Independent Vector Analysis for Molecular Data Fusion: Application to Property Prediction and Knowledge Discovery of Energetic Materials

Zois Boukouvalas*, Monica Puerto*, Daniel C. Elton†, Peter W. Chung‡, and Mark D. Fuge‡
*Dept. of Mathematics and Statistics, American University, Washington, DC 20016
†Imaging Biomarkers and Computer-Aided Diagnosis Lab, Radiology and Imaging Sciences,
National Institutes of Health Clinical Center, Bethesda, MD 20892
‡Dept. of Mechanical Engineering, University of Maryland, College Park, College Park, MD 20742

***Abstract*—Due to its high computational speed and accuracy compared to ab-initio quantum chemistry and forcefield modeling, the prediction of molecular properties using machine learning has received great attention in the fields of materials design and drug discovery. A main ingredient required for machine learning is a training dataset consisting of molecular features—for example fingerprint bits, chemical descriptors, etc. that adequately characterize the corresponding molecules. However, choosing features for any application is highly non-trivial, since no "universal" method for feature selection exists. In this work, we propose a data fusion framework that uses Independent Vector Analysis to uncover underlying complementary information contained in different molecular featurization methods. Our approach takes an arbitrary number of individual feature vectors and generates a low dimensional set of features—molecular signatures—that can be used for the prediction of molecular properties and for knowledge discovery. We demonstrate this on a small and diverse dataset consisting of energetic compounds for the prediction of several energetic properties as well as for demonstrating how to provide insights onto the relationships between molecular structures and properties.**

***Index Terms*—Data fusion, blind source separation, interpretability**

I. INTRODUCTION

Machine learning (ML) has recently been used for the prediction of molecular properties and studies have shown that it can provide accurate and computationally efficient solutions for this task [1]–[5]. The prediction ability of a ML model highly depends on the proper selection of a training dataset, via feature vectors, that can fully capture certain characteristics of a given set of molecules. A common way to represent a molecule is through a string representation called the Simplified Molecular-Input Line-Entry System (SMILES) string [6]. Although working directly with SMILES strings has shown to be effective in some ML tasks [7], most of the ML methods require vector or matrix variables rather than strings.

Basic classes of featurization methods that have been widely used in the literature include cheminformatic descriptors, molecular fingerprints, and custom graph convolution based fingerprints, where each featurization method provides different – though not necessarily unique – information about a molecule.

Thus, it is important to find answers to the question "how can disparate datasets, each associated with unique featurization methods, be integrated?" The question is motivated by the desire to create automated approaches where data generation techniques are integrable with ML models. *Data fusion* methods [8], [9] may serve this purpose, since they enable simultaneous study of multiple datasets by, for instance, exploiting alignments of data fragments where there is a common underlying latent space.

A naive approach could be to simply concatenate data that has been generated by different featurization methods. However, this typically leads to a curse of dimensionality, that could affect the performance of a ML algorithm and make it impossible to discover the features of greatest importance. Therefore, selecting a model that generates a set of molecular feature vectors by effectively exploiting complementary information among multiple datasets is an important issue.

Blind source separation (BSS) techniques enable the joint analysis of datasets and extraction of summary factors with few assumptions placed on the data. This is generally achieved through the use of a generative model. One of the most widely used BSS techniques is independent component analysis (ICA) [10], [11]. The popularity is likely because by only requiring statistical independence of the latent sources it can uniquely identify the true latent sources, subject only to scaling and permutation ambiguities. However, ICA can decompose only a single dataset. In many applications multiple sets of data are gathered about the same phenomenon. These sets can often be disjoint and merely misaligned fragments of a shared underlying latent space. Multiple datasets could therefore share dependencies. This motivates the application of methods that

Support for this work is gratefully acknowledged from the U.S. Office of Naval Research under grant number N00014-17-1-2108 and from the Energetics Technology Center under project number 2044-001. Partial support is also acknowledged from the Center for Engineering Concepts Development in the Department of Mechanical Engineering at the University of Maryland, College Park. We thank Dr. Ruth M. Doherty and Dr. Bill Wilson from the Energetics Technology Center for their encouragement, useful thoughts, and for proofreading the manuscript.

Daniel C. Elton contributed to this article in his personal capacity. The views expressed are his own and do not necessarily represent the views of the National Institutes of Health or the United States Government.

can jointly factorize multiple datasets, like independent vector analysis (IVA) [11]. IVA is a recent generalization of ICA to multiple datasets that can achieve improved performance over performing ICA on each dataset separately.

