The Covariance of Topological Indices that Depend on the Degree of a Vertex

Boris Hollas

Theoretische Informatik, Universität Ulm, D-89081 Ulm
E-mail: hollas@informatik.uni-ulm.de

November 6, 2018

Abstract

We consider topological indices $\mathcal{I}$ that are sums of $f(\deg(u))f(\deg(v))$, where $\{u, v\}$ are adjacent vertices and f is a function. The Randić connectivity index or the Zagreb group index are examples for indices of this kind. In earlier work on topological indices that are sums of independent random variables, we identified the correlation between $\mathcal{I}$ and the edge set of the molecular graph as the main cause for correlated indices. We prove a necessary and sufficient condition for $\mathcal{I}$ having zero covariance with the edge set.

1 Introduction

For quite some time it has been known that topological indices (graph invariants on molecular graphs) exhibit considerable mutual correlation [1, 2]. This is a major problem when performing structure-activity studies as the employed statistical methods may fail or give little meaningful results on sets of correlated data. Also, strong correlations among a set of topological indices raise doubt whether these indices describe different and meaningful biological, chemical or physical properties of molecules.

In an attempt to investigate the reasons for these correlations, we used random graphs [3] as a model for chemical graphs and for topological indices of the form

$$\mathcal{I}_{\mathbf{X}}(G) = \frac{1}{2} \sum_{\{u,v\} \in E} X_u X_v$$

where E is the edge set of the molecular graph $G = (V, E)$ and $\{X_v \mid v \in V\}$ is a set of independent random variables with a common expectation $\boldsymbol{E}(X)$ [4, 5, 6]. We proved that $\mathcal{I}_{\mathbf{X}}$, $\mathcal{I}_{\mathbf{Y}}$, and $\mathcal{I}_{\mathbf{1}}$ are *linearly dependent* for *independent* vertex properties X, Y with $\boldsymbol{E}(X)\,, \boldsymbol{E}(Y) > 0$ as the number of vertices tends to

infinity. For $\boldsymbol{E}(X) = \boldsymbol{E}(Y) = 0$ however these indices are uncorrelated. Here, $\mathcal{I}_{\mathbf{1}}$ denotes a topological index with $X_v = 1$ for all $v \in V$.

While the random graph model we used in [6] encompasses graphs of arbitrary structure, including chemical graphs, the notion of vertex (or atom) properties X_v that are *independent* of the molecular graph is a serious abstraction from computational chemistry where atom properties used for topological indices are a *function* of the graph or even the molecule.

In this paper, we use a slightly more general random graph model than the one used in [4]. In particular, we consider graphs on n vertices whose edges are chosen independently with a probability proportional to $1/n$. The latter ensures that the expected number of edges increases linearly in the number of vertices. We use this to model an approximately linear relation of bonds to vertices present in molecules. For example, homologous series of aliphatic or aromatic hydrocarbons with n atoms contain $n + c$ bonds for some constant c. Polyphenols contain $\frac{7}{6}n + c$ bonds as each monomer adds 6 atoms and 7 bonds. On the other hand, there is some variation in the number of bonds for a given number of atoms in a heterogenous set of molecules, which is also true for the random graph model.

As a more significant difference we consider the vertex properties X_v to be a function of the vertex degree instead of being independent. Thus, our results are valid for important topological indices such as the Randić connectivity index or Zagreb group index. We will focus on the crucial covariance between $\mathcal{I}_{\mathbf{X}}$ and $\mathcal{I}_{\mathbf{1}}$.

2 Preliminaries

First, we describe the random graph model. For a graph (V, E) let

$$1_{uv} = 1_{\{\{u,v\}\in E\}} = \begin{cases} 1 & \text{if } \{u, v\} \in E \\ 0 & \text{else} \end{cases}$$

be the indicator function for $\{\{u, v\} \in E\}$. For $V = \{1, \dots, n\}$ let 1_{uv} $(u, v \in V)$ be independent random variables with $P(1_{uv} = 1) = p$. The space of random graphs $\mathscr{G}(n, p)$ can be identified with the distribution of $(1_{uv})_{u,v\in V}$. We set $p = \alpha/n$ for a fixed parameter $\alpha > 0$ so that $\boldsymbol{E}|E| = \binom{n}{2}p \sim \frac{\alpha}{2}n$ as motivated in the introduction.

