Published in final edited form as:

Curr Comput Aided Drug Des. 2008 March 1; 4(1): 23–34.

Novel Algorithms for the Identification of Biologically Informative Chemical Diversity Metrics

Bhargav Theertham^{a,b}, Jenna. L. Wang^a, Jianwen Fang^c, and Gerald H. Lushington^{a,c,*}

^a Center of Excellence for Chemical Methodology and Library Development, The University of Kansas, Lawrence, KS 66045, USA

^b Department of Electrical Engineering and Computer Science, The University of Kansas, Lawrence, KS 66045, USA

^c Bioinformatics Core Facility, The University of Kansas, Lawrence, KS 66045, USA

Abstract

Despite great advances in the efficiency of analytical and synthetic chemistry, time and available starting material still limit the number of unique compounds that can be practically synthesized and evaluated as prospective therapeutics. Chemical diversity analysis (the capacity to identify finite diverse subsets that reliably represent greater manifolds of drug-like chemicals) thus remains an important resource in drug discovery. Despite an unproven track record, chemical diversity has also been used to posit, from preliminary screen hits, new compounds with similar or better activity. Identifying diversity metrics that demonstrably encode bioactivity trends is thus of substantial potential value for intelligent assembly of targeted screens. This paper reports novel algorithms designed to simultaneously reflect chemical similarity or diversity trends and apparent bioactivity in compound collections. An extensive set of descriptors are evaluated within large NCI screening data sets according to bioactivity differentiation capacities, quantified as the ability to co-localize known active species into bioactive-rich K-means clusters. One method tested for descriptor selection orders features according to relative variance across a set of training compounds, and samples increasingly finer subset meshes for descriptors whose exclusion from the model induces drastic drops in relative bioactive colocalization. This yields metrics with reasonable bioactive enrichment (greater than 50% of all bioactive compounds collected into clusters or cells with significantly enriched active/inactive rates) for each of the four data sets examined herein. A second method replaces variance by an active/inactive divergence score, achieving comparable enrichment *via* a much more efficient search process. Combinations of the above metrics are tested in 2D rectilinear diversity models, achieving similarly successful colocalization statistics, with metrics derived from the active/inactive divergence score typically outperforming those selected from the variance criterion and computed from the DiverseSolutions software.

1. INTRODUCTION

Within the voluminous, diverse, and constantly growing body of synthetically feasible organic chemistry lie the tools for potential treatment of most, if not all, health ailments that have tangible biochemical origins. The main obstacle in drug discovery is the complex task of identifying candidates from within this broad array of chemistry that provide suitable profiles for addressing the intricate combinations of pharmacological factors that govern the potency

*Address correspondence to this author at the Center of Excellence for Chemical Methodology and Library Development, and Bioinformatics Core Facility, The University of Kansas, Lawrence, KS 66045, USA; Tel: 785-864-1140; Fax: 785-864-5326; glushington@ku.edu.

of any given compound toward its intended biomolecular target. In recent years, laboratory automation has become a key tool in attempting to meet this challenge. Specifically, combinatorial parallel synthesis (CPS) has emerged as a mechanism for efficiently extending collections of available prospective therapeutics, while high throughput screening (HTS) is serving as the means for rapidly estimating the activity of these and other existing compounds toward known pharmacological targets. Tremendous public and private capital has been invested in these technologies in the hopes that they will unlock the pharmacological keys of human health.

Even with tremendous increases in automated throughput, a huge gap remains between the size of the proverbial haystack and the technological capacity to uncover needles. Even the best endowed and most efficient HTS facilities are generally still limited to screening tens of thousands ($\sim 10^4$) of compounds per day, whereas there are tens of millions ($\sim 10^7$) of potentially suitable organic chemicals already present and available in compound collections, thus one can not hope to exhaustively survey a given target. Furthermore, an estimated lower bound for the number of new distinct chemical species readily accessible *via* current CPS protocols (i.e., using known reactions and available reagents) is 10^{13} [1,2], and many practical estimates of accessible chemistry space are much higher. The effective implications of these numbers is that HTS has not yet made it possible to fully screen more than a small fraction of existing compounds, and that it is still very possible that the best candidates for pharmacological redress of any given target have not yet even been synthesized. There is thus substantial interest in being able to efficiently and systematically filter the ranks of currently available compound collections so as to select subsets that provide reasonable representation of the chemical diversity inherent in much broader compound collections, and there is comparable interest in being able to intuit which regions of the vast manifold of hitherto unsynthesized chemistry space have enhanced probabilities of yielding new bioactive compounds. Informatics methods are widely viewed as a particularly promising avenue for such tasks.

Molecular diversity analysis is the most widely applied scheme for the systematic selection of compound subsets from large (existing or hypothetical) chemical collections. In practice, this entails the computation and analysis of chemical diversity metrics capable of differentiating each compound according to functional similarity/dissimilarity relative to other species in the collection. This scheme is typically applied according to one of two diametrically opposed strategies: 1) identification of a diverse compound subset whose members are optimally distinct from each other so as to attain a subset that best represents the broad scope of the compound collection, or 2) identification of a focused subset of compounds whose members are most similar to a select number of compounds with known desirable attributes. The first of these strategies is generally applied in early stage screens wherein one may wish to identify a diverse subset of a larger compound collection that can be selected in order to perform an HTS study that should be representative of the whole library. The second scheme is more appropriate for identifying new synthesis targets similar to known actives, or for selection of existing compounds similar to known actives that could be used in targeted screens.

Molecular similarity is most commonly represented computationally either *via* the degree of overlap in molecular fingerprints (i.e., binary strings that encode the presence and arrangement of specific chemical groups) or *via* the spatial distance between any pair of molecules in a Cartesian space defined by property-based metrics. The use of fingerprint representations in chemical informatics originated as an efficient mechanism for database structure searching. Since the fingerprint overlap between two molecules also provides a reasonable measure for the extent of structural conservation in the two compounds, it has been widely adopted as an indicator of relative molecular similarity. Thus, a diverse subset of a large compound collection could theoretically be identified by choosing only those compounds whose fingerprint overlap

with any other selected compound was less than some threshold value (e.g., as accomplished by Bayada *et al.* [3]). Conversely, a focused subset could be identified by choosing only those species with overlaps greater than some minimum threshold relative to one or several compounds of interest. It is equally reasonable to implement both of these tasks *via* a property space. A diverse subset can thus be selected from a broad compound collection by partitioning the Cartesian space into cells defined according to a uniform rectilinear grid and selecting compounds uniformly across the array of occupied cells, whereas a targeted subset can be derived by choosing only compounds within a certain Cartesian distance of known actives. Both the fingerprint-based [4] and cell-based [5] schemes have been advocated as reasonable techniques for identifying diverse or targeted subsets.