In this paper, we propose a novel data fusion framework that uses IVA to exploit underlying complementary information contained across multiple datasets. Namely, we show that information that has been generated by different molecular featurization methods can be fused to improve learning of the chemical relationships.

Note here that achieving perfect regression error is not the the goal of this work. Instead, our goal is to determine how to generate feature vectors by combining datasets from multiple featurization methods to improve the learned response of the data-fused model over the performance of the individual feature vectors treated separately. Particularly noteworthy is that the proposed approach is *completely unsupervised*, *parameter free*, *computationally attractive* when compared with existing methods [5], easily *interpretable* due to the simplicity of the generative model, and *does not require a large amount of training samples* in order to achieve a desirable regression error.

The remainder of this paper is organized as follows. In Section (II), we provide a brief background on the IVA method and the regression procedure. Section (III) provides the results of the regression procedure and associated discussions. The conclusions and future research directions are presented in Section (IV).

II. MATERIALS AND METHODS

The BSS model is formulated as follows. Let $\mathbf{X} \in \mathbb{R}^{d \times N}$ be the observation matrix where d denotes the dimension of the feature vector and N denotes the total number of molecules. The noiseless BSS generative model is given by

$$\mathbf{X} = \mathbf{A}\mathbf{S}, \tag{1}$$

where $\mathbf{A} \in \mathbb{R}^{d \times P}$ is the mixing matrix, and $\mathbf{S} \in \mathbb{R}^{P \times N}$ is the matrix that contains the sources that need to be estimated and will be used as the new feature vector for the ML task. One of the most widely used methods for solving the BSS problem (1) is ICA and its basic assumption is that the source signals are statistically independent [11]–[13].

By rewriting (1) using random vector notation, we have $\mathbf{x}(n) = \mathbf{A}\mathbf{s}(n),\ n = 1, \ldots, N$, where n is the sample index denoting the nth molecule, $\mathbf{s}(n) \in \mathbb{R}^P$ are the unknown sources that need to be estimated, and $\mathbf{x}(n) \in \mathbb{R}^d$ are the mixtures. Our interest is in dealing with overdetermined problems where $d > P$. This can be reduced to the case where $d = P$ using a dimensionality reduction technique like principal component analysis (PCA). For the purpose of this work and for the rest of this paper we assume that the samples of each unknown source are independent and identically distributed and therefore, to simplify the notation we drop the sample index n.

Although ICA has been shown to be very useful in many applications, it decomposes a single dataset at a time. For practical applications that involve more than one dataset one could perform ICA separately on each dataset and align the subsequent results. However, this approach could be considered suboptimal since performing ICA individually on each dataset will ignore any dependencies that exist among them.

A. Independent Vector Analysis

It is common for datasets that have been generated from different featurization methods to have some inherent dependence among them. IVA generalizes the ICA problem by allowing for full exploitation of this dependence leading to improved performance beyond what is achievable by applying ICA separately to each dataset. Additionally, IVA automatically aligns dependent sources across the datasets, thus bypassing the need for a second permutation correction algorithm.

IVA is similar to ICA, except that now we have K datasets, $\mathbf{x}^{[k]},\ k = 1, ..., K$ where each dataset is a linear mixture of N statistically independent sources. Under the assumption that PCA preprocessing results in $d = P$, the noiseless IVA model is given by $\mathbf{x}^{[k]} = \mathbf{A}^{[k]}\mathbf{s}^{[k]},\ k = 1, ..., K$, where $\mathbf{A}^{[k]} \in \mathbb{R}^{P \times P},\ k = 1, ..., K$ are invertible mixing matrices and $\mathbf{s}^{[k]} = [s_1^{[k]}, ..., s_P^{[k]}]^\top$ is the vector of latent sources for the kth dataset. In the IVA model, the components within each $\mathbf{s}^{[k]}$ are assumed to be independent, while at the same time, dependence across corresponding components of $\mathbf{s}^{[k]}$ is allowed. To mathematically formulate the dependence across components that IVA can take into account, we define the source component vector (SCV) by vertically concatenating the pth source from each of the K dataset as

$$\mathbf{s}_p = [s_p^{[1]}, ..., s_p^{[K]}]^\top, \tag{2}$$

where $\mathbf{s}_p$ is a K-dimensional random vector. The goal in IVA is to estimate K demixing matrices to yield source estimates $\mathbf{y}^{[k]} = \mathbf{W}^{[k]}\mathbf{x}^{[k]}$, such that each SCV is maximally independent of all other SCVs. It should be mentioned that dependence across datasets is not a necessary condition for IVA to work. In the case where this type of statistical property does not exist, IVA reduces to individual ICAs on each dataset.