To describe the vertex properties, let $f : \mathbb{N}_0 \to \mathbb{R}$ be a function with $f(0) = 0$. We consider the topological index

$$\mathcal{I}_{\mathbf{X}} = \mathcal{I}_{\mathbf{X}}(G) = \frac{1}{2} \sum_{\{u,v\}\in E(G)} X_u X_v \tag{2.1}$$

with

$$X_v = f(\deg(v))$$

being the vertex properties and $G \in \mathscr{G}(n, \alpha/n)$ is a random graph. Thus, $f(0) = 0$ accounts for isolated vertices being ignored. Using indicators this can be written as

$$\mathcal{I}_{\mathbf{X}} = \sum_{u<v} X_u X_v 1_{uv} \tag{2.2}$$

which is better suited to employ the expectation operator.

We us the following *notations* throughout the text:

$O(f)$	denotes	a function g with $g(x) \leq cf(x)$ for all large x and some constant $c > 0$
$X_n \xrightarrow{\mathscr{D}} X$	denotes	that random variable X_n converges to X in distribution
$a_n \nearrow a$	denotes	that sequence (a_n) is monotonically increasing and converges to a

3 Expectations and Covariance

To determine expectation values, we have to eliminate the dependence among X_u and X_v in (2.2). This is achieved by conditioning for $\{1_{uv} = 1\}$. If the edge $\{u, v\}$ exists then the degree of u has no effect on the degree of v and vice versa:

Lemma 1.
Suppose $u < v$. Then the random variables $(1_{uu'})_{u'>u}$ and $(1_{vv'})_{v'>v}$ are independent with respect to the probability measure $P(\cdot \mid 1_{uv} = 1)$. The same claim holds for $\deg(u)$ *and* $\deg(v)$.

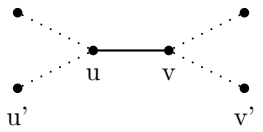

Figure 1: We fix edge $\{u, v\}$

Proof. Let $a_{uu'}, a_{vv'} \in \{0, 1\}$ for $u' > u$, $v' > v$ and $a_{uv} = 1$. We check that for $(a_{uu'})_{u'>u}$, $(a_{uu'})_{u'>u}$ holds

$$\begin{aligned}
&P((1_{uu'})_{u'>u} = (a_{uu'})_{u'>u} \wedge (1_{vv'})_{v'>v} = (a_{vv'})_{v'>v} \mid 1_{uv} = 1) \\
&\quad = \frac{P((1_{uu'})_{u'>u,u'\neq v} = (a_{uu'})_{u'>u} \wedge (1_{vv'})_{v'>v} = (a_{vv'})_{v'>v} \wedge 1_{uv} = 1)}{P(1_{uv} = 1)} \\
&\qquad = P((1_{uu'})_{u'>u,u'\neq v} = (a_{uu'})_{u'>u} \wedge (1_{vv'})_{v'>v} = (a_{vv'})_{v'>v}) \\
&\qquad = P((1_{uu'})_{u'>u,u'\neq v} = (a_{uu'})_{u'>u})\, P((1_{vv'})_{v'>v} = (a_{vv'})_{v'>v}) \\
&\qquad\qquad = P((1_{uu'})_{u'>u} = (a_{uu'})_{u'>u} \mid 1_{uv} = 1) \\
&\qquad\qquad\qquad \cdot P((1_{vv'})_{v'>v} = (a_{vv'})_{v'>v} \mid 1_{uv} = 1)
\end{aligned}$$

If $a_{uv} = 0$, both sides are zero. The second claim is a consequence of $\deg(u)$ or $\deg(v)$ being functions of $1_{uu'}$ or $1_{vv'}$, respectively. □

We are going to apply lemma 1 to conditional expectations. This motivates the definition

$$\delta_f^{(k)} = \boldsymbol{E}(X_1 \mid 1_{12}1_{13}\cdots 1_{1k+1} = 1)\,, \quad k > 0 \tag{3.1}$$

We shall see later why we also need $k > 1$. For symmetry reasons, this could as well be defined for a vertex $v \neq 1$ and any set of distinct vertices $\{u_2, \ldots, u_k\}$ different from v. As we shall see in section 4, $\lim_{n\to\infty} \delta_f^{(k)}$ exists and is a function of α if f satisfies a condition. Thus, we may regard $\delta_f^{(k)}$ as almost constant for large n.