However, recent skepticism has emerged regarding the true value of diversity analysis. In an interesting opinion article, Roth argues that the widely used concept of molecular diversity, as grounded in the structure and basic physico-chemical properties of the molecules, is at best pharmacologically superficial, and may be effectively irrelevant [6] because the measures of similarity and diversity are not trained to represent real trends in biological activity as measured relative to specific drug targets. The main implication of his analysis is that compounds very close together within a chemical diversity representation may well not have comparable activity toward a given target, while species far separated from each other might conversely both share comparable function.

Specific scenarios exemplifying such contrary cases are well established in the literature [6], however evidence also exists for legitimate trends relating chemical similarity measures with observed bioactivity tendencies. Martin and coworkers, for example, thoroughly surveyed the capacity of fingerprint-based similarity metrics to co-locate bioactive species close to other bioactive species [4]. By varying thresholds in Tanimoto similarity [7] as measured by relative overlap of 1024 bit Daylight fingerprints [8], they determined that compounds substantially similar ($\text{Tanimoto} \geq 0.85$) to known actives were approximately 30 times more likely to be active themselves (hit rate of ~ 30%) than subsets achieved by pure random selection (hit rate of ~1%). The breadth and extent of this survey (19,533 compounds analyzed in 115 secondary screens) affords a reasonable degree of confidence in these ratios. Similarly, in property-space-based analysis Pearlman and Smith demonstrated that known ACE (angiotensin converting enzyme) inhibitors appear to co-locate within a highly localized region of property space when mapped onto the framework of a broad diverse compound collection [5]. In particular, they found that by employing a simple overlapping spheres clustering scheme (i.e., varying the property-space radii R around molecules and assigning to the same cluster all molecules linked contiguously by nonzero sphere overlap) they were able to co-locate all ACE inhibitors within just three clusters, that collectively occupied only a small fraction of total Cartesian property space volume (0.19%; approximately equal to the ratio of known actives to total compounds). Based on this analysis, they asserted that computational diversity analysis is capable of partitioning property space according to regions containing active compounds and those with only inactive compounds. Such partitions provide a diversity-based scheme for intuiting compound activity according to the location of that compound in property space. A number of more recent studies have also probed this issue (as reviewed recently by Gorse [1]) with comparable findings.

Despite these achievements, Roth's criticism has tangible justification. Consider the numerous scenarios whereby minor structural variations (e.g., addition or deletion of a methyl) can dramatically affect observed bioactivity, rendering known actives completely inactive and vice versa. Many cases also exist where compounds with very distinct structures and physical properties are known to bind to the same target with comparable affinity. Application of most commonly employed similarity criteria is unlikely to properly account for such pronounced effects arising from such subtle dissimilarities, and it is possible that such discrepancies may

only be resolved with careful selection of diversity metrics according to assay-specific conditions. The question thus arises: if generic molecular diversity representations have demonstrated some capacity for discriminating between active and inactive compounds in space without target-specific training, how much discriminatory improvement might be gained by employing target-specific knowledge in development or refinement of the diversity model?

There are two main avenues for attempting to enhance the capacity of a diversity model to discriminate between active and inactive chemicals: tailoring the choice of chemical diversity metrics for greater activity/inactivity recognition, and applying specialized spatial partitioning schemes that can decompose chemical space into biologically fertile/infertile regions by extrapolating from the observed positions of known actives and inactives. While a substantial amount of thought has gone into the latter and a number of review articles have been generated on the topic (e.g., Schnecke and Bostrom [9]), relatively few studies have been pursued with respect to identifying specialized target-relevant metrics for diversity analysis. Whereas studies such as those of Bayada *et al.* [3] and Schuffenhauer *et al.* [10] have surveyed a variety of types of molecular descriptors for their capacity to distinguish between active and inactive regions of space across data arising from large collections of screens (38 and 160 unique assays respectively) and Gillet *et al.* explored the role of chemical substructure functionality in dictating global bioactivity (i.e., differentiating known drugs from known universal inactives) [11], we are not aware of any studies to date that have attempted to optimize the selection of diversity metrics based on assessment of activities for a single assay. This is somewhat surprising, since the formulation of quantitative structure-activity relationship (QSAR) models for specific targets has a well-established history of carefully filtering a manifold of different metrics in order to ascertain suitability for reproducing trends across activity data on that target [12]. In order to implement such a protocol within the framework of chemical diversity analysis, we have thus embarked on a preliminary exploration of methods for selecting and evaluating descriptors (and combinations of descriptors) according to their suitability for segregating active and inactive species with data sets arising from HTS analyses on specific pharmacological targets.

Our main goal in this work has been to develop and validate a systematic procedure for identifying chemical diversity metrics sensitive to the activity trends of specific targets. It is thus hoped that these metrics can be applied to the decomposition of chemical property space to highlight regions with enhanced likelihood of housing novel bioactive species. It is our intent to identify such descriptors based on trends within large screening sets (i.e., 10^4 – 10^5 compounds) such as those available in the PubChem Bioassay data repository [13], thus the metric elucidation protocol must balance predictive accuracy with reasonable computational expedience and tolerance of potentially noisy data. Furthermore, considering that even very large screening sets likely do not ensure representation of all possible biologically active chemical families for a given target, we believe that descriptors whose assessment of active regions of chemical space entail high accuracy but narrow scope (i.e., those producing very small false positive ratios, but large numbers of false negatives) are less desirable from the perspective of novel bioactive compound discovery than more general (albeit also more approximate) methods that, in addition to identifying known bioactively fertile regions, also suggest potentially under-explored regions with existing evidence of marginal bioactivity.

2. METHODS

The methodology chosen to pursue biologically sensitive metric identification entailed the following steps: **1)** procuring screening data for a given target of interest, including the chemical structures of all screened compounds and their measured bioactivity toward that target, **2)** computing a reasonable array of chemical descriptors for the screened compounds, **3)** eliminating from consideration all of those descriptors that are information-poor (i.e.,

demonstrated little variation across the compound set or are substantially redundant relative to other descriptors), **4**) identify specific descriptors according to apparent capacity for grouping active and inactive compounds in distinct regions of space, **5**) selecting N-dimensional (ND) descriptor sets according to propensity for co-locating active species predominantly in active-enriched locations of space Cartesian spaces whose, and **6**) validate those descriptor sets predicted in step 5 to be most effective in segregating according to bioactivity by validating them against test sets of activity data not considered in steps 4 and 5.

2.1. Data

In this study, we examined the following four distinct HTS data sets:

1. Growth inhibition constants of the IGROV1 human ovarian tumor cell line as a measure of potential anti-cancer activity (42,139 compounds)
2. Growth inhibition constants of the NCI-H322M human non-small cell lung tumor cell line as a measure of potential anti-cancer activity (41,160 compounds)
3. Growth inhibition of the KM12 human colon tumor cell line as a measure of potential anti-cancer activity (42,223 compounds)
4. Growth inhibition of the NCI/ADR-RES human ovarian tumor cell line, apparently analogous to OVCAR-8 [14] (29,048 compounds).