The IVA cost function can be defined in a similar manner as ICA. However, the optimization parameter is not just a single demixing matrix $\mathbf{W}$ as in the ICA case, but a set of demixing matrices $\mathbf{W}^{[1]}, \ldots, \mathbf{W}^{[K]}$, which can be collected into a three dimensional array $\mathcal{W} \in \mathbb{R}^{P \times P \times K}$. The IVA objective function is given by

$$J_{IVA}(\mathcal{W}) = \sum_{p=1}^{P} H(\mathbf{y}_p) - \sum_{k=1}^{K} \log \left| \det \left(\mathbf{W}^{[k]} \right) \right| - H(\mathbf{x}^{[1]}, ..., \mathbf{x}^{[K]}), \tag{3}$$

where $H(\mathbf{y}_p)$ denotes the differential entropy of the estimated pth SCV and the term $H(\mathbf{x}^{[1]}, ..., \mathbf{x}^{[K]})$ is a constant parameter where it can be treated as a constant for optimization purposes. By definition, the term $H(\mathbf{y}_p)$ equals $\sum_{k=1}^{K} H(y_p^k) - I(\mathbf{y}_p)$, where $I(\mathbf{y}_p)$ denotes the mutual information within the pth SCV. Therefore, it can be observed that minimization with respect to each demixing matrix $\mathbf{W}^{[k]}$ of (3) automatically increases the mutual information within the components of a SCV, revealing how IVA exploits this type of statistical property. It can be seen that without the mutual information term, the objective function (3) is equivalent to performing independent ICA separately on each dataset. For more information about the derivation of the IVA cost function and the optimization scheme, we refer the reader to [14].

B. Feature Extraction and Regression Procedure

The regression process consists of three stages. In the first stage, N molecules are randomly sub sampled in order to generate training and testing datasets, $\mathbf{X}^k_{\text{Train}}$ and $\mathbf{X}^k_{\text{Test}}$ respectively, for each $k = 1, \ldots, K$.

In the second stage, the mean from each dataset is removed so all datasets are centered and PCA is applied to each $\mathbf{X}^k_{\text{Train}}$ using an order P. This implies that the signal subspace contains the components that have higher variance. Then for each $k \in \{1, \ldots, K\}$, we generate $\hat{\mathbf{X}}^k_{\text{Train}} \in \mathbb{R}^{P \times N_{\text{Train}}}$ and by vertically concatenating each $\hat{\mathbf{X}}^k_{\text{Train}}$ we form a three dimensional array $\hat{\mathbf{X}}_{\text{Train}} \in \mathbb{R}^{P \times N \times K}$. IVA is performed on $\hat{\mathbf{X}}_{\text{Train}}$ resulting in a set of demixing matrices, $\{\mathbf{W}^1, \ldots, \mathbf{W}^K\}$, where each $\mathbf{W}^k \in \mathbb{R}^{P \times P}$. Using the demixing matrices we generate $\mathbf{Y}^k_{\text{Train}} = \mathbf{W}^k(\hat{\mathbf{X}}^k_{\text{Train}})^\top$ for each $k = 1 \ldots, K$ and the training dataset $\mathbf{Y}_{\text{Train}}$ is formed by vertically concatenating the SCVs, $\mathbf{y}_1, \ldots, \mathbf{y}_P$. Note here that $\mathbf{Y}_{\text{Train}}$ contains all the extracted molecular signatures for each training molecule and will be used for training the regression model.

The testing dataset is generated by removing the mean from each testing dataset and by using the PCA training transformation from the training phase, we generate $\hat{\mathbf{X}}^k_{\text{Test}}$, $k = 1 \ldots, K$. Each of the demixing matrices from the training phase is used to create testing datasets $\mathbf{Y}^k_{\text{Test}} \in \mathbb{R}^{P \times N_{\text{Test}}}$ by $\mathbf{Y}^k_{\text{Test}} = \mathbf{W}^k(\hat{\mathbf{X}}^k_{\text{Test}})^\top$ for each $k = 1 \ldots, K$. Finally, the testing dataset is formed by vertically concatenating the testing SCVs. Similar to the training phase, $\mathbf{Y}_{\text{Test}}$ will contain all the molecular signatures for all testing molecules.