Lemma 2.
$\boldsymbol{E}(\mathcal{I}_{\mathbf{X}}) = \left(\delta_f^{(1)}\right)^2 \boldsymbol{E}|E|$

Proof.

$$\begin{aligned}
\boldsymbol{E}(\mathcal{I}_{\mathbf{X}}) &= \sum_{u<v} \boldsymbol{E}(X_u X_v \mid 1_{uv} = 1)\, p && \text{by (2.2)} \\
&= \sum_{u<v} \boldsymbol{E}(X_u \mid 1_{uv} = 1)\, \boldsymbol{E}(X_v \mid 1_{uv} = 1)\, p && \text{by lemma 1} \\
&= \left(\delta_f^{(1)}\right)^2 \boldsymbol{E}|E| && \text{by (3.1)}
\end{aligned}$$

□

Lemma 3.

$$\boldsymbol{E}(\mathcal{I}_{\mathbf{X}}\mathcal{I}_{\mathbf{1}}) = \left[\left(\delta_f^{(1)}\right)^2 \binom{n-2}{2} p + 2\delta_f^{(1)}\delta_f^{(2)}(n-2)p + \left(\delta_f^{(1)}\right)^2\right] \boldsymbol{E}|E|$$

Proof. To dissect the sum

$$\boldsymbol{E}(\mathcal{I}_{\mathbf{X}}\mathcal{I}_{\mathbf{1}}) = \sum_{u<v} \sum_{u'<v'} \boldsymbol{E}(X_u X_v 1_{uv} 1_{u'v'})$$

according to $|\{u, v\} \cap \{u', v'\}|$, consider

$$S_k = \{(u, v, u', v') \mid u < v \wedge u' < v' \wedge |\{u, v\} \cap \{u', v'\}| = k\}, \quad 0 \leq k \leq 2$$

Then

$$|S_0| = \binom{n}{2}\binom{n-2}{2} \tag{3.2}$$

$$|S_1| = 6\binom{n}{3} \tag{3.3}$$

$$|S_2| = \binom{n}{2} \tag{3.4}$$

(3.2) and (3.4) are obvious. To verify (3.3) let $(u, v, u', v') \in S_1$. Exactly two numbers are equal as indicated in figure 2. Cases (a), (b) allow just one way to distribute three distinct numbers on u, v, u', v' while there are two ways for cases (c), (d). For symmetry reasons, $\boldsymbol{E}(X_u X_v 1_{uv} 1_{u'v'}) = \boldsymbol{E}(X_1 X_2 1_{12} 1_{13})$ for

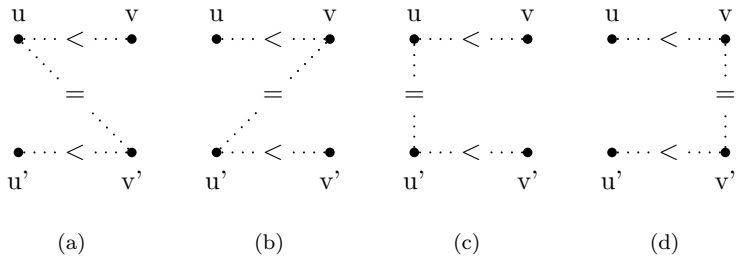

Figure 2: Possibilities for $(u, v, u', v') \in S_1$

all $(u, v, u', v') \in S_1$. Hence, we get

$$\begin{aligned}
\boldsymbol{E}(\mathcal{I}_{\mathbf{X}}\mathcal{I}_{\mathbf{1}}) &= |S_0|\boldsymbol{E}(X_1X_2 1_{12}1_{34}) \\
&\quad + |S_1|\boldsymbol{E}(X_1X_2 1_{12}1_{13}) \\
&\quad + |S_2|\boldsymbol{E}\big(X_1X_2 1_{12}^2\big) \\
&= |S_0|\boldsymbol{E}(X_1X_2 \mid 1_{12}=1)\, p^2 \\
&\quad + |S_1|\boldsymbol{E}(X_1X_2 1_{13} \mid 1_{12}=1)\, p \\
&\quad + |S_2|\boldsymbol{E}(X_1X_2 \mid 1_{12}=1)\, p \\
&= |S_0|\boldsymbol{E}(X_1 \mid 1_{12}=1)\, \boldsymbol{E}(X_2 \mid 1_{12}=1)\, p^2 \\
&\quad + |S_1|\boldsymbol{E}(X_1 1_{13} \mid 1_{12}=1)\, \boldsymbol{E}(X_2 \mid 1_{12}=1)\, p \\
&\quad + |S_2|\boldsymbol{E}(X_1 \mid 1_{12}=1)\, \boldsymbol{E}(X_2 \mid 1_{12}=1)\, p
\end{aligned} \tag{3.5}$$