In all cases, the data sets were obtained from the Pub-Chem BioAssay repository [13], with activity data being expressed in $-\log[\text{GI}_{50}]$ form. To expedite model training, we randomly chose subsets of 10,000 molecules from each of the full sets. In all of the above cases, the stated threshold for 'activity' was listed as 6, however in a desire for our models to account for marginally bioactive species, we chose a more liberal threshold ($-\log[\text{GI}_{50}] > 5$) that appeared to capture all species that were measurably above basal (typically 7–15% of the total data set). This has the likely disadvantage of introducing more false positives into the analysis, but should also provide a better chance of recognizing all potentially fertile regions of chemical space.

2.2. Descriptors

The descriptor set considered herein included the entire collection of 1664 properties implemented within the Dragon 5.0 suite of software [15] and documented in detail within the Handbook of Molecular Descriptors [16]. This collection includes simple count properties (i.e., by atom types, functional groups and fragments), 2D and 3D topological and geometric descriptors, plus a variety of physicochemical whole molecule properties. In order to properly compute the full slate of descriptors, the SDF files reporting chemical structures chemical structure information were downloaded from PubChem [13], all valences were saturated with protons, and the structures were converted from a 2D representation to a reasonable 3D conformer using the Concord [17] program (limits set on number of heavy atoms = 100, number of rotatable bonds = 35 and maximum ring size = 15; all other parameters left at default values). Note that any molecules for which Concord coordinate generation failed were discarded from consideration; this reduced the sizes of the subsets examined to 9550, 9559, 9561, and 9489 respectively for the four data sets listed in section 2.1.

2.3. Non-Informative Descriptor Filtering

In order to reduce the computational expense of our analysis, descriptors offering low levels of non-redundant information across the compound set were discarded from consideration. This included:

1. Descriptors with standard deviation lower than 0.0001 across the set of compounds, or those descriptors that have any missing values.

2. Descriptors with no more than two distinct values across the set of compounds.
3. Descriptors with a large correlation coefficient ($|R| \geq 0.95$) relative to any other retained descriptor (i.e., within a multitude of highly correlating properties, only one property would be retained for consideration).

For each of the four data sets examined herein, these criteria reduced the number of descriptors under consideration from 1664 down to a number between 750–800.

2.4. Enrichment-Based Single Metric Selection

The capacity of a given chemical diversity metric to achieve reasonable active/inactive (A/I) spatial partitioning can be evaluated in a number of ways, but typically it entails devising some way to logically partition chemical space, determining the extent to which bioactive species are co-located in the same spatial partitions, and evaluating the relative roles of each metric (i.e., spatial dimension) in achieving bioactive co-location.

Whereas N-dimensional rectilinear grids are a fairly standard spatial decomposition scheme in molecular diversity analysis and are useful for comparing specific molecules relative to the properties of a large collection [18,19], grid cell-based statistics are a non-optimal criterion when it comes to systematic metric evaluation across high dimensional space, since cell structures (i.e., definition, dimension and density) change substantially each time a metric is added or dropped, regardless of the importance of that metric to determining chemical similarity. Distance-based clustering schemes are inherently more suitable for such tasks, in that changes in inter-compound distances arising from the addition or deletion of a metric vary smoothly in proportion to the importance of that metric to the differentiation of the compounds being measured. Examples of computational clustering schemes implemented for such purposes are thoroughly surveyed elsewhere [18].

Systematically sorting the large number (750–800) of prospective metrics examined in this study requires a computationally expedient clustering scheme that affords physico-chemically logical spatial partitions. As a reasonable compromise between these potentially competing requirements, we chose the k-means method [20] as an efficient clustering technique with a reasonable recent history of application toward chemical diversity analysis [19,21–23]. In experimenting with different numbers of k-means cluster seeds we concluded that the method consistently yielded around ten natural clusters that consistently accounted for more than 90% of all compounds in these datasets, and that specifying greater number of seeds generally produced surplus clusters with few or no constituents. Greater granularity could doubtlessly be achieved with the specification of very large numbers of seeds but this would generally prove computationally impractical. Throughout the ensuing study, we thus consistently chose the minimum number of seeds necessary to resolve ten populous clusters. This relatively coarse granularity might be expected to discriminate against metrics that might offer fine grained distinctions between actives and inactives, but is hopefully adequate to elucidate others that reveal broader trends.

At the outset of this work it was unclear whether the metrics optimal for spatial A/I enrichment would function best in combination with a small number of other metrics (as per most QSAR models, whereby the most successful models are often linear combinations of a number of descriptors), or whether it would be possible to identify metrics in isolation (i.e., according to their intrinsic ability to perform successful A/I enrichment within a one dimensional space) and have them still function as reasonable metrics in combination in N-dimensional space. We thus chose distinct metric selection schemes for both scenarios.

In order to evaluate descriptors on an individual basis for their capacity to spatially segregate actives and inactives, we devised a simple criterion that we call the A/I divergence (*AID*) formalism, which can be conceptually expressed as:

$$AID = \int_L^U \left| \frac{A(D)}{N_A} - \frac{I(D)}{N_I} \right| dD$$

where N_A and N_I are the total numbers of actives and inactives in the full data set. U and L are the upper and lower bound values for descriptor D , and the quantities $A(D)$ and $I(D)$ are distributions of actives and inactives respectively as a function of D . In practice, $A(D)$ and $I(D)$ are discrete distributions rather than continuous functions, so the computation of *AID* is best accomplished *via* a numerical approach such as the trapezoidal rule. The *AID* score of a given descriptor is somewhat dependent on the density of the grid used for numerical integration, so we experimented with different grid spacings, $(U-L)/N_{\text{cells}}$, by varying N_{cells} by unit increments from 5 to 20. From this analysis, we determined that the descriptor ordering does not vary significantly for N_{cells} values of 15 or more, thus we consistently used the grid spacing $(U-L)/15$ for all *AID* computations.

2.5. Enrichment-Based Multivariate Model Detection

While the *AID* score as derived above should be a useful measure for characterizing a descriptor's inherent capacity for A/I partitioning within a 1D projection, experience with QSAR modeling suggests that models combining multiple descriptors usually provide better discrimination of A/I trends than even the best single descriptor models do. One would thus expect there to be some value in attempting systematic evaluation of prospective small multidimensional models according to their capacity to resolve actives and inactives into mutually distinct regions of space. Exhaustive sampling of all unique combinations of 800 descriptors is not computationally practical even for mere 2D ($800 \times 799 = 639,200$) representations, however, thus a more strategically selective sampling scheme is required.