In the third stage, we train the regression model using $(\mathbf{Y}_{\text{Train}})^\top$. Here, the specific form of the regression function is unimportant. But to demonstrate a concrete example, we use kernel ridge regression (KRR) with a Gaussian kernel, which follows from previous work [3] and [1]. Once the regression model has been trained we evaluate its performance by using the unseen data $(\mathbf{Y}_{\text{Test}})^\top$. For all of the experiments, hyperparameter optimization and model training and testing is done using a nested cross validation scheme. In the inner loop, the length scale parameter of the Gaussian kernel and the regularization parameter are optimized using grid search selection using 80% of the data to train the model and another 10% for validation. The remaining 10% is held out as a test set to estimate performance after hyperparameter optimization. The outer loop is done five times, corresponding to five folds. This entire process was repeated 30 times (with shuffling before each iteration) to generate well-converged statistics.

C. Data Sources and Methods

While modern deep learning methods have demonstrated superior performance for the prediction of molecular properties [15], [16], they require a large amount of data in order to effectively work and due to their mathematical nature they lose the ability to provide interpretable results. In this work, we use a small and diverse dataset in order to demonstrate that our proposed method does not require many training points in order to achieve desirable performance while retaining the possibility of interpreting the generated features to discover relationships between molecular structures and properties. For more numerical experiments using different types of molecules and properties we refer the reader to [17].

Our energetics dataset consists of 109 energetic compounds with properties computed by Huang and Massa [18] and 261 energetics related compounds with properties computed by Mathieu [19]. The 109 molecules are from ten distinct compound classes and models trained on it should be relatively general in their applicability. Therefore, the diverse nature of those data points enable us to demonstrate that our proposed method is applicable to a wide range of candidate energetic molecules. Huang and Massa calculated properties using a combination of DFT calculations and Chapman-Jouguet theory. Mathieu also used Chapman Jouguet theory but used different simulation techniques. A few molecules appear in both datasets and the properties calculated by both authors differ slightly (both are included though in training).

The properties common to all 370 molecules include the density, detonation pressure, and detonation velocity. From the energetics data, we generate five observation matrices, $\mathbf{X}^1, \ldots, \mathbf{X}^5$, using five different featurization methods: Coulomb matrices eigenspectra (CME) [1], sum over bonds (SOB) [3], custom descriptors [3], functional groups counts, and Estate fingerprint [3]. The Coulomb matrix of a molecule is specified by the 3-dimensional coordinates of the atoms as well as their atomic charges. Since the Coulomb matrix is not invariant under row or column permutations, the eigenspectra are used as a feature vector. The dimensionality of the Coulomb matrix eigenspectra ($d = 87$) is set by the molecule with the largest number of atoms in the dataset, with smaller eigenspectra zero padded. Sum over bonds is defined as a bond count and the feature vectors are generated by first enumerating all of the bond types in the dataset and then counting how many of each bond are present in each molecule ($d = 24$). Similarly, the "functional group" featurization is a sum of different functional groups which can be counted using RDKit (the total number found in our dataset was $d = 55$). Since most molecules only contain a few different types of functional groups, this featurization is very sparse. By a "descriptor" we mean any function that maps a molecule to a scalar value. Descriptors may range from simple atom counts to complex quantum mechanical descriptors that describe electron charge distributions and other subtle properties. For our work we several descriptors based on stoichiometry which were deemed important for energetic materials including fractions of nitrogen, oxygen, and other elements and oxygen balance. For this type of featurization method we have ($d = 21$). Finally, the Estate fingerprint is based on electrotopological state indices [20], which encode information about associated functional group, graph topology, and the Kier-Hall electronegativity of each atom. For the Estate featurization method we have ($d = 32$).

III. RESULTS

The energetics dataset study examines (III-A) regression performance with respect to five featurization methods and (III-B) interpretability properties of IVA.

A. *Regression Analysis*

The first five rows in Table (I) show the MAEs for out-of-sample predictions by the KRR model for each type of featurization method. Errors are reported for the three properties associated with the energetics dataset. From (I), we see that SOB provides the best results. Overall, Estate performs the worst.

