |

Table 2 The number of all and unique standardized compounds

| Compound set       | All compounds | Unique compounds |
|--------------------|---------------|------------------|
| DrugBank           | 6768          | 6496             |
| ChEMBL             | 1,666,863     | 1,512,302        |
| PubChem            | 91,221,617    | 69,081,967       |
| ZINC               | 285,732,863   | 157,914,301      |
| merged_dbs         | 378,628,111   | 213,777,358      |

Compounds are standardized using IMI eTox standardizer [36] and duplicates are identified using InChIKey calculated after compound standardization.

Results and discussion

Chemical database analysis

The number of all and unique standardized compounds in the DrugBank, ChEMBL, PubChem, ZINC and merged_dbs compound sets is shown in Table 2 and the overlaps between individual compound sets in Table 3.

PMI profiles of increasingly larger randomly selected ZINC subsets are shown in Fig. 3.
The presence of a higher amount of negative interrelations in ZINC ECFP profile (Fig. 4c, d) means that ZINC contains less co-occurring structural fragments than any other database. This indicates that, in terms of feature interrelations, ZINC contains the most diverse set of compounds. On the other hand, considering that the average database Tanimoto coefficient $T_C$ is calculated from 12,497,500 pairwise comparisons generated exhaustively from 5000 compounds, ZINC $T_C$ value of 0.14, which is the highest of all databases (Table 4), means that ZINC structures share 14% of ECFP features on average. ZINC can, thus, be considered as the least structurally diverse database. Seemingly contradictory conclusions regarding ZINC diversity are only the manifestation of the fact, that both measures capture different compound properties and reflect, thus, different views of reality. Tanimoto similarity quantifies how are individual features shared between compounds compared to all features present in a compound set $S$. On the other hand, PMI quantifies (Eq. 3) how often features $x$ and $y$ occur together in the same compound (given by the feature pair co-occurrence probability $p(x,y)$) compared to the chance that they appear in the same compound if they are independent (given as $p(x)p(y)$). So, if $x$ and $y$ are present in all compounds in $S$, they positively contribute to pairwise Tanimoto coefficients.
between structures in \( S \). However, their PMI will be zero because \( p(x,y) = 1 \), \( p(x) = 1 \), \( p(y) = 1 \) and \( \text{PMI} = \log_2 \frac{p(x,y)}{p(x)p(y)} = \log_2 1 = 0 \). This means that a compound set can have a high average Tanimoto similarity between the structures and, at the same time, low PMI values. In the case of ZINC compounds, while a high pairwise Tanimoto similarity indicates that they have, out of all studied compound sets, most fragments in common, their low PMI values mean that these fragments are less mutually interrelated.

Fig. 3 The dependence of PMI profile on compound set size. 5 randomly selected ZINC subsets that contain 8000, 32,000, 128,000, 512,000 and 2,048,000 compounds are profiled using MACCS, PubChemKey, ECFP4 and ECFP6 fingerprints.
Synthetic accessibility analysis

ZRFT, SAScore and SYBA distributions, of the nonpher, savi, scubideo and zinc_random compound sets are shown in Fig. 6.

While ZRFT profiles (Fig. 6a) and SYBA distributions (Fig. 6c) are smooth, SAScore distributions (Fig. 6b) shows more complex shapes that are likely the result of heuristic complexity penalty used in SAScore calculation.

Fig. 4 PMI profiles of the DrugBank, PubChem, ChEMBL and ZINC databases using MACCS, PubChemKey, ECFP4 and ECFP6 fingerprints. Dashed rectangles in MACCS and PubChemKey profiles highlight the regions where DrugBank significantly differs from other databases. In this region, 1674 interrelations not present in other databases were identified.
ZRFT profiles (Fig. 6a) show a clear separation between ZINC (i.e., ES) and Nonpher (i.e., HS) [45] compounds. ZRFT values of the computationally generated ES compounds sets, i.e. SAVI and Scubidoo, fall between those of Nonpher and ZINC, closer to ZINC. The same trends can be observed in SYBA and SAScore distributions, albeit SAScore distributions show less distinction between ZINC and SAVI compounds.