Thus to intelligently sample combinations of metrics that support strong A/I partitioning, we performed a hierarchical mesh sampling of descriptor subsets, as is rendered conceptually in Fig. 1. This scheme assumes that large descriptor sets, by virtue of their flexibility, will enable reasonable A/I partitioning, but that if one deletes descriptors that are key to differentiating between actives and inactives, the effective A/I enrichment will drop dramatically, thus we refer to the process as *Drastic Drop Isolation* (DDI). The descriptors within coarse mesh cells (i.e., sampled in groups of 50; top bar of Fig. 1) whose omission leads to a drastic A/I enrichment drop are thus retained for medium-grained sampling (i.e., sampled in groups of 10; middle bar of Fig. 1) and so forth. It was our assumption that careful ordering (a.k.a., prioritization) of the descriptors according to their likely level of information content would lead to enhanced likelihood of concentrating the most important metrics within the same coarse-mesh (and perhaps also medium-mesh) cells. Therefore, drawing an analogy to principle components analysis, we assumed that the normalized variance of descriptor values across sets of compounds would have been a reasonable reflection of their importance in reflecting trends within the compound set. We further assumed that clusters with the higher and lowest A/I ratios would collectively constitute a subset of compounds whose properties were likely to be particularly representative of the global differences between actives and inactives. Thus, the prioritization scheme by which we ordered the descriptors for the DDI scheme depicted in Fig. 1 entailed a rank according to the descriptors' normalized variance across the set of compounds present in the two clusters with highest and lowest A/I ratios. In addition to variance, we tried a similar protocol using the Fisher Criterion Score (FCS), and using the aforementioned *AID* term as prioritization scores.

Assuming that the descriptor prioritization scheme generally does somewhat correlate with A/I enrichment capacity and thus co-locates the most important descriptors within similar cells in the Fig. 1 hierarchical grid sampling scheme, it is possible to tailor the mesh spacings set for the coarse, medium and fine sampling process so as to minimize (on average) the total number of distinct sampling events required to focus in on a small dimensional subset of metrics suitable for diversity modeling. Specifically, in the event that the best descriptors were always most likely to be found within the same coarse mesh cell, spacings of 100 descriptors for the coarse mesh and 10 for the medium mesh would yield the smallest theoretical average for the number of distinct tests: $0.5 * ((800/100 + 100/10 + 10) = 14$ (or a maximum of 28). However, in the plausible event that the best descriptors spanned two or more cells in the coarse mesh, the expected number of medium mesh sampling points would increase rather steeply under the 100, 10, 1 mesh hierarchy, thus we instead opted for a compromise hierarchy of 50, 10 and 1.

As a simple alternative to the above mesh sampling scheme, we also implemented a simple strategy of testing the 2D combinations of the top metrics identified *via* the AID prioritization (as discussed in Section 2.4). In this case, for each of the four assays, all unique non-correlative (i.e., $R^2 < 0.50$) descriptor pairs from among those with top 10 AID scores were evaluated according to their capacity to co-locate bioactives into cells (within a Cartesian 10×10 grid) with strong A/I ratios.

2.6. Validation

In order to test the efficacy of our approach and to ensure that the resulting diversity space models not only explained bioactivity trends within trained data but also predicted trends within untrained data, we performed all of the above procedures in quintuplicate using leave-out-one-fold (LOOF) validation. In other words, we randomly partitioned the data set by selecting five equivalent-sized non-overlapping subsets (i.e., folds) and thus derived five distinct training sets, where each corresponded to the entirety of the set minus one of the folds. Each model trained on such a set was then tested for A/I enrichment capacity on its complementary untrained fold, and was then evaluated over the full data set.

Ultimately, we scored the performance of a given chemical diversity model based on the percentage of known actives in the training, test, and full data sets that were co-located within cluster that had an A/I enrichment factor of at least 1.20 relative to the global A/I ratio. This enrichment factor was chosen as a moderately challenging target based in part on observations that that only one clustering protocol (Ward's clustering of compounds mapped *via* 2D substructure fingerprints) in the study by Bayada *et al.* [3] achieved better than 20% A/I enrichment for diverse subset selections collectively accounting for about 50% of the total actives present within a merged set of 38 assays.

3. RESULTS AND DISCUSSION

3.1. AID-Based Metric Identification

Table 1 reports the results of the AID-based identification of descriptors demonstrating the strongest levels of A/I enrichment. Specifically, the top ten descriptors, as ranked by AID values, are reported for each of the four HTS data sets reported in section 2.1. Conceptually, AID scores thus numerically reflect the ratio of bioactives in excess of the mean that are located in disproportionately active regions plus the ratio of inactives in excess of the mean located in disproportionately inactive regions. Thus an AID score identical to 0.0 would reflect active and inactive distributions that are perfectly identical, whereas AID scores of 2.0 would reflect distributions that are completely mutually exclusive. AID scores of the magnitudes reflected in Table 1 reflect metrics that have substantially different spatial distributions of actives and inactives. Example distributions are depicted in Fig. 2 for the GGI1 and Mor06m descriptors

from the NCI-H322M data set. The effective discrimination between actives and inactives in the former case arises from compounds with smallest GGI1 values having a substantially higher rate of inactivity than average, whereas those with larger GGI1 values tending to have a greater tendencies toward bioactivity. For Mor06m, the compounds in general seem to display a standard Gaussian distribution, however the bioactives appear to have a pronounced shift in values toward smaller metric values, while inactives have a similar shift toward larger values.

3.2. Multivariate Model Search

Table 2 reports the results of the multivariate model search based on DDI. Specifically, it details the proportion of bioactive compounds found in clusters with enriched A/I ratios for the various chemical diversity models achieved in our analysis of the four HTS data sets reported in section 2.1. Results were reported for chemical diversity metric search strategies based on both variance- and *AID*-prioritization. While some effort was also invested in implementing a FCS-prioritization search, the latter failed to yield metrics with reasonable A/I enrichment. In retrospect, this is not particularly surprising: the FCS paradigm functions best given a gaussian distribution according to descriptor values, however in practice different descriptors have broadly varying distribution forms, many of which are not normal.

Initially, we had established a goal of being able to identify for any HTS assay a set of diversity metrics capable of resolving at least 50% of all actives into clusters with A/I enrichment ratios of at least 1.2. In this we have succeeded. Table 2 reflects that for all four assays examined thus far, the variance-based search yielded models with at least 59% of actives in enriched clusters, while the *AID*-based method yielded models with at least 54%. The average performance of the best variance-based models (65.6% of actives in enriched over the four assays) exceeded that of the *AID* counterparts (60.9% of actives enriched), however the *AID* model generation appears to yield much better stability as a function of training set. Specifically, from variance-based searches, the set of metrics selected from the five different folds of a data set varied substantially from fold to fold, with the same descriptor only rarely emerging as a key metric in the models produced from different folds. From *AID*-based analyses, however, we nearly always obtained the same model from fold to fold. In the only two exceptions to this trend, folds 3 and 4 from the IGROV1 data set, the primary metric from the consensus model (as determined from folds 1,2,5) also appeared as the lead component of the unique models, and the practical difference in performance between the consensus model and the unique models for these two folds was negligible. Specifically, for fold 3 the consensus model (descriptor 103 alone) sequestered 55.17% of bioactives into A/I enriched clusters, while the unique model (descriptors 103+348) sequestered 55.73%. For fold 4, the consensus model sequestered 55.12%, while the unique model achieved 55.88%.

