Chem Soc Rev

REVIEW ARTICLE

View Article Online
View Journal

Cite this: DOI: 10.1039/c5cs00227c

Facts and fictions about polymorphism†

Aurora J. Cruz-Cabeza,*[a] Susan M. Reutzel-Edens[b] and Joel Bernstein[cd]

Received 13th March 2015

DOI: 10.1039/c5cs00227c

www.rsc.org/chemsocrev

We present new facts about polymorphism based on (i) crystallographic data from the Cambridge Structural Database (CSD, a database built over 50 years of community effort), (ii) 229 solid form screens conducted at Hoffmann-La Roche and Eli Lilly and Company over the course of 8+ and 15+ years respectively and (iii) a dataset of 446 polymorphic crystals with energies and properties computed with modern DFT-d methods. We found that molecular flexibility or size has no correlation with the ability of a compound to be polymorphic. Chiral molecules, however, were found to be less prone to polymorphism than their achiral counterparts and compounds able to hydrogen bond exhibit only a slightly higher propensity to polymorphism than those which do not. Whilst the energy difference between polymorphs is usually less than 1 kcal mol^{-1}, conformational polymorphs are capable of differing by larger values (up to 2.5 kcal mol^{-1} in our dataset). As overall statistics, we found that one in three compounds in the CSD are polymorphic whilst at least one in two compounds from the Roche and Lilly set display polymorphism with a higher estimate of up to three in four when compounds are screened intensively. Whilst the statistics provide some guidance of expectations, each compound constitutes a new challenge and prediction and realization of targeted polymorphism still remains a holy grail of materials sciences.

1. Introduction

"Polymorphism has been mainly studied in its phenomenological aspects, while its structural and energetic aspects have been alluded to in diverse fields of research, but in spite of a large body of data, have never been considered in a systematic way... The control of crystal polymorphism has practical advantages in many branches of the chemical industry, in fact, all those which deal with the organic solid state."[1]

Since 1991, the phenomenon of polymorphism – the ability of a compound to crystallize in more than one crystal structure – has been the subject of growing interest (Fig. 1). A literature search on the topic renders over eleven thousand scientific publications in WebOfScience[2] and over six thousand patent documents worldwide from 1966–2013 in Espacenet[3] (Fig. 1). Although some key contributions to the subject were made in the late 60s and 70s, a significant communal interest in the subject did not occur until the early 90s (Fig. 1). The phenomenon of polymorphism, however, was already recognized almost 200 years ago,[4] and it has a somewhat turbulent history.

The first example of a polymorphic organic compound was benzamide, identified and studied by Liebig and Wohler in 1832.[5] Although the crystal structure of the stable form was determined as early as 1959,[6] a labile form was discovered in 2005[7] whilst the original metastable form resisted solution until 2007.[8,9]

The century following the Wohler–Liebig discovery witnessed considerable activity in the study of polymorphism.[10–16] For instance, the first issue of *Zeitschrift fur Kristallographie*, founded in 1877 by the legendary P. von Groth,[12] contained a paper by his student, Otto Lehmann, with a diagram of a hot stage microscope and a "time *versus* temperature" curve, clearly indicating the four polymorphs of ammonium nitrate.[17] Although the subject was not a molecular crystal, the study is a classic example of the recording of thermal events associated with transitions between polymorphic crystal forms.

Interest in structural polymorphism declined during the early decades of the development of structural crystallography and, although many instances of polymorphism had been documented based on thermal[11] and optical[14] data, in most cases their structural characterization awaited developments in rapid single crystal structure determination. That gap has been closed to a significant extent, especially for "classic" (*i.e.* iconic) molecules (*e.g.* benzene, benzamide, *etc.*). It is possible – indeed, not unlikely – that for at least some of those molecules the polymorphic landscape has not been fully mapped since

[a] *Roche Pharma Research and Early Development, Therapeutic Modalities, Roche Innovation Center Basel, Basel, Switzerland. E-mail: aurorajosecruz@gmail.com*
[b] *Small Molecule Design & Development, Eli Lilly and Company, Lilly Corporate Center, Indianapolis, IN 46285, USA*
[c] *Faculty of Natural Sciences, New York University Abu Dhabi, P.O. Box 129188, Abu Dhabi, United Arab Emirates*
[d] *Department of Chemistry, Ben-Gurion University of the Negev, P.O. Box 653, Beer Sheva 84105, Israel*

† In memory of our friend and mentor, the late Frank H. Allen, and in honor of the 50 years anniversary of the Cambridge Crystallographic Data Centre.