Fig. 5  Examples of MACCS and PubChemKey fragments. Generic and specific MACCS fragments (a), MACCS (b) and PubChemKey (c) fragments exclusive for DrugBank.
ZRFT profiles and SYBA and SAScore distributions of the nonpher_com, savi_com, scubidoo_com and gdb_com compound sets are shown in Fig. 7.

SYBA, SAScore and ZRFT distributions of the scubidoo_com compound set are shifted toward positive values and contain more values associated with synthetically accessible structures than any other complex compound set. Strong scubidoo_com peaks at ZRFT ~ 0.25 (Fig. 7a), SAScore ~ 3.7 (Fig. 7b) and SYBA ~ 10 (Fig. 7c) are composed mostly by the same 66 structures with five or six membered heterocycles. savi_com compounds are rated differently by all three methods with their SAScore and SYBA distributions being particularly irregular and widespread. Based on their high ZRFT (> 0.3) and SYBA (> 180) values (Fig. 7a, c), 499 SAVI complex compounds should be considered as ES. However, their SAScore higher than 8.5 (Fig. 7b) would rate these compounds as HS. A closer inspection reveals that all these compounds are oligopeptides (Fig. 8) and can be, therefore, synthesized by connecting individual amino acid residues [61]. Because SAScore is designed [48] to assess the SA of drug-like [62] compounds, oligopeptides lie outside its applicability domain. Their structural complexity, incorporated into SAScore using the heuristic complexityScore [48], outweighs individual fragment contributions and contributes unfavorably to their SAScore values. In contrast, both SYBA and ZRFT predict these compounds correctly as ES. Oligopeptides include a large number of fragments that are highly scored because they appear more often in ES than in HS compounds, which is reflected in their high SYBA values. Also, oligopeptides contain ECFP feature pair combinations that fit well within the ZRFT profile of known SA compounds in merged_dbs.

The smallest overlap between ZINC and Nonpher compounds and, therefore, the best differentiation between ES and HS compounds was achieved by the SYBA model, followed by SAScore and ZRFT (Fig. 9). ZRFT is strongly correlated (Fig. 9) both with SYBA (r = 0.82) and SAScore (r = -0.83) which demonstrates that ZRFT contains a significant amount of information about compound SA.

In addition, the separation between ES and HS compounds in ZRFT density plots (Fig. 9) suggests that ZRFT can be used as a classifier. The comparison between the RF, SYBA, SAScore and ZRFT classification of the TMC and TCP tests sets is given in Tables 5 and 6, respectively.

Though ZRFT classification is inferior to SYBA, SAScore and RF, its ability to distinguish, using the Youden index optimized threshold of 0.2, between ES and HS compounds is surprisingly high considering that ZRFT is a generic measurement based only on interrelations between structural feature pairs compared to the reference compound set, while SYBA and SAScore are dedicated models trained to estimate compound SA.

### Conclusions

The methodology of pointwise mutual information (PMI) profiling is introduced and its utility is demonstrated for the analysis of structural feature interrelations in publicly available chemical databases and for the analysis and prediction of synthetic accessibility of organic compounds. Interrelation profiles are constructed both from dictionary-based (MACCS and PubChemKey) and hashed circular fragments (ECFP). PMI interrelation profiles of the PubChem, ZINC, ChEMBL and DrugBank databases indicate the presence of both positive and negative feature interrelations. ECFP structural fragments are more suitable for fragment co-occurrence profiling than dictionary-based fragments as they provide more regular interrelation profiles. Unusual favorable fragment combinations of DrugBank compounds manifest themselves by the shift of DrugBank PMI profile to positive values meaning that DrugBank compounds have stronger positive feature interrelations than any other chemical database. Z-standardized relative feature tightness (ZRFT), a PMI-derived measure that quantifies how tightly the query compound set matches the reference compound set, is used to characterize five compound sets with varying degree of synthetic accessibility. Synthetically accessible compounds possess a higher amount of fragment pairs occurring in known molecules. ZRFT profiles are compared with the distributions of SYBA [37] and SAScore [48], two dedicated models for the estimation of synthetic accessibility. In addition, ZRFT is also applied to the classification of compounds as easy (ES) or hard (HS) to synthesize and compared to the results of the random forest (RF), SYBA and SAScore. Though ZRFT classification is inferior to SYBA, SAScore and RF, its ability to distinguish between ES and HS compounds is surprisingly high. Therefore, we may conclude that compound synthetic accessibility is given, to a large extent, by structural feature combinations that...
can be quantified by ZRFT. However, we would like to stress that ZRFT is not a dedicated measure of synthetic accessibility. Instead, ZRFT is a generic method that only detects interrelations between structural feature pairs and quantifies their match to interrelations in the reference compound set. ZRFT interpretation depends on the context. For example, comparing a compound with the interrelation profile of synthetically accessible compounds will be interpreted differently than comparing it with the interrelation profile of natural products.
For the comparison of chemical databases, PMI inter-relation profiles (Eq. 4) are favored over ZPMI profiles (Eq. 5) because Z-score standardization removes information about the absolute PMI values which is usually undesirable for this application. On the other hand, ZRFT is more suitable for the analysis and prediction of compound properties such as synthetic accessibility. While RFT (Eq. 6) captures the strength of...
interrelations in absolute numbers that can vary widely depending on the reference interrelation profile, ZRFT (Eq. 7) quantifies how well the observed feature pairings match the reference interrelation profile in the units of standard deviation, leading to better interpretability.