by lemma (1). With

$$\boldsymbol{E}(X_1 1_{13} \mid 1_{12}=1) = 1/p\boldsymbol{E}(X_1 1_{12} 1_{13}) = \delta_f^{(2)} p$$

$$\binom{n}{3} = \binom{n}{2}\frac{n-2}{3}$$

and (3.2)-(3.4), (3.5), we have

$$\begin{aligned}
\boldsymbol{E}(\mathcal{I}_{\mathbf{X}}\mathcal{I}_{\mathbf{1}}) &= \left(\delta_f^{(1)}\right)^2 \boldsymbol{E}|E| \binom{n-2}{2} p \\
&\quad + 2\delta_f^{(1)}\delta_f^{(2)}\boldsymbol{E}|E|(n-2)p \\
&\quad + \left(\delta_f^{(1)}\right)^2 \boldsymbol{E}|E|
\end{aligned}$$

□

Remark. With $f \equiv 1$, lemma 2 and the help of Mathematica follows $Var(\mathcal{I}_{\mathbf{1}}) = \boldsymbol{E}|E|(1-p)$, as it should be.

We combine the results of this section in

Theorem 4.
If $\delta_f^{(1)}$, $\delta_f^{(2)}$ are bounded in n then

$$Cov(\mathcal{I}_{\mathbf{X}}, \mathcal{I}_{\mathbf{1}}) = \begin{cases} 0 & \text{if } \delta_f^{(1)} = 0 \\ \left[\left(\delta_f^{(1)}\right)^2 \left(1 + 2\alpha\left(\frac{\delta_f^{(2)}}{\delta_f^{(1)}} - 1\right)\right) + O\left(\frac{1}{n}\right)\right] \boldsymbol{E}|E| & \text{else} \end{cases}$$

Proof. By lemma 2 and lemma 3,

$$Cov(\mathcal{I}_{\mathbf{X}}, \mathcal{I}_{\mathbf{1}}) = \Bigg[\left(\delta_f^{(1)}\right)^2 \binom{n-2}{2} p + 2\delta_f^{(1)}\delta_f^{(2)}(n-2)p + \left(\delta_f^{(1)}\right)^2 - \left(\delta_f^{(1)}\right)^2 \boldsymbol{E}|E|\Bigg] \boldsymbol{E}|E|$$

Using $\binom{n-2}{2} - \binom{n}{2} = 3 - 2n$, this can be written as

$$Cov(\mathcal{I}_{\mathbf{X}}, \mathcal{I}_{\mathbf{1}}) = \left[\left(\delta_f^{(1)}\right)^2 (1 + (3-2n)p) + 2\delta_f^{(1)}\delta_f^{(2)}(n-2)p\right] \boldsymbol{E}|E|$$

$$= \begin{cases} 0 & \text{if } \delta_f^{(1)} = 0 \\ \left[\left(\delta_f^{(1)}\right)^2 \left(1 + 2p\left(\frac{\delta_f^{(2)}}{\delta_f^{(1)}} n - n\right)\right) + \left(\delta_f^{(1)}\right)^2 p \left(3 - 2\frac{\delta_f^{(2)}}{\delta_f^{(1)}}\right)\right] & \\ \qquad \cdot \boldsymbol{E}|E| & \text{else} \end{cases}$$

The assertion follows with $p = \alpha/n$. □

Remark. We will prove in theorem 5 in section 4 that all $\delta_f^{(k)}$ are in fact bounded in n if $f \in O(x)$.

Yet, it is not clear whether $Cov(\mathcal{I}_{\mathbf{X}}, \mathcal{I}_{\mathbf{1}}) \neq 0$ for $\delta_f^{(1)} \neq 0$. This is dealt with in the next section.

4 The Poisson Distribution and $\delta_f^{(k)}$

For the proof of the following theorem recall that for random variables X_n, X holds

$$X_n \xrightarrow{\mathscr{D}} X$$

iff

$$\boldsymbol{E}(f(X_n)) \to \boldsymbol{E}(f(X))$$

for all bounded and continuous functions $f : \mathbb{R} \to \mathbb{R}$. This does not hold for arbitrary unbounded functions f. Therefore, we require that $f \in O(x)$ in this section. While this does not seem to be the most general restriction, it facilitates the following elaborations.