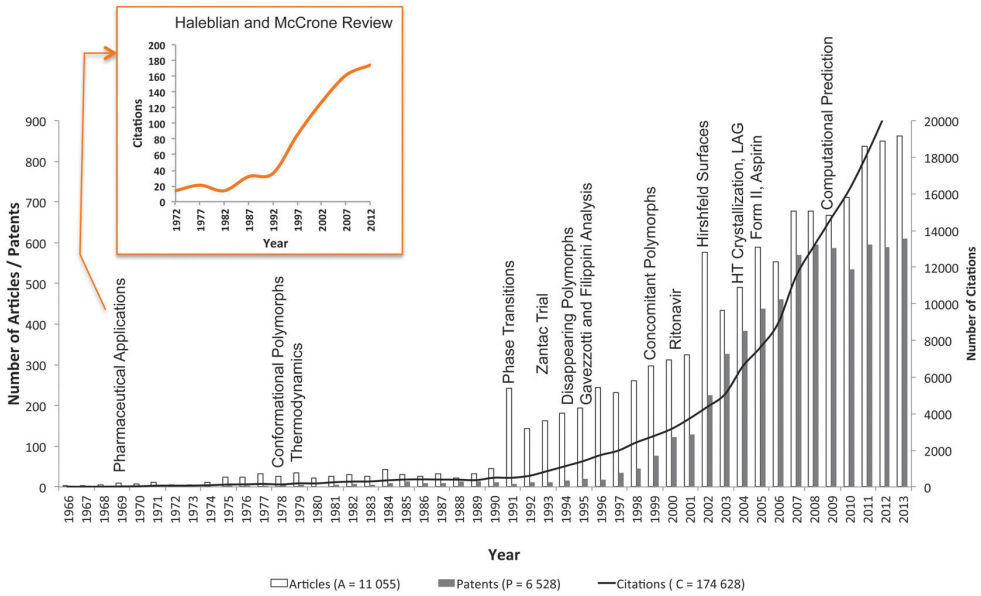

Fig. 1 Number of publications, citations to those and patents related to polymorphism. Landmark contributions are indicated and commented further in the text. Inner graph corresponds to the citations history of the McCrone & Haleblian, *J. Pharm. Sci.*, 1969 review (669 citations in Google Scholar, five-year bins).

sufficient "time and money"[18] have not been invested into exploring their polymorphs.

Polymorphism, after all, has not always been a sought phenomenon and it has been overlooked in the past, especially in the early days of structural crystallography. There are a number of reasons for this relative neglect. First, for many years carrying out a crystal structure determination was a major task; hence, the time and effort involved in solving the crystal structure of another crystal form of the same molecule often was not justified. Furthermore, prior to the early 1970's determination of non-centrosymmetric crystal structures, structures with $Z' > 1$, structures with disorder or even those that did not grow into "good" single crystals, presented major challenges for the crystallographer and were often abandoned. This is most likely

From left to right: Aurora J. Cruz-Cabeza, Susan M. Reutzel-Edens and Joel Bernstein

Aurora J. Cruz-Cabeza is a Postdoctoral Fellow at F. Hoffmann-La Roche Ltd in Basel (Switzerland). After her BSc degree in chemistry from the University of Jaén (2002), she earned a Masters degree in Heterogeneous Catalysis from the University of Córdoba (2004) and a PhD in Physical Chemistry from the University of Cambridge (2008). Aurora has worked as a researcher in several pharmaceutical companies (Pfizer and Roche), the University of Amsterdam and the Cambridge Crystallographic Data Centre.

Susan M. Reutzel-Edens is a Senior Research Advisor in Small Molecule Design & Development at Eli Lilly and Company. She obtained her BS degree in chemistry from Winona State University (1987), then earned her PhD in organic chemistry at the University of Minnesota (1991). Susan brought her experience in hydrogen-bond directed co-crystallization and interest in crystal polymorphism to Eli Lilly, where she developed Lilly's solid form design program and for two decades led a team of cross-functional scientists charged with finding commercially-viable crystalline forms for small-molecule drug products.