The results presented in the current work indicate that structural feature co-occurrence, quantified by PMI or ZRFT profiles, contains a significant amount of information relevant to physico-chemical properties of organic compounds. It must be stressed that neither PMI nor ZRFT are models. PMI is simply the representation of interrelations between feature pairs within a compound set and ZRFT is the measure of a similarity, in terms of feature co-occurrences, between two compound sets (though ZRFT is not a metric as it is not symmetric). This is akin to structural fingerprints, where a fingerprint is the representation of structural features within a compound and the Tanimoto coefficient is the measure of similarity between two fingerprints. The possible use cases of interrelation profiles will be, due to these conceptual parallels, similar to these of binary fingerprints. Consequently, feature interrelation profiles can be potentially used to introduce additional information rich layer to established fingerprint-based methodologies. However, the construction of meaningful interrelation profiles is computationally intensive, which we perceive as one of the biggest limitations of feature interrelation profiling.

The study of the influence of the number of compounds on the interrelation profile (Fig. 3) indicates that the number of compounds necessary to yield a meaningful interrelation profile is in the order of $10^5$–$10^6$ for ECFP feature vectors. Finally, the interrelation profile is defined by the choice of a feature vector. For an intended use, it may not be straightforward to choose the appropriate feature vector and it may be required to construct a multitude of different interrelation profiles for different feature vectors.

In the future, we plan to further experiment with different types of feature vectors and to adapt the methodology of sparse vectors and matrices in order to decrease computational demands. Furthermore, we will design feature vectors with structural features corresponding to specific functional groups, pharmacophore features etc. with the aim to improve the interpretability of the resulting interrelation profiles. Later, we will also investigate the utility of hybrid feature vectors containing interrelation profiles concatenated with, for example, QAFFP biological fingerprints [63, 64] or with other features of interest. We plan to use interrelation profiling in various cheminformatics applications, such as in biological activity classification or potency prediction, focused chemical library construction, diversity data selection or ensemble modeling using RFT together with domain-specific models for, e.g., natural product likeness assessment [65–67]. Given that
Fig. 9  Correlation between ZRFT, SYBA and SAScore. On the diagonal, distributions of individual SA scores for the nonpher (i.e., HS) and zinc_random (i.e., ES) compound sets are plotted. Above the diagonal, correlations between all SA score pairs are shown. Below the diagonal, pairwise kernel density estimations between all SA score pairs are depicted. Distributions were calculated for 10 000 randomly selected compounds from Nonpher and ZINC databases.
Table 5 The performance of classification models for the manually curated TMC test set

| Model  | AUC   | Acc  | SN   | SP   |
|--------|-------|------|------|------|
| SYBA   | 0.903 | 0.871| 0.902| 0.840|
| SAScore| 0.865 | 0.859| 0.799| 0.919|
| ZRFT   | 0.871 | 0.831| 0.827| 0.836|
| RF     | 0.875 | 0.842| 0.855| 0.828|

Quality measures AUC, Acc, SN and SP are reported as their average values over 30 TMC instances. Results for SYBA, SAScore and RF classification are taken from SYBA publication [37].