Several interesting trends emerge from Table 2. Firstly, both the variance-based and *AID*-based metric search schemes select the same descriptor for multiple data sets: hypnotic-80 achieves the best variance-based model (in terms of actives segregation) for both the IGROV1 and NCI-H322M data sets, while MPC09 attains the based *AID*-based model for IGROV1, NCI-H322M and KM12. This suggests that descriptors that achieve good bioactive segregation for one cancer model may also be reasonable metrics for multiple other cancer models. This observation bodes well for the application of our method toward selection of diversity metrics appropriate to broader bioactivity classifications. Also interesting is the high degree of consistency across folds with which the *AID*-based search selects metrics. While variance-based models often achieve better A/I enrichment ratios than their *AID* counterparts, the model derived based on training to one fold is usually different from that determined from other folds, thus implying substantial dependence on the training set choice. For subsets with imperfect representativity, the variance method may thus be susceptible to generating over-trained models that account very well for the data at hand, but are less appropriate for compounds

differing significantly from those in the training set. Conversely, the consistency of the *AID*-based metric selection suggests a greater inherent sensitivity to the true distribution of bioactive vs inactive compounds within these data sets.

Another advantage of the *AID* search scheme comes with methodological uniformity. In all *AID* searches, the most important descriptors were consistently found among those with high *AID* scores (in these cases always within the top 10). By contrast, in the case of the variance search scheme substantial variation exists in the precise coarse mesh range of prioritized descriptors wherein the drastic drop occurs. For example, in the case of NCI-H322M data set, the drastic drop occurs for folds 1–3 when discarding descriptors 100–150, but in the 50–100 range for fold 4, and between 150–200 range for fold 5. Thus while the convoluted hierarchical search procedure described in section 2.4 is likely unavoidable in a variance-based search, the *AID* criterion may enable effective circumvention by preselecting a small number of descriptors with high *AID* values and then evaluating a small number of permutations.

The most surprising result from our attempt to resolve multivariate models was the fact, for both the variance- and *AID*-based prioritization schemes, that the strict criterion of maximizing A/I enrichment usually leads to isolation of 1d models. Given the reasonable assertion by Bayada *et al.* that “no single descriptor can possibly be sufficient” to describe chemical diversity [3], we must thus admit that our DDI scheme does not directly elicit the best combinations of metrics that lead to informative A/I enriched chemical diversity models. Nonetheless, the individual metrics isolated by this scheme clearly do all contain important information about the attributes of molecules that distinguish actives from inactives. The salient question thus is whether these metrics, when used in combination to define multidimensional chemical property spaces, still support meaningful spatial discrimination between active and inactive compounds.

3.3. A/I Enrichment for Combinations of Metrics

Most conventional chemical diversity schemes currently available involve the construction of an N-dimensional chemical property space whose structure should co-localize functionally similar molecules in similar regions of space. Metrics that have been chosen for their sensitivity to bioactivity trends should hopefully also collect active species in spatial regions somewhat distinct from those with the greatest proportional density of inactive species. In Fig. 3, we report a series of 2d chemical property space distributions for the IGROV1 data set, using metrics from among the top descriptors recommended by the DiverseSolutions (DVS) suite, our own variance-based analysis, and our *AID* prioritization scheme.

Net compound density distributions within the three plots vary somewhat from the DVS to variance to *AID* models, thus suggesting that each model elicits a different representation of chemical similarity and diversity. The DVS model exhibits several distinct clusters, which is a favorable attribute in chemical diversity modeling, where greater dispersion throughout space afford greater functional differentiation among compounds. The variance-based metrics, by contrast, have an intense peak at the coordinate origin and quickly drop off to a fairly uniformly sparse distribution throughout the remainder of the space. The *AID* formalism affords a pattern somewhere between the previous two, with one primary dense cluster region at the origin, but also pronounced regions of density moving out toward the periphery.

In each of the three representations, the regions of greatest overall compound density all correspond to areas that have A/I ratios that tend moderately or strongly toward inactivity. Within these models, those representations that yield the best overall A/I enrichment characteristics are those that have some higher A/I cells in regions somewhat proximate to the more dense regions of space. Specifically, in the case of the *AID* model, cells (8,2), (9,3), (9,4), (9,5) and (9,6) (as indexed by analogy with increasing Cartesian coordinates: x increasing from

left to right and y from bottom to top) all have A/I ratios greater than 1.2 times the mean and are all located in cells with total populations that are a reasonable fraction (> 10%) of the most dense cell. As a result, it is not particularly surprising to discover that this model is found to place 62% of all actives in cells with A/I ratios greater than 1.2 times the mean. It is further worth noting that the *AID* model has a substantial contiguous band of cells in the upper right and lower right quadrants whose A/I ratios are consistently greater than the mean.

The variance model has a very different distribution, in that it has only two cells, (3,2) and (3,3), with significantly enhanced A/I ratios that simultaneously have an overall compound density reasonably proximate to that of the most dense cell. In compensation, there is a wide swath of contiguous chemical space that, although sparse in terms of total compound density, has a consistently very strong A/I ratio, leading to co-location of 53% of active in significantly A/I enriched cells.

Interestingly, the DVS model, whose metrics were not explicitly selected with specific bioactivity co-location in mind, is actually quite successful in grouping actives into A/I enriched regions (58%), largely by virtue of locating numerous enriched cells in fairly dense regions of space: (6,4), (8,4), (5,5), (6,5), (8,5) and (6,7). The regions of enhanced activities tend to be more broadly distributed throughout DVS chemical space than for either the *AID* or variance models, with smaller regions of contiguous enhanced A/I ratio.

In practical terms, the extent to which models exhibit regions of contiguous A/I enrichment may prove strategically important when attempting to select compound subsets with strong A/I enrichment. Assuming that the computed metrics on which the model is based are inherently approximate, it is reasonable to expect that the best chances for selecting subsets that have optimal A/I ratios will be attained by selecting from regions of space that not only are predicted to have high A/I enrichment but also are surrounded by other regions with high A/I enrichment. Conversely, selecting from very localized regions with high A/I ratio but with proximate pockets of low activity may yield lower success rates due to uncertainty in the exact locations (i.e., as defined by metric values) of specific compounds.

In view of the success of the *AID* method in yielding metrics that co-locate most IGROV1 active species in regions with both solid A/I ratios and fair contiguity of enriched cells, it is worth examining the performance of 2D combinations of *AID* metrics relative to the other screens. Table 4 thus reports the capacity of the best performing *AID* metric pairs for all four screens, while Fig. 4 depicts the spatial distribution of total compounds and A/I enrichment trends in the NCI-H322M, KM12, and NCI/ADR-RES data sets resulting from these pairs.