Joel Bernstein obtained a BA degree at Cornell University and a PhD in physical chemistry at Yale University. Following postdoctoral stints in X-ray crystallography at UCLA and the Weizmann Institute of Science, he joined the faculty of Ben-Gurion University where he was the incumbent of the Carol and Barry Kaye Professorship of Applied Science until 2010 and is now Professor Emeritus. Currently, Joel is a professor at the newly founded New York University Abu Dhabi. He has published over 180 research and review articles and is the sole author of a book entitled "Polymorphism in Molecular Crystals" (Oxford University Press).

the reason that the structures of the four polymorphs of the relatively simple molecule benzidine (with $Z' = 4.5$, 3.0. 1.5 and 4.5 respectively) remained unreported until well into the 21st century.[19] Another important factor for the decline in interest in polymorphism was the advance of other analytical methods that readily provided increasingly precise and reproducible data for characterizing and defining compounds and materials. During that period numerical data became the mode for defining materials. That is still very much the case, but the relative ease and decreasing cost in time and money of carrying out crystal structure determinations, combined with the facility of publishing digital color photos of crystals and crystal structures in the chemical literature has considerably aided in nurturing the renaissance of interest and activity in polymorphism and its manifestations.

In spite of the lack of widespread activity in polymorphism research during the middle decades of the 20th century, there were two active groups that made important contributions in that period. One was the group led by Ludwig Kofler at the University of Innsbruck (followed successively by Marie Kuhnert-Brandstatter and Artur Burger) and Walter McCrone, originally at Cornell, and later as an independent consultant. Both published books in the 1950's with major emphasis on the polymorphism of organic materials.[20,21] A 1980 translation of the Kofler book by Walter C. McCrone is available from McCrone Associates, Inc. McCrone's 1965 chapter on polymorphism[18] remains one of the classic papers on the subject together with his seminal 1969 *Journal of Pharmaceutical Sciences* review with Haleblian,[22] the first specific review publication relating polymorphism to the pharmaceutical industry, see Fig. 1. The citation history of that publication is illustrative of the evolving interest in polymorphism catalyzed to a great extent by the pharmaceutical industry. From the inner graph in Fig. 1 it can be seen that following an initial rise in citations of the 1969 McCrone/Haleblian paper during the 1970's, a pattern normally expected for a review, interest apparently waned until a renewal marked by a fairly steep rise in the number of citations starting around 1995. This rapid rise in interest in the McCrone/Haleblian paper is likely related to the high profile 1991 patent litigation on ranitidine hydrochloride (Zantac®)[23,24] which at the time was the world's largest selling drug ($3.45 billion year^{-1} – nearly twice as much as the next largest selling drug) and dealt to a large extent with various issues surrounding the polymorphism of the active ingredient. In support of this contention there was a parallel increase in scientific publications dealing with polymorphism, echoed by an increase in the number of patents issued containing "polymorph" in the title or abstract after 1991 (Fig. 1). Interestingly, there is also a spike in the number of publications in 2002, following the 1998–1999 saga of the removal and subsequent relaunch of Abbott's reformulated drug ritonavir due to the formation of an undesirable new polymorph.[25,26]

As interest in polymorphism has increased, many aspects of the subject have been addressed either directly or in passing. Some representative (but by no means comprehensive) scientific contributions towards our understanding of polymorphism (Fig. 1) after the Haleblian and McCrone review include:

(i) reports on conformational (1978),[27] disappearing (1995)[28] and concomitant polymorphs (1999),[29]

(ii) contributions towards our fundamental understanding of the thermodynamics (1979)[30,31] and phase transitions (1991)[32] in polymorphs,

(iii) the structural and energetic study of polymorphs under room temperature conditions by Gavezzotti and Filippini (1995),[1]

(iv) studies on polymorphism in the context of crystal engineering,[33–35]

(v) numerous studies on polymorphism in the context of pharmaceutical materials[7,8,12–15,36,37] including studies of landmark polymorphic drug systems such as ritonavir[25,26] and aspirin,[38–41]

(vi) applications of Hirshfeld surfaces[42] and computational chemistry[43] to the study of polymorphism, and

(vii) the development of new methods for surveying the crystal forms landscape, among them automation for high-throughput crystallization,[37] the liquid assisted grinding technique[44] or crystallizations in the presence of polymers.[45]

In spite of an impressive array of contributions across a broad spectrum of their chemical and physical aspects, polymorphic systems are in many ways still enigmatic, echoing the 1937 observation by Buerger and Bloom "with the accumulation of data, there is developing a gradual realization of the generality of polymorphic behavior, but to many chemists polymorphism is still a strange and unusual phenomenon."[46]