Table 6 The performance of classification models for the computationally picked TCP test set

| Model  | AUC   | Acc  | SN   | SP   |
|--------|-------|------|------|------|
| SYBA   | 0.998 | 0.988| 0.985| 0.991|
| SAScore| 0.999 | 0.990| 0.986| 0.993|
| ZRFT   | 0.987 | 0.945| 0.933| 0.956|
| RF     | 0.995 | 0.973| 0.960| 0.986|

Results for SYBA, SAScore and RF classification are taken from SYBA publication [37].

interrelation profiles are matrices of numeric values, they can also be used to train machine learning models and to identify and leverage specific feature interrelations that provide most information about the estimated property.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13321-020-00483-y.

Acknowledgements

Computational resources were supplied by the project “e-Infrastruktura CZ” (e-INFRA LM2018140) provided within the program Projects of Large Research, Development and Innovations Infrastructures.

Authors’ contributions

IC conceptualized the problem and developed, implemented and validated the methodology. IC also maintains Feature Interrelation Profiling library (fip) GitHub repository of and its Conda packages. MV provided SA related data sets and contributed to methodology development. DS supervised the study and prepared the manuscript with the active participation of IC and MV. All authors read and approved the final manuscript.

Funding

This work was supported from the Ministry of Education of the Czech Republic (RVO 68378050-KAV-NPUI and LM2018130).

Availability of data and materials

fip, Python library for interrelation feature profiling is available at https://github.com/lich-uct/fip. fip GitHub repository contains Python code, tutorial in the form of Jupyter notebook and pre-computed CORM matrices of ZINC, PubChem, ChEMBL, DrugBank and merged compound sets. fip is also available as Conda package at https://anaconda.org/LICH/fip.

Competing interests

The authors declare that they have no competing interests.

Author details

1 CZ-OPENSSCREEN National Infrastructure for Chemical Biology, Department of Informatics and Chemistry, Faculty of Chemical Technology, University of Chemistry and Technology Prague, Technická 5, 166 28 Prague, Czech Republic 2 CZ-OPENSSCREEN National Infrastructure for Chemical Biology, Institute of Molecular Genetics of the ASCR v. v. i., Vídeňská 1083, 142 20 Prague 4, Czech Republic.