From Table 4 it is evident that selection of metric pairs from within the list of top *AID* scoring descriptors leads to 2D models with strong A/I enrichment in all of the data sets examined herein, and that this performance is fairly reproducible when metrics are chosen based on performance on one set of compounds (training set) and then applied blindly to a distinct test set. From Fig. 4, however, one may ascertain that the active/inactive profiles from one screen to another can be quite different. In particular, qualitative similarities do exist between the NCI-H322M and KM12 data sets and the IGROV1 data set reported in Fig. 3 in that all three also have large contiguous regions of enhanced bioactivity along the periphery of dense regions of compound space, all three show predominantly active cells in areas of fairly substantial density, but in each case actives are generally somewhat depleted from regions of maximal density. The NCI-ADR-RES data set, however, is very distinct in that four of the most dense cells of the entire compound set, (9,4), (9,5), (10,5) and (10,6) all have elevated A/I ratios. However, far fewer cells in the full space slant toward enhanced bioactivity ratios, and relatively little contiguity exists among the different bioactively enhanced cells. This distinct nature of this latter data set, which echoes earlier observation that *AID* scores for NCI/ADR-

RES descriptors were markedly lower than those of the other three data sets (as per Table 1), and that the metrics consistently resolved for the three other data sets *via* the DDI method were not among the more important for NCI/ADR-RES (as per Table 2).

The fact that the NCI/ADR-RES profile differs so markedly from IGROV1 in particular is fascinating as both are nominally screens for ovarian cancer growth inhibition. Indeed, while there is clear experimental evidence of similar activity trends (e.g., a correlation of $R^2 = 0.76$ in $-\log(\text{GI50})$ values between the two data sets), there are also significant differences in that 312 molecules in the data set that are found to either be active in IGROV1 and inactive in NCI/ADR-RES or vice versa (as compared with a fairly modest 796 that are found to be active in both screens). These incongruities are perhaps not all that surprising given the fact that fundamental features of the cell lines such as structural and karyotypic complexity vary substantially between the two [27]. Given that our method should treat low level random differences between data sets as minor perturbations that should lead to only minor differences in metric *AID* scores and model structure and performance, the fact that such a pronounced analytical differences emerge between our models for NCI/ADR-RES and IGROV1 suggests to us that whatever differences exist between the two may well be systematic rather than random. Further study of the differences in bioactivity profiles of these two similar but apparently distinct targets is planned. Regardless of the precise origin of trend differences between IGROV1 and NCI/ADR-RES, the fact that our method clearly detects these incongruities that might not be intuited from the correlation analysis alone is evidence of the method's potential utility in characterizing and differentiating activity profiles of assays.

4. SUMMARY AND CONCLUSIONS

The analyses that we have performed in this study, aimed at developing chemical diversity models with direct relevance toward specific biological targets, have produced a number of metrics that yield reasonable enrichment of actives in localized regions of chemical property space. Both our variance-based multi-descriptor search scheme as well as our computationally simpler *AID*-based prioritization seem capable of resolving descriptors that perform well both as a function of A/I enrichment in k-means clusters and in rectilinear Cartesian grid cells. Both methods appear to be generally applicable to reasonable-sized sets HTS screening data (~10,000 compounds), with neither failing to yield a good A/I partitioning model (at least 50% of actives located within clusters whose A/I enrichment ratio is at least 1.2 relative to the mean) on any data set attempted thus far. Both methods have produced models that withstood rigorous validation (five-fold LOOF cross validation), however metrics selected *via* the *AID* scheme appear to be much more likely to be consistently identified from one fold to the next -- an indication that the *AID* method is capable of uncovering metrics whose properties more reliably reflect the characteristic distribution of bioactive compounds for a given target.

While these achievements are indicative of a promising methodology, it is important to view this work as preliminary in nature, with ample room for improvement in further study. For example, attempts to resolve multidimensional diversity models *via* the DDI method typically tend to collapse to 1D models. The corresponding 1D descriptors do nonetheless generate reasonable multidimensional A/I enrichment profiles when used in combination with other high scoring metrics, however it remains a distinct possibility that unresolved combinations of lower-ranking descriptors might actually perform better than those pairings that we have detected herein.

It is our suspicion that the tendency we have observed toward a collapse by the DDI method to single metric models might both be intrinsically related to the choice of k-means as a clustering tool. The k-means method may be viewed in part as a boundary detection scheme, whereby it partitions space by identifying places where sizeable gaps exist between distinct

clumps of compounds. Within data sets that have approximately homogeneous distribution throughout space, this is a useful way of logically dividing regions, however in highly non-uniform distributions such as chemical screening compounds frequently are (i.e., dominant collections of Lipinski-compliant [28] drug-like species all collected together in a very dense region, with much broader and more diffuse regions of unique and more specialized species) the k-means method tends to introduce fewer logical boundaries in the dense drug-like space (where gaps are relatively small) and many more in the more sparse and likely less relevant peripheral areas. In such a case, the k-means method may be inadequately granular for diversity analysis, and methods that cluster according to approximate target cluster populations instead of approximate cluster volume, may prove more successful.

While we believe that our choice of an extensive set of molecular descriptors such as those presented in the Dragon Suite [15] should prove to be sufficiently diverse to afford a successful proof of concept for the algorithms examined herein, it is worth noting that there are other very distinct descriptor sets that may also be very relevant to such analyses. One obvious choice includes VolSurf descriptors [29] that encode molecular volumetric and solvent-accessible surface profiles into distinct parameters for QSAR analyses. Another entails the recognition of characteristic substructural features such as may be identified and classified *via* methods such as frequent subgraph mining [30].

At this point we are tentatively quite confident that our *AID* formalism will prove to be a successful scheme for feature selection as applied to biologically sensitive chemical diversity analysis. Given the importance of the application, however, it is very much worth exploring whether other feature selection techniques pursued for the sake of active/inactive classification (e.g., as per Wegner *et al.* [31]) might usefully supplement the capacity of *AID* for tasks such as this.

Finally, it is also possible that the predictive capacity of a method such as ours might be enhanced by extending the focus from single targets to trends across multiple targets within known gene families -- a concept that has gained a reasonable amount of interest within the chemical informatics community [32]. The fact that we observed multiple instances where variables achieved strong A/I enrichment scores for more than one distinct data set suggests that our method should be applicable to analysis of data sets involving a merger of results from multiple screens across related targets, and should be well suited to resolution of specific metrics that recognize activity patterns that are conserved within the family. Extension of the method toward multiple sources of bioactivity would have the added benefit of providing a formalism by which one may gauge the more general prospective of value of specific compounds that one may wish to synthesize or otherwise acquire for a screening collection.

Acknowledgments

This work was supported by the Kansas IDeA Network of Biomedical Research Excellence (P20 RR016475) and the KU Center of Excellence in Chemical Methodologies and Library Development (P50 GM069663). We would also like to express our gratitude to John Gauch for valuable discussions.