The last two rows of Table (I) list the MAEs when different techniques have been used to generate molecular signatures for training the regression model. The so-called Regular approach denotes the procedure where the SOB, Fun.group, Estate, Custom descriptors, and CME are vertically concatenated resulting in d feature vectors of dimension 198. The IVA approach shares the same feature extraction and regression procedures described in (II-B). For the IVA approach the resulting molecular signatures are of dimension 50. From Table (I), we observe that IVA leads to an improved MAE for all properties when compared to the "Regular" approach. This is likely due to IVA's ability to exploit the complementary information that is shared across them.

We next evaluate the performance of the IVA approach as a function of the number of datasets that are fused. Figure (1) displays the MAEs of the density, detonation pressure, and detonation velocity as a function of the number of featurization methods used to train the regression model. We chose the dimension of the vectors created by IVA to be $P = 10$ for each dataset. The MAEs in Figure (1) show the improvement in the regression error with an increasing number of fused datasets. For the sake of comparison, the errors associated with the Regular approach are also shown. When $K = 2$ the median of the Regular approach when predicting the density is lower than the IVA approach. However, IVA reduces the variation in the MAEs which reveals how it produces compact features. As expected, as the number of datasets is increased, IVA performs better than the Regular approach, due to the fact that IVA exploits complementary information among the different featurization methods.

B. *Knowledge Discovery*

An important property of IVA is that it can also provide chemical or physical interpretations through the estimated mixing matrix. Once the demixing matrices, $\mathbf{W}^{[k]}$ have been estimated using the approach in Section II-B, we can estimate the mixing matrix by performing back-reconstruction. In order to describe this procedure remember that before IVA is performed, PCA is applied to each $\mathbf{X}^k$ using a presumed order P. This provides the data reduction matrix $\mathbf{F}^k$ where each $\mathbf{F}^k$ is formed by the eigenvectors with the first P highest eigenvalues of the corresponding $\mathbf{X}^k$. An estimate of the mixing matrix $\mathbf{A}^k$ is computed as $\hat{\mathbf{A}}^k = (\mathbf{F}^k)^\dagger(\mathbf{W}^k)^{-1}$, where $(\cdot)^\dagger$ denotes the pseudo-inverse of a matrix.

TABLE I: Average MAE for different featurization methods and different approaches to generate molecular signatures for three target properties. The properties are density, detonation pressure, and detonation velocity.

	$\rho(g/cc)$	$P(GPa)$	$V_{det}(km/s)$
SOB	0.037	2.125	0.258
Fun.group	0.075	3.873	0.451
Estate	0.082	4.022	0.456
Custom Desc.	0.069	2.643	0.275
CME	0.061	2.982	0.381
Regular	0.034	1.919	0.205
IVA	**0.028**	**1.458**	**0.173**

Each row of the kth estimated mixing matrix therefore represents the weights for the estimated sources of the kth dataset. The values of the weights can reveal relationships between certain characteristics of a given set of molecules. To demonstrate this further we sort the magnitudes of the weights on each estimated column of mixing matrix $\hat{\mathbf{A}}^k$ and the corresponding features used in the analysis.

From Table (II), we see the most representative features that belong to each low dimensional molecular signature. Some of these rankings make some chemical sense - for instance the count of nitro groups is very important for energetic properties, and the $C : C$ bond (aromatic carbon bond) designates the presence of a benzene ring. Molecules with benzene rings are often planar which means they can often achieve higher crystalline density. In addition, it is interesting to observe that the functional groups and types of bonds that are heavily weighted indicate types of atomic groupings that are likely to appear in as-yet unknown energetic molecules. For example bonds involving oxygen and nitrogen are well known in a variety of energetic materials (e.g., nitrate esters, and nitro compounds), and the atomic groupings involving carbon and nitrogen are commonly found in energetic aromatic compounds. This may suggest that the weights, or some function of the weights, may constitute a new descriptor that could be used in the construction of new physically meaningful features for the fused data. This proof of concept feature ranking shows that our method remains interpretable, allowing for the illumination of potentially interesting structure-property relationships and also allowing for easier debugging of the model.