References

1. Shannon CE (1948) A mathematical theory of communication. Bell Syst Tech J 27(4):623–656
2. Quinlan JR (1986) Induction of decision trees. Mach Learn 1:26
3. Everet S (2005) The statistics of word cooccurrences: word pairs and colocations. Universität Stuttgart, Universität Stuttgart
4. Everet S (2006) The co-occurrence network: a tool for understanding complex datasets. Universität Stuttgart, Universität Stuttgart
5. Godden JW, Bajorath J (2000) Shannon entropy—a novel concept in molecular descriptor and diversity analysis. J Mol Graph Model 18(1):73–76
10. Vogt M, Wassermann AM, Bajorath J (2010) Application of information-theoretic concepts in chemoinformatics. Information 1(2):14
11. Godden JW, Stahura FL, Bajorath J (2000) Variability of molecular descriptors in compound databases revealed by Shannon entropy calculations. J Chem Inf Comput Sci 40(3):796–800
12. Gregori-Puigjane E, Mestre J (2006) SHED: Shannon entropy descriptors from topological feature distributions. J Chem Inf Model 46(4):1615–1622
13. Xue L, Godden JW, Stahura FL, Bajorath J (2003) Design and evaluation of a molecular fingerprint involving the transformation of property descriptor values into a binary classification scheme. J Chem Inf Comput Sci 43(6):1151–1157
14. Bonchev D, Kamenski D, Kamenska V (1976) Symmetry and information-content of chemical structures. B Math Biol 38(2):119–133
15. Fernandez-de Gortari E, Garcia-Jacas CR, Martinez-Mayorga K, Medina-Franco JL (2017) Database fingerprint (DFP): an approach to represent molecular databases. J Cheminform 9:1–9
16. Wang Y, Geppert H, Bajorath J (2009) Shannon entropy-based fingerprint similarity search strategy. J Chem Inf Model 49(7):1687–1691
17. Bender A, Mussa HY, Glen RC, Reiling S (2004) Evaluation of mutual information and genetic programming for feature selection in QSAR. J Chem Inf Comput Sci 44(5):1686–1692
18. Martinez MJ, Ponsoni I, Diaz MF, Vazquez GE, Soto AJ (2015) Similarity searching of chemical databases using atom environment descriptors (MOLPRINT 2D): Evaluation of performance. J Chem Inf Model 44(5):1708–1718
19. Venkatraman V, Dalby AR, Yang ZR (2004) Evaluation of mutual information and genetic programming for feature selection in QSAR. J Chem Inf Comput Sci 44(5):1686–1692
20. Martinez MJ, Ponsoni I, Diaz MF, Vazquez GE, Soto AJ (2015) Visual analytics in cheminformatics: user-supervised descriptor selection for QSAR methods. J Cheminform 7:39
21. Barigye SJ, Marrero-Ponce Y, Martinez-Lopez Y, Torres F, Arilles-Martinez LM, Pino-Urías RW, Martinez-Santiago O (2013) Relations frequency hypermatrices in mutual, conditional and joint entropy-based information indices. J Comput Chem 34(9):259–274
22. Gaulton A, Hensey A, Nowotka M, Bento AP, Chambers J, Mendez D, Mutowoo P, Atkinson F, Bellsi L, Cibrarian-Halate E et al (2017) The ChEMBL database in 2017. Nucleic Acids Res 45(D1):D945–D954
23. Mendez D, Gaulton A, Bento AP, Chambers J, De Veij M, Felix E, Magarinos MP, Mosquera JF, Mutowoo P, Nowotka M et al (2019) ChEMBL: towards direct deposition of bioassembly data. Nucleic Acids Res 47(D1):D930–D940
24. Kim S, Chen J, Cheng T, Gindulyte A, He J, He S, Li Q, Shoemaker BA, Thiessen PA, Yu B et al (2019) PubChem 2019 update: improved access to chemical data. Nucleic Acids Res 47(D1):D1102–D1109
25. Sterling T, Irvin JJ (2015) ZINC 15—Ligand discovery for everyone. J Cheminform 55(11):2324–2337
26. PubChem/CACVT5/structure keys. ftp://ftp.ncbi.nlm.nih.gov/pubchem/specifications/pubchem_fingerprints.txt. Accessed 21 Feb 2020.
27. Durant JL, Leland BA, Henry DR, Nourse JG (2002) Reoptimization of MDL DES. 27(8):675–679
28. Kim S, Chen J, Cheng T, Gindulyte A, He J, He S, Li Q, Shoemaker BA, Thiessen PA, Yu B et al (2019) PubChem 2019 update: improved access to chemical data. Nucleic Acids Res 47(D1):D930–D940
29. Rogers D, Hahn M (2010) Extended-connectivity fingerprints. J Chem Inform Model 50(5):742–754
30. Church KW, Hanks P (1990) Word-assocation norms, mutual information, and lexicography. In: 27th Annual Meeting of the Association for Computational Linguistics, pp 76–83
31. Willett P, Barnard JM, Downs GM (1998) Chemical similarity searching. J Chem Inf Comput Sci 38(6):983–996
32. Bajorath J (2001) Selected concepts and investigations in compound classification, molecular descriptor analysis, and virtual screening. J Chem Inf Comput Sci 41(2):233–245
33. Cereto-Massague A, Ojeda MJ, Valls C, Mulero M, Garcia-Vallve S, Pujadas G (2015) Molecular fingerprint similarity search in virtual screening. Methods 71:58–63
34. RDKit: Open-source cheminformatics. http://www.rdkit.org. Accessed 24 Jan 2020.
64. Cortes-Ciriano I, Skuta C, Bender A, Svozil D (2020) QSAR-derived affinity fingerprints (part 2): modeling performance for potency prediction. J Cheminform 12:41
65. Chen Y, Kirchnair J (2020) Cheminformatics in natural product-based drug discovery. Mol Inform 39:2000171
66. Jayaseelan KV, Moreno P, Truszkowski A, Ertl P, Steinbeck C (2012) Natural product-likeness score revisited: an open-source, open-data implementation. BMC Bioinformatics 13:106
67. Seo M, Shin HK, Myung Y, Hwang S, No KT (2020) Development of Natural Compound Molecular Fingerprint (NC-MFP) with the Dictionary of Natural Products (DNP) for natural product-based drug development. J Cheminform 12:6

Publisher’s Note
Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.