References

1. Gorse DA. *Curr Topics Med Chem* 2006;6:3–18.
2. Nikitin S, Zaitseva N, Demina O, Solovieva V, Mazin E, Mikhalev S, Smolov M, Rubinov A, Vlasov P, Lepikhin D, Khachko D, Fokin V, Queen C, Zosimov V. *J Comput Aided Mol Des* 2005;19:47–63. [PubMed: 16059666]
3. Bayada DM, Hamersma H, van Geerestein VJ. *J Chem Inf Comput Sci* 1999;39:1–10.
4. Martin YC, Kofron JL, Traphagen LM. *J Med Chem* 2002;45:4350–4358. [PubMed: 12213076]
5. Pearlman RS, Smith KM. *J Chem Inf Comp Sci* 1999;39:28–35. and references therein

6. Roth HJ. *Curr Opin Chem Biol* 2005;9:293–295. [PubMed: 15939331]
7. Willett P, Barnard JM, Downs GM. *J Chem Inf Comput Sci* 1998;38:983–996.
8. Thor; Merlin. Version 4.62. Daylight Chemical Information Systems Inc; Irvine, CA: 2002.
9. Schnecke V, Bostrom J. *Drug Discov Today* 2006;11:43–50. [PubMed: 16478690]
10. Schuffenhauer A, Brown N, Selzer P, Ertl P, Jacoby E. *J Chem Inf Model* 2006;46:525–535. [PubMed: 16562980]
11. Gillet VJ, Willett P, Bradshaw J. *J Chem Inf Comput Sci* 1998;38:165–179. [PubMed: 9538517]
12. Dudek AZ, Arodz Tomasz GJ. *Comb Chem & High Throughput Screen* 2006;9:213–228. [PubMed: 16533155]
13. The PubChem Project, National Institutes of Health/NCBI. 2006. <http://pubchem.ncbi.nlm.nih.gov/>
14. Liscovitch M, Ravid D. *Cancer Lett* 2007;245:350–352. [PubMed: 16504380]
15. Dragon 5.0. Talete SRL; Milan, IT: <http://www.disat.unimib.it/chm/Dragon.htm>
16. Todeschini, R.; Consonni, V. *Handbook of Molecular Descriptors*. In: Mannhold, R.; Kubinyi, H.; Timmerman, H., editors. *Series of Methods and Principles in Medicinal Chemistry*. Vol. 11. Weinheim: Wiley VCH; 2000. and references therein
17. Concord 2.07. The Tripos Associates; St. Louis, MO: 2005.
18. Maldonado AG, Doucet JP, Petitjean M, Fan BT. *Mol Diversity* 2006;10:39–79.
19. Fontaine F, Pastor M, Gutierrez-de-Teran H, Lozano JJ, Sanz F. *Sampling of molecular libraries*. *Mol Diversity* 2003;6:135–147.
20. MacQueen, JB. *Some Methods for classification and Analysis of Multivariate Observations*. *Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability*; Berkeley: University of California Press; 1967. p. 281–297.
21. Matero S, Lahtela-Kakkonen M, Korhonen O, Ketolainen J, Lappalainen R, Poso A. *Chemometrics Intell Lab Sys* 2006;84:134–141.
22. Ashton M, Barnard J, Casset F, Charlton M, Downs G, Gorse D, Holliday J, Lahana R, Willett P. *Quant Struct-Act Relat* 2002;21:598–604.
23. Gutierrez-de-Teran H, Lozano JJ, Segarra V, Sanz F. *Comb Chem and High Throughput Screen* 2002;5:49–57. [PubMed: 11860339]
24. Ghose AK, Viswanadhan VN, Wendoloski JJ. *J Comb Chem* 1999;1:55–68. [PubMed: 10746014]
25. Wiener H. *J Am Chem Soc* 1947;69:2636–8.
26. Randic M, Jurs PC. *Quant Struc-Acti Relationships* 1989;8:39–48.
27. Roschke AV, Tonon G, Gehlhaus KS, McTyre N, Bussey KJ, Lababidi S, Scudiero DA, Weinstein JN, Kirsch IR. *Cancer Res* 2003;63:8634–8647. [PubMed: 14695175]
28. Lipinski CA, Lombardo F, Dominy BW, Feeney PJ. *Adv Drug Del Rev* 2001;46:3–26.
29. Cruciani G, Pastor M, Guba W. *Eur J Pharm Sci* 2000;11:S29–S39. [PubMed: 11033425]
30. Huan, J.; Wang, W.; Prins, J. *Efficient Mining of Frequent Subgraph in the Presence of Isomorphism*. *Proceedings of the 3rd IEEE International Conference on Data Mining (ICDM)*; 2003. p. 549–552.
31. Wegner JK, Frölich H, Zell A. *J Chem Inf Comput Sci* 2004;44:921–930. [PubMed: 15154758]
32. Miller J. *Curr Topics in Med Chem* 2006;6:19–29.

Fig. 1.

Example of the hierarchical mesh sampling scheme. In the coarse mesh (top) the chemistry space A/I enrichment performance is sampled across metric spaces entailing the top 800, 750, 700, etc., descriptors (as ranked according to priority). In the medium mesh, those 50 descriptors deemed most important to A/I enrichment are then evaluated according to metric subsets of size 50, 40, 30, etc. (again ranked by priority). Finally all unique combinations of those 10 descriptors deemed most important in the medium mesh analysis were tested for A/I enrichment capacity.

Fig. 2. Normalized distributions of active (grey) and inactive (black) compounds over the BELp1 (top) and Mor05v (bottom) metrics for the NCI-H322M compound set.

Fig. 3.

Compound density and A/I enrichment profiles for 2d chemical space models of the IGROV1 data set as evaluated using top metrics derived from DiverseSolutions (DVS) suite, our own variance-based analysis, and our AID scheme. In each case, metrics were chosen for optimal capacity to produce optimal A/I enrichment rates within cells defined by a 10×10 grid over the two chosen metrics. The axes for these grids correspond to: DVS(x): $0.75 < bcut_hdonor_S_invdist_000.120_K_H < 1.25$; DVS(y): $2.0 < bcut_tabpolar_S_invdist_000.250_K_H < 7.0$; var(x): $0 < RDF150u < 76000$; var (y): $0 < Hypnotic-80 < 410$; AID(x): $0.0 < ATS8e < 10.0$; AID(y): $0.0 < MPC09 < 450.0$. All relevant metrics are defined and cited in ref. [14]. Full blackness in the density distribution is defined as the density of the most populous cell, while less dense cells are rendered proportionally in grey scale (down to pure white for empty or near-empty cells). In the A/I enrichment profiles, bluish cells have A/I ratios greater than the mean (pure blue is defined as having an A/I ratio 2.4 times the mean) while reddish cells are similarly slanted toward inactivity.