This contribution presents a systematic study of polymorphism from diverse sources. The first of these is based on the Cambridge Structural Database (CSD). We analyze the data and attempt to correct for certain biases in order to extract meaningful statistics. We also compute energetics for 215 polymorphic families with modern DFT-d techniques. These structure-based statistics are then compared to experimental polymorph screening statistics from 229 studies conducted at F. Hoffmann-La Roche Ltd (hereafter Roche) and Eli Lilly and Company (hereafter Lilly) over more than nine and fifteen years respectively. In this article, we address several fundamental aspects of the phenomenon and we question previous assertions promoted in the literature, many based on chemical intuition rather than scientific evidence. These lead to the correction of some common misconceptions that have been perpetuated in the polymorphism literature and suggest that the facts about polymorphism lie beyond chemical intuition and predictability.

2. Datasets and methods

2.1 Datasets derived from the Cambridge Structural Database (CSD)

2.1.1 Retrieval of the polymorphic datasets. Crystallographic data were retrieved from the CSD *vs.* 5.33 (Nov 2011) using Conquest.[47] The structure searches were restricted to organic molecules containing only the most common atomic elements (C, H/D, N, O, S and halogens). Crystal structures with all atomic coordinates determined (with the exception of hydrogen atoms) were retrieved and no polymeric structures were allowed. Only crystal structures containing the keyword

"polymorph" (*i.e.* the compound was described in the literature as being polymorphic) were kept.

In the CSD, a REFCODE consists of a six-letter code followed by two numbers. A REFCODE family (the 6 letter code) contains all determined crystal structures for a given compound (including polymorphic crystals and structure redeterminations). For the initial statistics of the CSD, we worked with REFCODE families (the six letter code). A REFCODE family corresponds to a unique composition (*e.g.* a unique compound or a unique mixture of compounds in a particular stoichiometry).

Three different polymorphic datasets were constructed:

• Polymorphic dataset of neutral single components (POL): REFCODE families of single component crystal structures – 2048 polymorphic families.

• Polymorphic dataset of neutral multicomponents (MULTI-POL): REFCODE families containing at least 2 or more components, all in neutral form – 303 polymorphic families.

• Polymorphic dataset of salts (SALTS-POL): REFCODE families containing at least two ionised components – 347 polymorphic families.

2.1.2 Retrieval of the monomorphic datasets. In building the monomorphic datasets, the same search criteria used for the polymorphic searches were applied to the entire CSD *vs.* 5.33 (Nov 2011). REFCODE families belonging to the polymorphic sets were then removed.

Three different monomorphic datasets were constructed:

• Monomorphic dataset of neutral single components (MONO): REFCODE families of single component crystal structures – 105 601 monomorphic families.

• Monomorphic dataset of neutral multicomponents (MULTI-MONO): REFCODE families containing at least 2 or more components, all in neutral form – 21 622 monomorphic families.

• Monomorphic dataset of salts (SALTS-MONO): REFCODE families containing at least two ionised components – 16 285 monomorphic families.

2.1.3 Molecular geometries and descriptors. Molecular geometries were retrieved from the crystal structures and exported as molecular files using Conquest.[47] OpenBabel was used for molecular format conversions and the addition of hydrogen atoms[48] to molecules with unresolved hydrogen atom positions.

Molecular descriptors were calculated using the ChemAxon cheminformatics plugin.[49] Properties such as number of atoms, molecular weight (M_w), number of asymmetric centers, number of aliphatic/aromatic rings or number of hydrogen bond donors and acceptors were calculated. In addition, we defined and calculated a descriptor referred to as DOFlex (or molecular degrees of flexibility) as the sum of: (1) the number of acyclic rotatable bonds, (2) the number of groups attached to triple bonds and (3) the number of aliphatic rings which could potentially also change their geometry. A compound was defined to be drug-like if it satisfied the Lipinski rule-of-five criteria: $M_w \leq 500$, $\log P \leq 5$, H-bond donors ≤ 5 and H-bond acceptors ≤ 10.[50]

2.1.4 Polymorphic subset for optimization. The subset of polymorphic molecules and crystals taken for calculations was constructed by searching the best *R*-factor list of the CSD *vs.* 5.33 (Nov 2011) using the same criteria as for the POL subset. Only different polymorphic crystals are kept in the best *R*-factor list, hence there are no redeterminations. Only REFCODE families with more than one REFCODE were kept.

Since the calculation of lattice energies with accurate methods requires a considerable amount of computational time, we applied further filtering criteria in order to obtain a manageable subset.