IV. DISCUSSION

The success of the proposed method raises several interesting questions that can be explored in future work. Depending on the nature of the data we can use or develop new algorithms that take different statistical properties into account such as sparsity [21]. In addition, we can also compare the regression performance with other IVA algorithms or other methods such

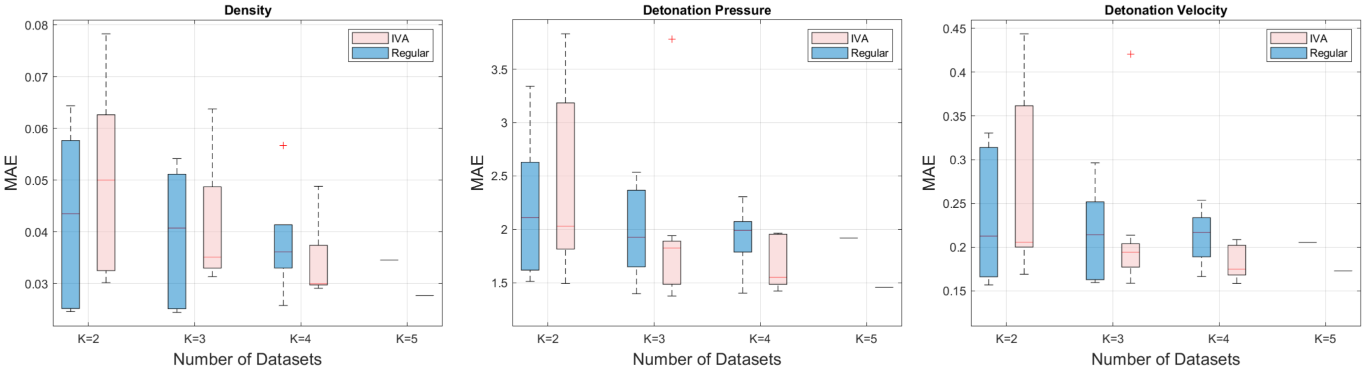

Fig. 1: Boxplots of the MAEs as a function of different number of datasets used in order to train the regression model. Boxplots provide statistics for the MAE when combining different featurization methods. For instance, when $K = 2$ we combine $\frac{5!}{2!3!}$ featurization methods yielding ten MAEs, similar for $K = 3$, and when $K = 4$ we combine $\frac{5!}{4!}$ featurization methods yielding five MAEs.

Signature Fun.group	Signature SOB	Signature Custom Desc.
Enamine	C:C	n_{NO}
Nitro	C-N	n_{CNO2}
benzene	N-O	OB_{100}
Azole	N=O	$n_{\mathrm{NNO_2}}$
pyro ring	C:N	n_{NOC}

TABLE II: The five most representative features for three extracted molecular signatures that correspond to different featurization methods. The order of the features is from the highest to the lowest magnitude.

as dictionary learning [22], [23] and non-negative matrix or tensor factorization [24]. Lastly, order selection techniques can be used to determine the order of the reduced IVA space.

REFERENCES

[1] M. Rupp, A. Tkatchenko, K.-R. Müller, and O. A. Von Lilienfeld, "Fast and accurate modeling of molecular atomization energies with machine learning," *Physical review letters*, vol. 108, no. 5, p. 058301, 2012.

[2] K. Hansen, F. Biegler, R. Ramakrishnan, W. Pronobis, O. A. Von Lilienfeld, K.-R. Müller, and A. Tkatchenko, "Machine learning predictions of molecular properties: Accurate many-body potentials and nonlocality in chemical space," *The journal of physical chemistry letters*, vol. 6, no. 12, pp. 2326–2331, 2015.

[3] D. C. Elton, Z. Boukouvalas, M. S. Butrico, M. D. Fuge, and P. W. Chung, "Applying machine learning techniques to predict the properties of energetic materials," *Scientific Reports*, vol. 8, no. 1, jun 2018. [Online]. Available: https://doi.org/10.1038/s41598-018-27344-x

[4] B. C. Barnes, D. C. Elton, Z. Boukouvalas, D. E. Taylor, W. D. Mattson, M. D. Fuge, and P. Chung, "Machine learning of energetic material properties," in *Proceedings of the 16th International Detonation Symposium, Cambridge, Maryland, USA*, 2018. [Online]. Available: https://arxiv.org/abs/1807.06156

[5] G. Montavon, M. Rupp, V. Gobre, A. Vazquez-Mayagoitia, K. Hansen, A. Tkatchenko, K.-R. Müller, and O. A. von Lilienfeld, "Machine learning of molecular electronic properties in chemical compound space," *New Journal of Physics*, vol. 15, no. 9, p. 095003, 2013. [Online]. Available: http://stacks.iop.org/1367-2630/15/i=9/a=095003

[6] D. Weininger, "Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules," *Journal of chemical information and computer sciences*, vol. 28, no. 1, pp. 31–36, 1988.