Theorem 5.
For all $f \in O(x)$ and all k holds

$$\lim_{n\to\infty} \delta_f^{(k)} = \boldsymbol{E}(f(k+\mathscr{P}_\alpha)) = \sum_{j=0}^{\infty} f(k+j)\frac{\alpha^j}{j!}e^{-\alpha}$$

where $\mathscr{P}_\alpha$ is the Poisson distribution with parameter α.

Proof. By definition (3.1),

$$\begin{aligned}\delta_f^{(k)} &= \boldsymbol{E}(f(\deg(1)) \mid 1_{12}1_{13}\cdots 1_{1k+1} = 1)\\ &= \boldsymbol{E}\left(f\left(k + \sum_{j=k+2}^{n} 1_{1j}\right)\right) \qquad (4.1)\end{aligned}$$

Since $p = \alpha/n$, Poisson's limit theorem gives

$$\sum_{j=k+2}^{n} 1_{1j} \xrightarrow{\mathscr{D}} \mathscr{P}_\alpha \quad (n\to\infty)$$

The function $f : \mathbb{N}_0 \to \mathbb{R}$ can be extended to a continuous function $f : \mathbb{R} \to \mathbb{R}$ in an arbitrary way. Hence, the continuity theorem gives

$$f\left(k + \sum_{j=k+2}^{n} 1_{1j}\right) \xrightarrow{\mathscr{D}} f(k+\mathscr{P}_\alpha) \quad (n\to\infty)$$

For all bounded and continuous functions f^* follows by (4.1)

$$\delta_{f^*}^{(k)} \to \boldsymbol{E}(f^*(k+\mathscr{P}_\alpha)) \quad (n\to\infty) \qquad (4.2)$$

If f is also bounded the claim follows. To prove (4.2) for unbounded f we cut f off above a limit to divide f into a bounded and an unbounded part. We show that the unbounded part tends to zero as the limit tends to infinity.

To begin with, let $|f(x)| \le x$ for all x and let f be unbounded. Then there is a sequence of integers (m_l) such that without loss of generality$f(m_l) \nearrow \infty$ for $l \to \infty$ and $f(m_l) > 0$ for all l. Let be

$$c_m(x) = \begin{cases} x & \text{if } |x| < m \\ 0 & \text{else} \end{cases}$$

and

$$\tilde{c}_m(x) = \begin{cases} 0 & \text{if } |x| < m \\ x & \text{else} \end{cases}$$

Let $S_n := k + \sum_{j=k+2}^{n} 1_{1j}$. Then

$$\begin{aligned}|\boldsymbol{E}((\tilde{c}_{m_l} \circ f)(S_n))\,| &= \left|\boldsymbol{E}\big(f(S_n)1_{\{f(S_n)>m_l\}}\big)\right| \\ &\le \boldsymbol{E}\big(S_n 1_{\{f(S_n)>f(m_l)\}}\big)\end{aligned}$$

since $0 \leq f(m_l) \leq m_l$

$$= \boldsymbol{E}\big(S_n 1_{\{S_n > m_l\}}\big)$$

since $f(m_l)$ increases monotonically

$$\begin{aligned} &< \boldsymbol{E}\left(S_n \frac{S_n}{m_l}\right) \\ &= \frac{1}{m_l}\left[Var(S_n) + (\boldsymbol{E}(S_n))^2\right] \\ &= \frac{1}{m_l}\left[O(n)\, p(1-p) + (O(n)\, p)^2\right] \\ &= O(1/m_l) \end{aligned}$$

By linearity of expectation follows

$$\boldsymbol{E}((\tilde{c}_{m_l} \circ f)(S_n)) = O(1/m_l) \tag{4.3}$$

for all $f \in O(x)$. Thus,

$$\lim_{n\to\infty} \delta_f^{(k)} = \lim_{l\to\infty} \lim_{n\to\infty} \boldsymbol{E}(f(S_n))$$

by (4.1)

$$\begin{aligned} &= \lim_{l\to\infty} \lim_{n\to\infty} [\boldsymbol{E}((c_{m_l} \circ f)(S_n)) + \boldsymbol{E}((\tilde{c}_{m_l} \circ f)(S_n))] \\ &= \lim_{l\to\infty} [\boldsymbol{E}((c_{m_l} \circ f)(k + \mathscr{P}_\alpha)) + O(1/m_l)] \end{aligned}$$

by (4.2) and (4.3)

$$= \boldsymbol{E}(f(k + \mathscr{P}_\alpha))$$

by the convergence theorem of Lebesgue. □

With the help of theorem 5 we are able to answer the question raised at the end of section 3:

Theorem 6.
For $n \to \infty$ and $f \in O(x)$ holds: $\mathcal{I}_{\mathbf{X}}$ and $\mathcal{I}_{\mathbf{1}}$ have covariance zero if and only if $\lim_{n\to\infty} \delta_f^{(1)} = 0$.