(1) Only structures with an *R* factor $< 5\%$.

(2) Only structures with resolved hydrogen atom positions.

(3) Only polymorphic families containing structures with less than 192 atoms per unit cell.

(4) Only ambient pressure polymorphic forms.

The subset used for optimization (POL_{calc}) consisted of 289 polymorphic molecules and 596 crystal structures.

2.1.5 Calculation of lattice energies. We used periodic density functional theory with van der Waals corrections (DFT-d) for geometry relaxations of the polymorphic structures in the POL_{calc} subset. The PBE functional[51] was used with PAW pseudopotentials[52,53] and the Grimme's van der Waals corrections (d2)[54] as implemented in the VASP code (version 5.3.3).[55–58] A kinetic energy cut-off of 520 eV was used. The Brillouin zone was sampled using the Monkhorst–Pack approximation[59] on a grid of *k*-points separated by approximately 0.07 Å (the minimum *k*-point sampling used was $2 \times 2 \times 2$ *k*-points). All atoms and unit cell parameters were allowed to optimize and structural relaxations were halted when the calculated force on every atom was less than 0.003 eV $Å^{-1}$.

Energies obtained from DFT-d codes are normally given per unit cell. We normalized the energies to the number of molecules in the unit cell so that energies can be compared per molecule across the polymorphs. We will refer to the calculated energies per mol as $E_{DFT\text{-}d}$.

2.1.6 Optimised subset of polymorphic structures ($POL_{DFT\text{-}d}$). After attempting the optimization of the 596 crystal structures, some additional filtering was applied. Polymorphic families with structures that did not converge in the optimization procedure were removed. The converged crystal structures were compared with the experimental structures (used as input in the optimization procedure) using the COMPACK algorithm[60] with a 20 molecule cluster and the standard settings. Some of the optimized structures deviated considerably from the experimentally determined ones. This could be due to errors in the experimental structures. In fact, previous studies have used DFT-d calculations to assess the correctness of experimental crystal structures.[61] We removed polymorphic families containing optimized structures that deviated considerably from the experimental X-ray structures. These included structures not matching 20 out of 20 molecules in the COMPACK comparison or structures matching 20 molecules but having an $rmsd_{20}[r] > 0.45$ Å.

After the above-mentioned filtering, 215 polymorphic families containing 446 crystal structures remained. We will refer to this subset as $POL_{DFT\text{-}d}$ and use it for further calculations and data analysis.

2.2 Datasets from Roche & Lilly

As evidenced very much by the historical record in Fig. 1 and discussed earlier, much of the progress in understanding the

chemistry of polymorphism, its manifestations and ramifications has been driven by practical demands and considerations. The rapidly increasing volume of literature on this subject contains many examples of individual studies on polymorphic systems.[62] However, since every compound represents totally unknown territory in terms of the crystal landscape, there is perhaps no better means for demonstrating the variety and vagaries of polymorphism behavior than the cumulative record of two groups of experienced practitioners. Thus we have compiled solid form statistics from 229 solid form screens conducted by Roche and Lilly comprising screens of 145 structurally diverse parent compounds (72 Roche & 73 Lilly) and 84 different salts (Lilly). The screenings were generally conducted early in drug product development to support commercial form selection and ranged in scope from limited to comprehensive. As might be expected in industrial settings, screens were most often limited by design, time or material supply, though material quality (purity) may arguably have also been a factor. Some screenings would have been stopped because of project termination whilst other compounds would have been screened for polymorphs several times at different stages of development. All of the compounds were screened by conventional (manual + semi-automated) methods, with screen designs tailored to the solubility properties of the starting materials, when appropriate. Small subsets of the Roche and Lilly datasets were also subjected to high-throughput methods to pilot the use of automation for polymorph screening. If the high-throughput method yielded a new XRD-pattern, follow up experiments would be repeated in a manual way.

In constructing the Roche and Lilly datasets, crystalline forms were counted only when sufficient physical and chemical data were acquired to support their existence. However, the criteria for establishing a new crystal form were slightly different at the two locations. Whereas a new solid form is designated at Roche only if it can be obtained at least twice and is characterized by various analytical techniques, at Lilly, a single occurrence of a new form might be sufficient, provided the supporting data are unequivocal (*e.g.* a crystal structure from a single crystal isolated from a batch of material). Importantly, amorphous forms and unconfirmed crystalline forms, of which there were many, have not been included in the survey for either the Roche or the Lilly datasets. As such, this tabulation of solid form diversity among typical pharmaceuticals must be considered conservative.