Fig. 4.

Compound density and A/I enrichment profiles for 2D chemical space models of the NCI-H322M, KM12 and NCI/ADR-RES data sets as evaluated using metrics derived *via* pairing top *AID*-scoring descriptors (Table 1). In each case, metrics were chosen for optimal capacity to produce optimal A/I enrichment rates within cells defined by a 10×10 grid over the two chosen metrics. The axes for these grids correspond to: NCI-H322M(x): $-10.0 < \text{Mor05v} < 0.0$; NCI-H322M(y): $3.5 < \text{BEHv1} < 4.1$; KM12(x): $8.0 < \text{BID} < 50.0$; KM12(y): $0.0 < \text{RDF050m} < 35.0$; NCI/ADR-RES(x): $0 < \text{piID} < 7.5$; NCI/ADR-RES(y): $2.5 < \text{RDF070m} < 6.0$. All relevant metrics are defined and cited in ref. [14]. Full blackness in the density distribution is defined as the density of the most populous cell, while less dense cells are rendered proportionally in grey scale (down to pure white for empty or near-empty cells). In the A/I enrichment profiles, bluish cells have A/I ratios greater than the mean (pure blue is defined as having an A/I ratio 2.4 times the mean) while reddish cells are similarly slanted toward inactivity.

Table 1
AID Scores for the Top 10 Descriptors for Each of the Four HTS Data Sets

IGROV1		NCI-H322M		KM12		NCI/ADR-RES	
Desc.	<i>AID</i>	Desc.	<i>AID</i>	Desc.	<i>AID</i>	Desc.	<i>AID</i>
piPC10	0.49	BELp1	0.56	MPC09	0.50	piID	0.34
MPC09	0.46	Mor05v	0.56	Mor05v	0.50	C-026	0.34
nR10	0.45	Mor05m	0.55	BEHv1	0.48	piPC08	0.33
piPC06	0.45	piPC10	0.54	nR10	0.47	RDF070m	0.31
Mor05v	0.43	BEHv1	0.54	nR06	0.47	PCR	0.31
Mor13v	0.43	D/Dr10	0.54	Mor08v	0.46	nR07	0.30
BEHv1	0.42	Mor08m	0.54	piPC10	0.45	Mor14v	0.30
nR06	0.42	MPC09	0.54	D/Dr10	0.45	X0A	0.29
ATS8e	0.41	nR10	0.53	BID	0.44	MSD	0.29
RDF065m	0.40	HGM	0.53	RDF050m	0.44	nCconj	0.29

Percent of Actives Co-Located Within A/I Enriched Clusters for Strongly A/I Segregative Chemical Diversity Models for the Four HTS Assay Data Sets Described in Section 2.1

Set - Fold	Dimensions: Descriptors	Train. %	Test. %	Tot. %	Dimensions: Descriptors	Train. %	Test. %	Tot. %
1 - 1	1d: 81	51.23	52.28	51.44	1d: 197	56.93	51.38	55.82
1 - 2	1d: 77	41.90	85.45	50.61	1d: 197	54.22	62.22	55.82
1 - 3	1d: 1656	73.24	71.24	72.84	1d: 197,939	55.73	53.61	55.31
1 - 4	4d: 129,1646,781,1182	51.90	49.60	51.44	1d: 197,1082	55.88	59.47	56.60
1 - 5	1d: 1656	72.96	72.36	72.84	1d: 197	56.30	53.90	55.82
2 - 1	2d: 143,855	51.79	71.74	55.78	1d: 197	59.85	81.10	64.10
2 - 2	1d: 694	64.53	65.83	64.79	1d: 197	67.22	51.62	64.10
2 - 3	1d: 796	55.12	75.82	59.26	1d: 197	67.86	49.06	64.10
2 - 4	1d: 1656	77.11	72.81	76.25	1d: 197	61.64	73.94	64.10
2 - 5	2d: 786,791	61.76	62.76	61.96	1d: 197	63.93	64.78	64.10
3 - 1	1d: 701	45.66	32.46	43.02	1d: 197	57.05	53.26	56.29
3 - 2	1d: 84	62.97	40.07	58.39	1d: 197	54.29	64.30	56.29
3 - 3	2d: 1290,125	60.22	45.15	57.21	1d: 197	56.68	54.74	56.29
3 - 4	2d: 134,797	40.34	23.13	36.90	1d: 197	56.86	54.02	56.29
3 - 5	1d: 795	50.70	55.60	51.68	1d: 197	56.58	55.14	56.29
4 - 1	1d: 1652	27.23	26.22	27.05	1d: 206	67.10	75.14	68.71
4 - 2	1d: 697	62.23	56.43	61.07	1d: 206	68.96	67.70	68.71
4 - 3	3d: 871,751,1185	54.60	69.15	57.51	1d: 206	70.76	60.50	68.71
4 - 4	1d: 697	59.30	68.15	61.07	1d: 206	68.51	69.50	68.71
4 - 5	1d: 84	31.34	33.54	31.78	1d: 206	68.21	70.70	68.71

For each assay, results from 5 distinct training folds are reported, with the resulting chemical diversity model being listed according to model dimensionality and the specific metrics used (metrics numbered according to Dragon [15]). For each model, A/I segregation is reported for the training set on which the models was based, the corresponding test set, and total set (including both the training and test components). For each assay, the identity of the best performing model was identified as that giving the best total set segregation, and is written in bold. More detail for some key descriptors is provided in Table 3.

Table 3

Descriptor Definitions for Table 2 Models with Optimal A/I Enrichment Values

Assay	Variance-Based Metric	AID-Based Metric
IGROV1	1656(hypnotic-80) = Ghose-Viswanadhan-Wendoloski hypnotic- like index at 80% [24]	197(MPC09) = molecular path count of order 09 [16]
NCI-H322M	1656(hypnotic-80) = Ghose-Viswanadhan-Wendoloski hypnotic- like index at 80% [24]	197(MPC09) = molecular path count of order 09 [16]
KM12	84(Wap) = all-path Wiener index [25]	197(MPC09) = molecular path count of order 09 [16]
NCI/ADR-RES	697(QZZv) = Qzz COMMA2 value/weighted by atomic volumes [16]	206(piPC08) = molecular multiple path count of order 08 [26]

Table 4

A/I Enrichment Performance of Pairs of Descriptors, Selected from Table 1 for Optimal Ability to Co-Localize Actives into Predominantly Active Cells Within 10×10 Rectilinear 2D Cartesian Cells

Screen	Metrics	Fold	A/I Enrichment (Training Set)	A/I Enrichment (Testing Set)
IGROV1	nr07, nr11	1	61.8%	50.8%
NCI-H322M	piPC05, BEHv3	4	60.2%	54.5%
KM12	Mor06v, Mor09v	2	62.7%	58.1%
NIC/ADR-RES	piID, RDF070m	2	62.1%	60.6%