Proof. Assume that $\lim_{n\to\infty} \delta_f^{(1)} \neq 0$ and $\lim_{n\to\infty} Cov(\mathcal{I}_{\mathbf{X}}, \mathcal{I}_{\mathbf{1}}) = 0$. By theorem 4 follows

$$\lim_{n\to\infty} \frac{\delta_f^{(2)}}{\delta_f^{(1)}} = 1 - \frac{2}{\alpha}$$

With theorem 5 we get

$$\sum_{j=0}^{\infty} f(2+j)\frac{\alpha^j}{j!} = \left(1-\frac{1}{2\alpha}\right)\sum_{j=0}^{\infty} f(1+j)\frac{\alpha^j}{j!}$$
$$= \sum_{j=0}^{\infty} f(1+j)\frac{\alpha^j}{j!} - \frac{1}{2}\sum_{j=0}^{\infty} f(1+j)\frac{\alpha^{j-1}}{j!}$$

We multiply by α and substitute j with $j-1$ in the first two series to get

$$\sum_{j=1}^{\infty} f(1+j)\frac{\alpha^j}{(j-1)!} = \sum_{j=1}^{\infty} f(j)\frac{\alpha^j}{(j-1)!} - \frac{1}{2}\sum_{j=0}^{\infty} f(1+j)\frac{\alpha^j}{j!}$$

Hence,

$$\frac{1}{2}f(1)\alpha^0 + \sum_{j=1}^{\infty}\left[f(1+j)\left(1+\frac{1}{2j}\right) - f(j)\right]\frac{\alpha^j}{(j-1)!} = 0$$

By theorem 5, this series converges for all $\alpha > 0$. By the identity theorem for power series follows that all coefficients are zero. By induction thus follows $f \equiv 0$, which contradicts $\lim_{n\to\infty} \delta_f^{(1)} \neq 0$.

The opposite direction follows by theorem 4 and theorem 5. □

5 Discussion

We have seen that $\delta_f^{(1)}$ is an important quantity for the covariance of the topological indices we consider. Theorem 5 shows that $\delta_f^{(k)}$ does not depend on n for large n. This justifies definition (3.1) since we do not want $\delta_f^{(k)}$ to be very different for graphs of different size. Also, theorem 5 provides a way to approximately compute $\delta_f^{(k)}$. If we substitute X_v by $X_v - \delta_f^{(1)}$ in (2.1), the resulting index is uncorrelated to $\mathcal{I}_\mathbf{1}$.

As a drawback, we require $f \in O(x)$ in section 4. Theorem 5 may not be valid if f increases very steeply. However, it should be possible to derive an upper limit similar to (4.3) for functions f with a higher rate of growth than $O(x)$.

In [6], we proved that topological indices (with independent vertex properties) are necessarily correlated if the vertex properties have expectations not equal to zero. Theorem 6 does not give this result as it is an assertion on covariance only. The next step will therefore be an examination of correlations within this setting.

References

[1] I. Motoc, A. Balaban, O. Mekenyan, and D. Bonchev. Topological indices: Inter-relations and composition. *MATCH - Commun. Math. Comput. Chem.*, 13:369–404, 1982.

[2] S. Basak, V. Magnuson, G. Niemi, R. Regal, and G. Veith. Topological indices: Their nature, mutual relatedness, and applications. *Math. Model.*, 8:300–305, 1987.

[3] B. Bollobás. *Modern Graph Theory*. Springer, 1998.

[4] B. Hollas. Correlation properties of the autocorrelation descriptor for molecules. *MATCH - Commun. Math. Comput. Chem.*, 45:27–33, 2002.

[5] B. Hollas. An analysis of the autocorrelation descriptor for molecules. *J. Math. Chem.*, 33(2):91–101, 2003.

[6] B. Hollas. Correlations in distance-based descriptors. *MATCH - Commun. Math. Comput. Chem.*, 47:79–86, 2003.