3. Choosing representative monomorphic structures for the CSD datasets

To obtain meaningful statistics of polymorphism in the CSD, it is important to define suitable monomorphic datasets as a basis for comparison. The fact that a given compound only has one unique crystal structure recorded in the CSD says very little about its tendency to exhibit polymorphism. Thus as a general *caveat*, it must be remembered that any statistical analysis based on the CSD relates only to *reported crystal structures*. Many compounds in the CSD have been studied only once crystallographically and often crystal structure determination has been a means for molecular structure validation.

One might initially expect that structures in the polymorphic and monomorphic datasets would span a similar range of molecular size. However, if we plot the normalized distributions of the structures in the POL and MONO datasets as a function of molecular size, *i.e.*, the number of atoms (Fig. 2a), we observe large and apparently significant differences. The maximum of the distribution for the POL dataset is located around 30 atoms whilst that of the MONO dataset appears at approximately 40 atoms. Thus if the complete MONO dataset of the CSD is used for statistical analysis, one might conclude that smaller molecules are far more likely to be polymorphic than larger molecules and that the occurrence of polymorphism in the CSD is 1.9%. A much more likely explanation is that our polymorphic datasets are somewhat biased for smaller molecules. In other words, on average, smaller molecules serve as model compounds for studies concerning polymorphism whilst larger molecules (which are generally less likely to be commercially-available and are harder to synthesize) are more often studied by X-ray single crystal diffraction determination only once.

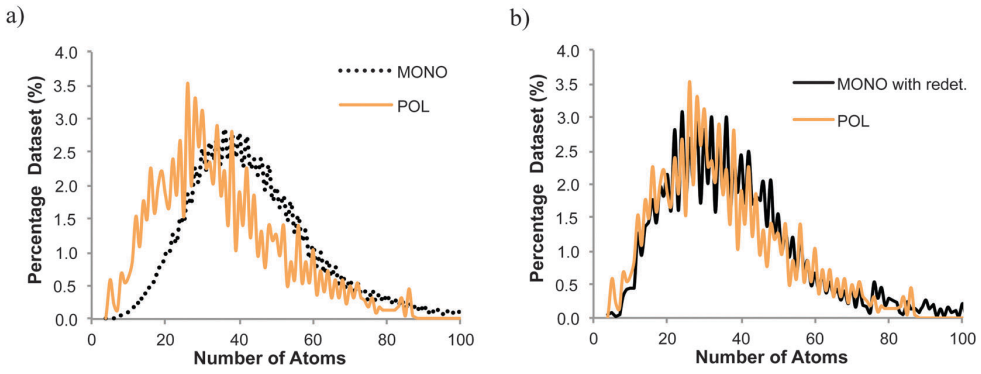

Fig. 2 Normalized distributions for the POL (orange) and MONO (black) datasets (a) and the POL and MONO subset with redeterminations (b) as a function of the number with atoms.

DOCKET ALARM

Explore Litigation Insights

Docket Alarm provides insights to develop a more informed litigation strategy and the peace of mind of knowing you're on top of things.

Real-Time Litigation Alerts

Keep your litigation team up-to-date with **real-time alerts** and advanced team management tools built for the enterprise, all while greatly reducing PACER spend.

Our comprehensive service means we can handle Federal, State, and Administrative courts across the country.

Advanced Docket Research

With over 230 million records, Docket Alarm's cloud-native docket research platform finds what other services can't. Coverage includes Federal, State, plus PTAB, TTAB, ITC and NLRB decisions, all in one place.

Identify arguments that have been successful in the past with full text, pinpoint searching. Link to case law cited within any court document via Fastcase.

Analytics At Your Fingertips

Learn what happened the last time a particular judge, opposing counsel or company faced cases similar to yours.

Advanced out-of-the-box PTAB and TTAB analytics are always at your fingertips.

API

Docket Alarm offers a powerful API (application programming interface) to developers that want to integrate case filings into their apps.

LAW FIRMS
Build custom dashboards for your attorneys and clients with live data direct from the court.

Automate many repetitive legal tasks like conflict checks, document management, and marketing.

FINANCIAL INSTITUTIONS
Litigation and bankruptcy checks for companies and debtors.

E-DISCOVERY AND LEGAL VENDORS
Sync your system to PACER to automate legal marketing.