[7] B. Sanchez-Lengeling and A. Aspuru-Guzik, "Inverse molecular design using machine learning: Generative models for matter engineering," *Science*, vol. 361, no. 6400, pp. 360–365, 2018. [Online]. Available: http://science.sciencemag.org/content/361/6400/360

[8] D. L. Hall and J. Llinas, "An introduction to multisensor data fusion," *Proceedings of the IEEE*, vol. 85, no. 1, pp. 6–23, 1997.

[9] F. Castanedo, "A review of data fusion techniques," *The Scientific World Journal*, vol. 2013, 2013.

[10] P. Comon and C. Jutten, *Handbook of Blind Source Separation: Independent component analysis and applications*. Academic press, 2010.

[11] T. Adalı, M. Anderson, and G.-S. Fu, "Diversity in independent component and vector analyses: Identifiability, algorithms, and applications in medical imaging," *Signal Processing Magazine, IEEE*, vol. 31, no. 3, pp. 18–33, May 2014.

[12] A. J. Bell and T. J. Sejnowski, "The "independent components" of natural scenes are edge filters," *Vision research*, vol. 37, no. 23, pp. 3327–3338, 1997.

[13] A. Hyvärinen and E. Oja, "Independent component analysis: algorithms and applications," *Neural networks*, vol. 13, no. 4, pp. 411–430, 2000.

[14] Z. Boukouvalas, "Development of ICA and IVA Algorithms with Application to Medical Image Analysis," *ArXiv e-prints*, Jan. 2018.

[15] D. C. Elton, Z. Boukouvalas, M. D. Fuge, and P. W. Chung, "Deep learning for molecular design—a review of the state of the art," *Molecular Systems Design & Engineering*, vol. 4, no. 4, pp. 828–849, 2019.

[16] T. F. G. G. Cova and A. A. C. C. Pais, "Deep learning for deep chemistry: Optimizing the prediction of chemical patterns," *Frontiers in Chemistry*, vol. 7, p. 809, 2019. [Online]. Available: https://www.frontiersin.org/article/10.3389/fchem.2019.00809

[17] Z. Boukouvalas, D. C. Elton, P. W. Chung, and M. D. Fuge, "Independent vector analysis for data fusion prior to molecular property prediction with machine learning," *arXiv preprint arXiv:1811.00628*, 2018.

[18] L. Huang and L. Massa, "Applications of energetic materials by a theoretical method (discover energetic materials by a theoretical method)," *International Journal of Energetic Materials and Chemical Propulsion*, vol. 12, no. 3, 2013.

[19] D. Mathieu, "Sensitivity of energetic materials: Theoretical relationships to detonation performance and molecular structure," *Industrial & Engineering Chemistry Research*, vol. 56, no. 29, pp. 8191–8201, 2017.

[20] L. B. Kier and L. H. Hall, "An electrotopological-state index for atoms in molecules," *Pharmaceutical research*, vol. 7, no. 8, pp. 801–807, 1990.

[21] Z. Boukouvalas, Y. Levin-Schwartz, V. D. Calhoun, and T. Adalı, "Sparsity and independence: Balancing two objectives in optimization for source separation with application to fmri analysis," *Journal of the Franklin Institute*, vol. 355, no. 4, pp. 1873–1887, 2018.

[22] J. Mairal, F. Bach, J. Ponce, and G. Sapiro, "Online dictionary learning for sparse coding," in *Proceedings of the 26th annual international conference on machine learning*. ACM, 2009, pp. 689–696.

[23] I. Tosic and P. Frossard, "Dictionary learning," *IEEE Signal Processing Magazine*, vol. 28, no. 2, pp. 27–38, March 2011.

[24] A. Cichocki, R. Zdunek, A. H. Phan, and S.-i. Amari, *Nonnegative matrix and tensor factorizations: applications to exploratory multi-way data analysis and blind source separation*. John Wiley & Sons, 2009.